Variability in sentence processing performance in German people with aphasia and unimpaired German native speakers

Doctoral Thesis submitted to the Faculty of Human Sciences at the University of Potsdam in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Cognitive Science

First Supervisor:
PD Dr. phil. habil. Frank Burchert
Second Supervisor:
Dr. Nicole Stadie

University of Potsdam
Department of Linguistics

submitted by

Dorothea Pregla

15 February 2023

Published online on the
Publication Server of the University of Potsdam:
https://doi.org/10.25932/publishup-61420
https://nbn-resolving.org/urn:nbn:de:kobv:517-opus4-614201

Abstract

Individuals with aphasia vary in the speed and accuracy they perform sentence comprehension tasks. Previous results indicate that the performance patterns of individuals with aphasia vary between tasks (e.g., Caplan, DeDe, & Michaud, 2006; Caplan, Michaud, & Hufford, 2013a). Similarly, it has been found that the comprehension performance of individuals with aphasia varies between homogeneous test sentences within and between sessions (e.g., McNeil, Hageman, & Matthews, 2005). These studies ascribed the variability in the performance of individuals with aphasia to random noise. This conclusion would be in line with an influential theory on sentence comprehension in aphasia, the resource reduction hypothesis (Caplan, 2012). However, previous studies did not directly compare variability in language-impaired and language-unimpaired adults. Thus, it is still unclear how the variability in sentence comprehension differs between individuals with and without aphasia. Furthermore, the previous studies were exclusively carried out in English. Therefore, the findings on variability in sentence processing in English still need to be replicated in a different language.

This dissertation aims to give a systematic overview of the patterns of variability in sentence comprehension performance in aphasia in German and, based on this overview, to put the resource reduction hypothesis to the test. In order to reach the first aim, variability was considered on three different dimensions (persons, measures, and occasions) following the classification by Hultsch, Strauss, Hunter, and MacDonald (2011). At the dimension of persons, the thesis compared the performance of individuals with aphasia and language-unimpaired adults. At the dimension of measures, this work explored the performance across different sentence comprehension tasks (object manipulation, sentence-picture matching). Finally, at the dimension of occasions, this work compared the performance in each task between two test sessions. Several methods were combined to study variability to gain a large and diverse database. In addition to the offline comprehension tasks, the self-paced-listening paradigm and the visual world eye-tracking paradigm were used in this work.

The findings are in line with the previous results. As in the previous studies, variability in sentence comprehension in individuals with aphasia emerged between test sessions and between tasks. Additionally, it was possible to characterize the variability further using hierarchical Bayesian models. For individuals with aphasia, it was shown that both between-task and between-session variability are unsystematic. In contrast to that, language-unimpaired individuals exhibited systematic differences between measures and between sessions. However, these systematic differences occurred only in the offline tasks. Hence, variability in sentence comprehension differed between language-impaired and language-unimpaired adults, and this difference could be narrowed down to the offline measures.

Based on this overview of the patterns of variability, the resource reduction hypothesis was evaluated. According to the hypothesis, the variability in the performance of individuals with aphasia can be ascribed to random fluctuations in the resources available for sentence processing. Given that the performance of the individuals with aphasia varied unsystematically, the results support the resource reduction hypothesis. Furthermore, the thesis proposes that the differences in variability between language-impaired and language-unimpaired adults can also be explained by the resource reduction hypothesis. More specifically, it is suggested that the systematic changes in the performance of language-unimpaired adults are due to decreasing fluctuations in available processing resources. In parallel, the unsystematic variability in the performance of individuals with aphasia could be due to constant fluctuations in available processing resources. In conclusion, the systematic investigation of variability contributes to a better understanding of language processing in aphasia and thus enriches aphasia research.

Acknowledgements

First and foremost, I would like to thank Nicole Stadie and Frank Burchert, who kindly took over the supervision of my work and were always available to discuss the progress of the project and give me feedback. I thank them for all their support, for their trust in me, and for granting me the scientific freedom to make this project my very own.

Furthermore, I would like to thank Shravan Vasishth for his contribution to the progress of the project and my work. He and the members of his lab group strongly influenced me in developing an approach to analyzing my data, interpreting my results, and developing my scientific mindset. In this group, I was inspired by Serine Avetisyan, Sol Lago, Anna Laurinavichyute, Daniela Mertzen, Bruno Nicenboim, Dario Paape, Daniel Schad, Pia Schoknecht, Garrett Smith, Kate Stone, João Veríssimo, Titus von der Malsburg, and Himanshu Yadav. Among this group, I would like to highlight Paula Lissón, with whom I worked closely not only spatially in an office, but also thematically.

This project would not have been possible without the support of the Deutsche Forschungsgemeinschaft and the CRC 1287 - Limits of Variability in Language: Cognitive, Computational, and Grammatical Aspects. The funding not only allowed me to conduct the experiments, but also to participate in conferences and workshops. In addition, the CRC provided me with opportunities for a variety of exchanges with other PhD students and PostDocs on linguistic topics and issues related to scientific work. I would like to thank Clara Huttenlauch, Carola de Beer, Mareike Philipp, Yulia Clausen, and Marc Hullebus, among others, for the fruitful conversations and mutual support. My colleague Leonie Lampe deserves special mention. I thank her for the intensive, trustful and very productive collaboration and for the support and understanding with which she went through the last phase of my PhD thesis with me.

My sincere thanks go to all the participants who completed many hours of experiments with a lot of enthusiasm. Many of them have grown on me over the sessions. In this context, I would also like to thank the student assistants Silke Böttger, Sarah Düring, and Therese Mayr for their commitment, and for helping me with the testing.

I would like to thank my husband Andreas for encouraging me to become a student assistant in Patholinguistics during my undergraduate studies. Without him, I probably would not have followed the path to academia so consistently and would only enjoy it half as much now. I thank him for his support with countless questions about my experiments, for proofreading my scientific texts and the inspiring discussions about linguistics.

Last but not least, I would like to thank my family, who always cared about how everything was going with me and my work, and who always stood behind me. I especially thank my parents for their love and support. It has been a great help that they have repeatedly provided a space for me to work in a focused and well-cared-for way.

Contents

Introduction

Imagine the following situation: An experimenter presents a person with the two pictures in Figure 1 and the sentence in (1), and asks the person to point to the picture that matches the sentence.

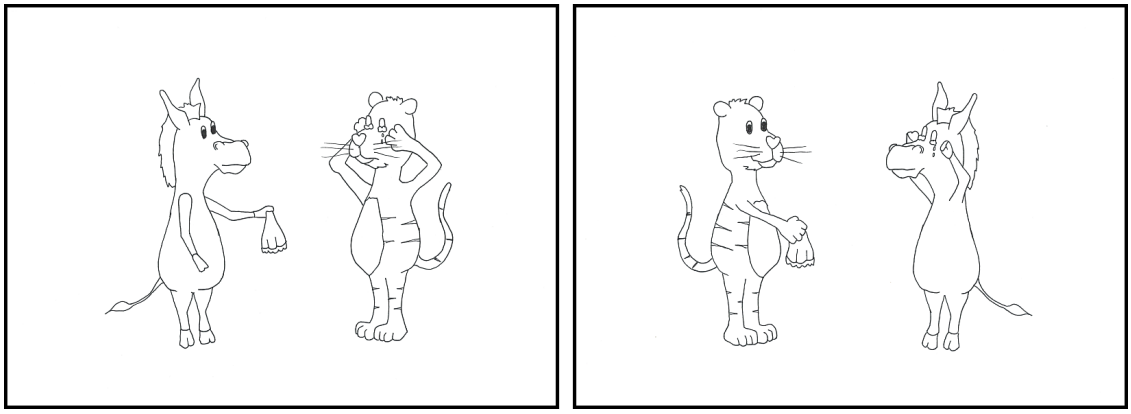

Figure 1: Example pictures of an experiment.

(1) Here is the tiger that the donkey comforts.

How would an "ideal speaker-listener" respond, "who knows its language perfectly and is unaffected by (...) memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of this language in actual performance." (Chomsky, 1965, p. 3)? This ideal person should always understand the sentence correctly and, thus, point to the correct picture immediately and reliably. That is, no matter how often the task is repeated, the comprehension of the ideal speaker-listener should always be correct. Also, if the task demands are increased, e.g., by adding more words to the sentence, the ideal speaker-listener should always process the sentence correctly and give the correct answer immediately.

However, the ideal speaker-listener does not exist. Hence, errors in language processing and limitations in non-linguistic factors such as memory, attention, or task execution influence language processing performance. Due to these influences, a real person will sometimes take a while to figure out the meaning of the sentence, leading to delays in picture selection. Additionally, a real person will occasionally misinterpret the sentence and choose the incorrect picture. That is, sentence processing performance is variable within individuals.

The performance will vary even more if an adult carries out the task who has difficulties in sentence comprehension. For example, individuals with aphasia (IWA), an acquired language disorder resulting from brain damage, will frequently experience difficulties in sentence processing that result in incorrect or delayed responses. This is different from language-unimpaired individuals who will most of the time process the given sentence successfully and choose the correct response. Thus, sentence processing performance is also variable between the speakers of a language.

This variability in sentence processing performance within and between IWA and language-unimpaired adults is the topic of this thesis. More precisely, this work examines the variability in syntactic processing, particularly in thematic role assignment, i.e., the ability to determine "who did what to whom" (Caplan et al., 2007, p.117). This focus is chosen since thematic role assignment is often impaired in aphasia (Caplan et al., 2006; Caramazza & Zurif, 1976). There is potential for variability research in the area of syntactic processing because the language impairment does not lead to stable reductions in sentence comprehension performance. Rather, performance varies within IWA, and the impairment leads to varying performance levels between IWA (Caplan et al., 2006; Caplan et al., 2007; Caramazza et al., 2005; McNeil et al., 2005). This thesis aims to gain a systematic overview of the patterns of variability in syntactic processing of IWA and language-unimpaired adults and, based on this overview, to find an explanation for the patterns of variability in sentence comprehension in aphasia.

In order to systematically investigate variability, a couple of experiments was carried out between July 2018 and August 2020 as part of the collaborative research center 1287 at the University of Potsdam, Germany. The experiments included 21 IWA (mean age = 60.2, range = 38–78 years, 1–26 years post onset) and 50 language-unimpaired adults (mean age = 48, range = 19–83 years), all native speakers of German. All participants were exposed six times to 120 sentences of varying structure, namely declarative sentences (n = 20), relative clauses (n = 60), and control structures with a pronoun or PRO (n = 40). To test for variability in sentence processing due to changes in task demands, the comprehension of these sentence structures was probed with three different tasks, namely an object manipulation task, and two versions of an auditory sentence-picture matching task with two pictures as shown in Figure 1. To test for variability due to changes in sentence processing over time, the performance in each task was compared across two test phases spaced approximately two months apart. Data were analyzed using hierarchical Bayesian models. The Bayesian analysis was chosen since the variance between and within participants could be estimated simultaneously and without convergence issues.

The results of the experiments were published in two articles in the journal *Brain & Language*. These two articles are part of this cumulative dissertation and form two broad sections of the thesis called Study 1 and Study 2. The third broad section of

this thesis, the Synopsis, is a review paper that merges the results of the two published articles, presents additional unpublished data, and provides an overarching explanation for the findings.

The synopsis will begin by highlighting the relevance of studying variability in sentence comprehension in aphasia (Chapter 1). The most critical studies on this topic are introduced, and the gaps in the available research are identified. Additionally, an outlook is given how this thesis intends to fill these gaps. Subsequently, it is outlined how the research on variability in aphasia has developed (Chapter 2). This chapter provides a rough overview of the state of research and different explanations for the variability in aphasia from the beginning of aphasia research until today. Based on this overview of previous research, a working definition of variability in sentence comprehension in aphasia is developed for the thesis (Chapter 3). This definition of variability will then be operationalized (Chapter 4). For this purpose, different ways to statistically analyze variability are discussed, and the most appropriate way to analyze the data for the purpose of this thesis is selected. At this point, the prerequisites for the formulation of the exact questions of the thesis are met, which are introduced in Chapter 5. The aim of the following chapters is to answer these questions. To this end, the published findings presented in Study 1 and Study 2 are first summarized (Chapter 6). Then, Study 3 presents further unpublished data that add new results regarding variability in sentence comprehension in aphasia (Chapter 7). Thus, Chapters 6 and 7 bring together all the findings of this research project. Based on this broad data basis, the following discussion (Chapter 8) can give a differentiated answer to the question of which patterns of variability in sentence comprehension exist in aphasia. Furthermore, it is discussed how variability in sentence comprehension differs between people with and without aphasia, and several suggestions are provided why these differences in variability occur. Finally, the research questions formulated in the thesis are answered once again briefly and precisely (Chapter 9).

Synopsis

1 Motivation for studying variability in aphasia

Traditionally, neuropsychological research focuses on the mean performance of an individual or a group. In such investigations, variability in performance is considered a nuisance and not an object to be studied. However, emerging research suggests that in addition to mean performance, variability in performance can also be characteristic of an individual or a group (Hultsch et al., 2011). According to Hultsch et al. (2011), various neurological diseases (traumatic brain injury, Parkinson's disease, dementia) are associated with increased variability in performance relative to neurotypical performance. This suggests that variability may be indicative of a disturbance in cognitive functioning.

Given that variability is a frequent symptom of neurological disorders, high variability in performance would also be expected in aphasia since it is an acquired language disorder that occurs as a result of neurological damage to the brain. Indeed, both the affected individuals themselves and speech-language therapists frequently report a high variability in language performance of IWA (Mack et al., 2016). These subjective observations are corroborated by objective findings of variability in aphasia at the non-linguistic level (Villard & Kiran, 2015, 2018), at the word level (Ciccone, 2003; Freed et al., 1996), at the level of sentence comprehension (Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 2007; Caplan et al., 1997; Hageman et al., 1982; McNeil, 1983, 1988; McNeil et al., 2005; McNeil et al., 2015), and the discourse level (Boyle, 2014; Brookshire & Nicholas, 1994). McNeil and Pratt (2001) and Odell et al. (1995) even consider variability in performance "a hallmark and defining feature of aphasia" (McNeil & Pratt, 2001, p. 909). However, despite a large body of reports suggesting the presence of variability in aphasia, the number of studies defining the scope, relevance, and causes of variability in aphasia is still small (McNeil et al., 2005; Nespoulous, 2000; Zakariás & Lukács, 2021). The presence of variability in the performance of IWA and the sparsity of studies investigating this variability motivate the research on variability in aphasia reported in this thesis.

While variability can be a sign of a neurological impairment, it can also be a sign of preserved abilities. For example, neurolinguists researching variability repeatedly argued that a damage of linguistic representations should lead to stable performance

patterns. A variable performance pattern, however, indicates that the underlying representations of linguistic knowledge are intact (Caplan et al., 2006; Caplan et al., 2007; Caplan et al., 1997; Hula et al., 2007; Kolk & Van Grunsven, 1985; McNeil, 1983, 1988; McNeil et al., 2005; McNeil et al., 1991). This is because a performance which is at times impaired and at times unimpaired demonstrates that an individual is, in principle, able to perform a task, although not reliably. An intermittent unimpaired performance should not be possible if the linguistic knowledge needed for successful task performance was permanently lost. Therefore, a variable performance rather speaks for an impairment in the access to linguistic knowledge than for an impairment in the linguistic knowledge itself (Caplan et al., 2006; Caplan et al., 2007; Caplan et al., 1997; Hula et al., 2007; Kolk & Van Grunsven, 1985; McNeil, 1983, 1988; McNeil et al., 2005; McNeil et al., 1991). Thus, investigating variability can yield information about the integrity of linguistic representations in aphasia, and, on a broader level, it can be an indicator of the performance potential of IWA. Thus, a second motivation for this thesis is to reveal the potential of IWA which might be underestimated by looking at the average performance only.

This thesis will specifically focus on the variability in aphasia in sentence comprehension performance. Two groups of researchers have already investigated variability in sentence comprehension in aphasia. McNeil and his colleagues focused on the variability in sentence comprehension performance of IWA within and between test sessions (Hageman et al., 1982; Hula & McNeil, 2008; McNeil et al., 2005; McNeil et al., 2015). Caplan and his colleagues investigated the variability in sentence comprehension performance of IWA between tasks (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan et al., 1997). To motivate a further investigation of variability in sentence comprehension in aphasia, the work of these groups of authors will be briefly introduced below. A more extended discussion of these authors' works can be found in Study 1, Chapter 1.1 and 1.2.

McNeil and his colleagues studied the variability in IWA by comparing their performance within or between sessions within a task. The prerequisite for the investigation of the variability within a task is high item homogeneity (McNeil et al., 2005). Therefore, the authors employed their previously developed Revised Token Test (RTT, McNeil & Prescott, 1978), which has both high internal consistency and high test-retest reliability (McNeil et al., 2015). Variability between items of the RTT was defined as the change in scores between consecutive items (McNeil et al., 1982). Responses in the RTT consist of pointing gestures (e.g., *touch the red circle and the blue square*). Participants are scored with a system in which responses are ranked from *no response* (1 point) over *rejection of the task* (5 points) and *reversal of the nouns* (10 points) to *correct response* (15 points, McNeil et al., 1989, Table 1). The authors showed that IWA's performance was highly variable within subtests of the RTT, between subtests of the RTT, and between two completions of the RTT in a test-retest design (Hageman et al., 1982; Hula & McNeil,

2008; McNeil et al., 2005; McNeil et al., 2015). The authors took this variability to indicate that language mechanisms are preserved in IWA since they are intermittently capable of carrying out the RTT correctly (Hula & McNeil, 2008; McNeil et al., 2005). The authors ascribed the variability in the performance to an impairment in the attentional system that allocates an insufficient amount of resources to the language task (Hula & McNeil, 2008; McNeil et al., 2005).

In contrast to McNeil et al., who examined variability *within* a task, Caplan et al. tested variability in IWA *between* different tasks, e.g., sentence-picture matching and object manipulation. Furthermore, Caplan et al. compared sentence structures with varying syntactic complexity, i.e., structures which do (simple) or do not (complex) require a particular syntactic operation (Caplan et al., 2015, 2013a; Caplan et al., 2007). The authors predicted that IWA should consistently show more sentence comprehension difficulties in syntactically complex versus simple sentences. However, in two large group studies including around 100 IWA, there was not a single IWA who consistently showed more difficulties for complex sentences across tasks (Caplan et al., 2013a; Caplan et al., 2007). Instead, the performance varied within IWA between tasks and sentence structures. The variability in sentence comprehension performance was used as an argument against a specific syntactic deficit in IWA (Caplan, 2012; Caplan et al., 2013a; Caplan et al., 2007). The authors explained the results by *resource reduction*. The resource reduction hypothesis is presented in Study 2, Chapter 1.1 in detail. In brief, the resource reduction hypothesis states that sentence comprehension is impaired when the task demands exceed the resources of the IWA. Variability in performance may originate from a noise-based fluctuation in the resources available to the IWA (Caplan, 2012).

This thesis will combine the approaches to studying variability of McNeil and his colleagues and Caplan and his colleagues. More specifically, both the variability *between* tasks and the variability between sessions *within* a task will be investigated. This will be done in order to gain a detailed picture of the variability in sentence comprehension performance in aphasia.

The thesis aims to replicate and extend the findings of McNeil and his colleagues and Caplan and his colleagues. Those previouse studies were exclusively carried out with English-speaking participants. Therefore, this work will expand the knowledge about variability by examining a different language than the previous studies, namely German. Studying a new language allows for investigating additional syntactic phenomena. McNeil et al. examined variability in sentence comprehension using declarative sentences of increasing length (McNeil & Prescott, 1978). Caplan et al. used sentence structures with varying syntactic complexity, namely relative clauses, clefts, passives, control structures, pronouns, and reflexives (e.g., Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 1997). The present investigation replicates the previous studies by also examining declarative sentences, relative clauses, pronouns and control structures. Addi-

tionally, in contrast to the earlier studies, declarative sentences with subject-object and object-subject word order will be compared. This comparison is impossible in English monoclausal declarative sentences due to rigid word order. Furthermore, different from the earlier studies, the influence of grammatical gender on pronoun resolution will be examined, which is not possible in English because it has no grammatical gender. Thus, by studying German, it can be shown whether the findings on variability in syntactic processing in English can be replicated in another language and generalized to other syntactic phenomena.

Furthermore, the thesis aims to extend the previous studies' findings by collecting new measures. McNeil et al. measured response time, accuracy, and reading time (e.g., McNeil et al., 2005; McNeil et al., 2015). Caplan et al. measured accuracy in object manipulation, response time and accuracy in sentence-picture-matching and grammaticality judgment, and listening times in self-paced listening (e.g., Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 1997). The current investigation uses similar measures as McNeil et al. and Caplan et al. by measuring accuracy in object manipulation, response time and accuracy in sentence-picture-matching and listening times in self-paced listening. Additionally, the present work extends the measures by collecting eye-movement data using the visual world paradigm. This additional measure provides temporally more accurate insights into sentence processing than the other measures since eye movements do not require a conscious decision, initiation of a response, and execution of a hand movement (Dickey et al., 2007). Furthermore, the collection of different online (eye-movements, listening times) and offline (response times, accuracy) measures makes it possible to compare the variability in online and offline processing. Such a comparison has not been carried out in previous studies and possibly allows a more precise statement about variability in sentence processing in aphasia.

Finally, the thesis aims to expand on the earlier results by comparing the variability in sentence processing in IWA to language-unimpaired adults. While both McNeil et al. and Caplan et al. included language-unimpaired control groups in their studies (e.g., Caplan et al., 2015; McNeil et al., 2015), they did not focus on the variability in the control group or they did not directly compare the variability in the language-impaired and language-unimpaired groups. One reason for excluding such comparisons might be that language-unimpaired participants show ceiling effects in accuracy scores in the comprehension tasks. While similar ceiling effects are expected for the accuracy in this thesis as well, the other response measures (i.e., reaction times, listening times, eye-tracking measures) can still vary in the control group. For example, James et al. (2018) found that in a group of more than 100 language-unimpaired adults with ages ranging from 18 to 67 years the within-participant consistency was low in sentence processing as measured with a self-paced reading task. Furthermore, Hultsch et al. (2002) showed that the variability in lexical decision times is increased in older (54–94 years) compared

to younger (19–36 years) adults. Thus, a certain degree of variability seems to be present not only in aphasia but also in unimpaired language processing. The comparison of the variability in IWA with the variability in language-unimpaired individuals in the present work makes it possible to determine whether, and in what way, the variability in aphasia deviates from the variability in the language-unimpaired population.

In summary, investigating variability in aphasia at the sentence level is warranted because prior studies suggest that variable behavior exists at this level. The existing knowledge will be extended by comprehensively investigating variability with new sentence structures and measures in language-impaired and language-unimpaired adults.

2 Tracing variability in the history of aphasia research

The high degree of variability between IWA has already been documented in the 19th century. For example, the "father of English neurology" (Tesak & Code, 2008, p.54) John Hughlings Jackson writes in one of the first issues of *Brain* (Hughlings Jackson, 1878):

> [D]ifferent amounts of nervous arrangements in different positions are destroyed with different rapidity in different persons. There is, then, no single well-defined "entity" — loss of speech or aphasia — and thus, to state the matter for a particular practical purpose, such a question as, "Can an aphasic make a will?" cannot be answered any more than the question, "Will a piece of string reach across this room?" can be answered. The question should be, "Can this or that aphasic person make a will?"
>
> (Hughlings Jackson, 1878, p.314)

With his view, Hughlings Jackson differed from the localizationists that dominated in his time (Tesak & Code, 2008). The aim of the localizationists was to find the "seat of language" in the brain (Broca, 1861) based on IWA with similar lesions. Hughlings Jackson, however, considered it unlikely that such a "seat of language" exists given the variability between IWA (Hughlings Jackson, 1878).

The localizationists identified syndromes based on groups of IWA that exhibited relatively homogenous symptoms. Each syndrome was associated with a lesion in a certain area of the brain (Wernicke, 1874). The approach of determining syndromes and assigning them to brain lesions is called the *syndrome approach*. The syndrome approach was criticized, amongst others, by Henry Head, supporter of Hughlings Jackson's theory, who stressed that individuals with the same lesion differed in their preserved language ability (Caplan, 1987). In addition, Head described another form of variability, the variation within a single IWA:

> An inconstant response is one of the most striking results produced by a lesion of the cerebral cortex. [...] It is not a sufficient test to hold up some object, and ask the patient to name it; at one time he may be able to do so, at another he fails completely. No conclusion can be drawn from one or two questions put in this way; his power of responding must be tested by a series of observations in which the same task recurs on two or more occasions. (Head, 1920, p.89–90)

Head's observation of variability within IWA was picked up by Kurt Goldstein (Caplan, 1987). Goldstein reasoned that IWA avoid tasks whose demands exceed their abilities because these tasks might lead to a "catastrophic condition", in which the IWA's "expression [becomes] one of helplessness, desperate, angry" (Goldstein, 1948, p.260). Ac-

cording to Goldstein, variability results from the avoidance behavior of the IWA and from changes in task demands (Goldstein, 1948).

Also Goldstein's student Egon Weigl addressed the variability in IWA. According to Weigl, the variability suggests that the linguistic competence of IWA is intact but their linguistic performance is impaired (Weigl & Bierwisch, 1970). The division into competence and performance goes back to Noam Chomsky, who introduced generative linguistics in the late 1950s and 1960s (Chomsky, 1965). Competence refers to the abstract knowledge about the grammar of a language, performance refers to the actual use of language (Chomsky, 1965). Weigl and Bierwisch (1970) list three reasons for preserved linguistic competence in IWA: 1) If competence were lost, each performance would have to be associated with an own competence to explain differences in performance between IWA. 2) Furthermore, if competence were lost, fluctuations in performance should not occur within IWA. 3) Finally, if competence were lost, deblocking should be impossible in treatment. Thus, both the variability between IWA as well as the variability within IWA played an important role in arguing for preserved linguistic competence in IWA.

Another theory that drew on variability between and within IWA to argue for preserved linguistic competence was introduced by Kolk and Van Grunsven (1985). The authors discussed several explanations for variable performance in sentence comprehension and production. The first option they considered was a partial loss of linguistic competence, i.e., a loss of a specific syntactic rule. A loss of a specific rule can explain variability between different syntactic structures, because the rule might affect comprehension of structure A but not structure B. However, the loss of a specific rule cannot explain variability in the comprehension of "one and the *same* type of construction"(Kolk & Van Grunsven, 1985, p.366). Therefore, partial loss was rejected as an explanation for variability. A second option discussed by Kolk and Van Grunsven (1985) were general limitations in cognitive abilities, such as working memory. However, according to the authors, this explanation is too unspecific to derive predictions regarding the linguistic behavior of IWA. Additionally, the authors ruled out the theory that variability is caused by random noise and an increased threshold in the access of linguistic knowledge. Alternatively, Kolk and Van Grunsven (1985) proposed adaptation as a strategy of IWA that leads to variability. That is, IWA would be able to adapt their linguistic behavior depending on the purpose of communication and the severity of their language disorder (Kolk & Van Grunsven, 1985). One way that IWA adapt sentence comprehension would be that they rely more on pragmatic context than on syntactic structure. One way that IWA adapt sentence production would be that they omit syntactic elements during conversation (telegraphic speech) to compensate for their difficulty with syntactic structure building. However, IWA can also choose to produce a more complex syntactic structure and, e.g., reduce the speech rate instead, depending on their "weighing of costs and benefits" (Kolk & Van Grunsven, 1985, p. 376). This flexibility in the adaptive behavior of the

IWA in a given communicative situation leads to the observed variability in performance according to Kolk and Van Grunsven (1985).

Around the time generative linguistics was introduced by Noam Chomsky, the syndrome approach was revived by Norman Geschwind (Tesak & Code, 2008). Group studies were conducted to investigate aphasic syndromes in detail, and to draw inferences from the functional deficits to the location of brain damage (Tesak & Code, 2008). In this context, the variability debate emerged in the 1970s which was concerned with the validity and reliability of dividing IWA into groups according to their syndromes (Tesak & Code, 2008). The usefulness of group studies was called into question because of the large variability between individuals subsumed under a syndrom (Caramazza, 1986). As an alternative, researchers advocated for the single case approach (Caramazza, 1986). On the basis of dissociations in performance between single IWA, components in language models were identified that can be selectively impaired (Tesak & Code, 2008).

The variability debate also took place in the context of sentence comprehension impairments. The basis for the debate was a study by Caramazza and Zurif (1976), who demonstrated that individuals with Broca's aphasia had difficulties understanding reversible object relative clauses, in which either of two nouns could be a plausible agent of the subclause (Caramazza & Zurif, 1976). This was an important finding because it disconfirmed the established view that individuals with Broca's aphasia have intact sentence comprehension abilities (Tesak & Code, 2008). Caramazza and Zurif (1976) concluded that individuals with Broca's aphasia can no longer apply syntactic rules and rely on semantic plausibility and heuristics for sentence interpretation (Caramazza & Zurif, 1976).

In subsequent studies, it was confirmed that individuals with Broca's aphasia can have sentence comprehension impairments, but the explanation that the syntactic rules are entirely lost turned out to be too strong (Caplan, 1987). One alternative explanation, that was discussed intensively in the context of the variability debate, is the trace deletion hypothesis by Yosef Grodzinsky (Grodzinsky, 1986, 1995, 2000). Grodzinsky argued that a specific part of syntactic structure building is impaired in individuals with Broca's aphasia, namely the ability to link syntactic traces with their antecedents (Grodzinsky, 2000). This specific syntactic deficit would lead to a selective impairment of sentences with moved arguments and reversed thematic order, such as object relative clauses (Grodzinsky, 2000). The comprehension would be normal in sentences without movement or with linear thematic order, such as subject relative clauses (Grodzinsky, 2000). Importantly, this performance pattern was predicted for all individuals with Broca's aphasia (Grodzinsky, 2000). Variations within and between individuals were interpreted as mere artifacts of guessing behavior, a "statistical property of chance performance" (Grodzinsky et al., 1999, p.144). Due to guessing behavior, the results of a single IWA were predicted to deviate randomly from the pattern which emerges at the

group level (Drai & Grodzinsky, 2006; Grodzinsky et al., 1999). Therefore, Grodzinsky and his colleagues concluded that group studies are necessary because "the findings from any one patient, without the context of a group, may give a distorted picture of the pathological reality" (Grodzinsky et al., 1999, p.135). Furthermore, they concluded that "the group's performance is stable, and well-delineated, despite intersubject variation" (Grodzinsky et al., 1999, p.134).

The conclusion that individuals with Broca's aphasia demonstrate a uniform performance pattern was criticized by several authors (e.g., Berndt et al., 1996; Caplan, 2001; Caramazza et al., 2001; De Bleser et al., 2006; Toraldo & Luzzatti, 2006). It was pointed out that Grodzinsky's group level analysis masks meaningful qualitative differences between individuals with Broca's aphasia (Toraldo & Luzzatti, 2006). For example, Berndt et al. (1996) found in a meta-analysis that only some individuals showed the performance pattern predicted by the trace deletion hypothesis, whereas a number of individuals had no problems in sentence comprehension or had similar problems with all sentence types. Given this heterogeneity, Caramazza et al. (2005) argued in opposition to Grodzinsky and his colleagues:

> Therefore, it makes little sense to ask what is the cause (singular) of the comprehension impairment in agrammatic Broca's aphasia. This is because there is not a single type of comprehension impairment associated with agrammatic Broca's aphasia. Instead, different types of comprehension performance are found in such patients, most likely reflecting different mixtures of damage to the various cognitive and linguistic mechanisms involved in sentence processing. (Caramazza et al., 2005, p.51)

Thus, variability between IWA was the basis to argue against a single specific syntactic impairment in individuals with Broca's aphasia. The variability was rather interpreted to suggest that various linguistic and general cognitive deficits may lead to impairments in sentence comprehension.

The variability debate in the context of sentence comprehension impairments continued into the 21st century. For instance, Drai and Grodzinsky (2006) stated:

> Thus, while variability exists [...] the robust structure we uncovered in the data, and its relation to clinical diagnostic tests of Broca's aphasia, are clear. Still, the variability debate is unlikely to stop here.
>
> (Drai and Grodzinsky, 2006, p.125)

Despite the frequent references to variability, little research directly investigated variability at the sentence level in IWA in a systematic manner. Two notable exceptions are the studies by Malcolm McNeil and David Caplan and their colleagues, which were already introduced briefly above, and which are discussed in greater detail in Study 1,

Chapter 1.1 and 1.2 (Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 2007; Caplan et al., 1997; Hageman et al., 1982; McNeil, 1983, 1988; McNeil et al., 2005; McNeil et al., 2015). Both groups of authors included IWA with different aphasic syndromes and lesion locations in their studies. They consistently found that there was no relationship between aphasia syndrome or lesion location and the variability between IWA in sentence comprehension performance (Caplan et al., 1997; McNeil et al., 1991).

Apart from the studies by McNeil et al. and Caplan et al., the topic of variability in aphasia received attention mainly outside the area of sentence processing in the last twenty years. Two important topics discussed at the moment are first, finding new methods to handle inter-individual variability in group studies, and second, implementing inter-individual variability in cognitive theories and models (Nickels et al., 2011; Schwartz & Dell, 2010; Shallice, 2015). The journal Cortex (2017) devoted a special issue to these topics to raise awareness for inter-individual variability within the cognitive neuropsychology community (De Schotten & Shallice, 2017). In this special issue, Halai et al. (2017) consider different methods to account for inter-individual variability in experiments. Concerning group studies, the authors remind the audience that group means can mask meaningful differences between participants. Concerning single case studies, the authors critize the low generalizability of the results. As a solution, Halai et al. (2017) offer principal component analysis and illustrate their approach using data from 31 IWA. In the same special issue, Friedman and Miyake (2017) present a cognitive framework taking inter-individual variability into account, called the unity/diversity framework for individual differences in executive functions. The authors demonstrate how inter-individual differences in executive function tasks and clusters in participants' performance can be useful to identify parameters in a cognitive model.

Besides *inter*-individual variability, recent investigations also focused on *intra*-individual variability in IWA. Authors examined intra-individual variability as a possible predictor of treatment response. For example, Duncan et al. (2016) reported that IWA with large pre-treatment variability in imitating words and phrases improved more in imitation than IWA with small pre-treatment variability. Based on these results, the authors suggest that variable pre-treatment performance indicates high learning potential, while stable pre-treatment performance reflects the maximum of an individual's performance (Duncan et al., 2016). Furthermore, intra-individual variability could shed light on the question whether IWA have a general cognitive impairment (Laures, 2005; Perez Naranjo et al., 2018; Villard & Kiran, 2015, 2018). To explore this question, authors used different linguistic and non-linguistic attention tasks where keys should be pressed for auditory or visual targets (e.g., sounds, words, or letters) but not for distractors (Laures, 2005; Perez Naranjo et al., 2018; Villard & Kiran, 2015, 2018). Compared to language-unimpaired control participants, IWA were more variable in their responses, and their variability increased more when task complexity increased (Laures, 2005; Perez Naranjo

et al., 2018; Villard & Kiran, 2015, 2018). The authors attributed this intra-individual variability to an inefficient activation of linguistic and non-linguistic information caused by disturbances in attention (Laures, 2005; Perez Naranjo et al., 2018; Villard & Kiran, 2015, 2018). The most recent investigation on intra-individual variability in aphasia is currently underway at Eötvös Loránd University in Budapest (Zakariás & Lukács, 2021). The study examines the variability in response accuracy and reaction time in phoneme identification, lexical decision, and semantic decision tasks between six sessions. Preliminary results for 13 IWA indicate a negative correlation between mean accuracy and intra-invididual variability in accuracy and a negative correlation between performance on standardized language tests and intra-invididual variability in accuracy (Zakariás & Lukács, 2021). These results may imply that intra-individual variability is increased in severely versus mildly impaired IWA (Zakariás & Lukács, 2021).

To conclude this chapter, the observation of variability in aphasia is as old as the investigation of aphasia itself. Yet, systematic examinations of variability in aphasia are rare. Variability is raised as a major criticism of prominent approaches in aphasia research based on the homogeneity of IWA, such as the syndrome approach. Researchers consistently invoked variability as an argument for preserved linguistic knowledge (Caplan et al., 2007; McNeil et al., 1991; Villard & Kiran, 2018; Weigl & Bierwisch, 1970). There is less agreement among researchers on the source of variability in IWA. For example, it could originate from adaptation of the IWA to the linguistic context (Kolk & Van Grunsven, 1985), an impairment in attention leading to an insufficient amount of resource allocation to linguistic processing (Hula & McNeil, 2008), or random fluctuation in the resources available to the IWA (Caplan, 2012). Most of the recent studies ascribe variability to a general cognitive deficit in the attention allocation mechanism (Laures, 2005; Perez Naranjo et al., 2018; Villard & Kiran, 2015, 2018).

3 Definition of variability

There is no fixed term in the aphasia literature to refer to variability in IWA. In addition to the term variability (Nespoulous, 2000), other terms used are variation (Grodzinsky, 2000), noise (Drai & Grodzinsky, 1999), individual differences (Villard & Kiran, 2015), dissociation (Caplan et al., 2013a), heterogeneity (Nickels et al., 2011), inconsistency (Perez Naranjo et al., 2018), or fluctuation (McNeil et al., 2005). According to Hultsch et al. (2011), research on variability can span three dimensions: *persons*, *measures*, and *occasions*. The variability in the dimension of *persons* is studied by comparing the performance of different individuals. Variabiliy in persons is also referred to as between-participant variability, inter-individual differences or inter-individual variability. In contrast, the variability in one individual is called within-participant variability or intra-individual variability. Intra-individual variability can be investigated on the dimensions of measures and occasions. Variability in the dimension of the *measures* is studied by comparing the performance of the same participant(s) across different tasks. Variability in the measure is also referred to as between-task or across-task variability (Nespoulous, 2000). In aphasia research, between-task variability has for example been examined by Caplan et al. (2007, 2013a, 2015). Finally, variability in the dimension of the *occasions* is studied by comparing the performance of the same participant(s) in the same task across different time points. Variability in the occasions is also referred to as within- or between-session variability (Villard & Kiran, 2015, 2018). In aphasia research, within- and between-session variability has, for example, been investigated by McNeil et al. (2005, 2015). In addition, it is common to simultaneously vary the dimensions (i.e., persons, measures, and occasions) to gain insight into different forms of variability in one study (Hultsch et al., 2011). This approach will be adopted in this thesis. That is, all three dimensions are varied in order to investigate the variability in sentence comprehension in aphasia within and between participants, between tasks, and between test phases.

In Chapter 2, it was demonstrated that there are different views in aphasia research about what variability is in IWA. According to some researchers, variability in the performance of IWA is random noise (Caplan, 2012; Caplan et al., 2006). According to other researchers, at least some of the variability in the performance of IWA is systematic and meaningful (Johnson & Cannizzaro, 2009; Mack et al., 2016; Nespoulous, 2000). The fact that researchers adopt two fundamentally different views of variability becomes clear, for example, in the variability debate. While one side of researchers held a broad view that variability could be partly random and partly systematic and meaningful (Caramazza, 1986; Caramazza et al., 2001), the other side of researchers held a narrow view that variability is only random noise (Grodzinsky, 2000; Grodzinsky et al., 1999).

Bearing the different views on variability in mind, one aim of this thesis will

be to gain insights into the question whether variability in aphasia is systematic or unsystematic. A starting point to approach this question will be the resource reduction hypothesis (Caplan, 2012) which was already mentioned briefly in the previous chapters and will be described in greater detail in Study 2, Chapter 1.1. The hypothesis was chosen as a starting point because it is a major theory of sentence comprehension in aphasia in which variability in the performance is an integral component. Specifically, the resource reduction hypothesis assumes that the resources needed for sentence processing fluctuate randomly due to "noise in the system" (Caplan, 2012, p. 47). These fluctuations in resources cause the variability in sentence comprehension. As discussed in Chapter 1.1 of Study 1 and 1.1 of Study 2, the resource reduction hypothesis is underspecified with respect to the nature of the resources and the reason for the noise in the system. However, the crucial point is that, according to the resource reduction hypothesis, variability in aphasia should be unsystematic.

Although it is an important theory of sentence comprehension in aphasia, the resource reduction hypothesis has only been investigated in English so far. Furthermore, the assumption that variability is unsystematic is only based on the studies of Caplan's own research group (Caplan et al., 2015, 2013a; Caplan et al., 2007). Additionally, Caplan and his colleagues only investigated variability between measures but not the variability between occasions, which might be systematic. To overcome these issues, the present work investigates the resource reduction hypothesis in a different language and including different test phases. Especially, the thesis evaluates the assumption of the resource reduction hypothesis that the variability in sentence comprehension performance in aphasia is unsystematic. This requires taking a step back and considering the occurrence of systematic and unsystematic variability as equally valid possibilities. Therefore, the present work will adopt the broad view that the variability in sentence comprehension performance of IWA might be systematic or unsystematic. Examples of systematic differences would be inter-individual differences relating to aphasia severity, between-task differences relating to task complexity, and within- or between-session differences relating to practice or fatigue effects, adoption of strategies, adaptive behavior, or learning (Halai et al., 2017). Examples of unsystematic differences would be measurement error and random variations in performance due to noise (Halai et al., 2017).

Based on the classification of Hultsch et al. (2011), the following working definition of variability is established for this thesis: Variability denotes a difference in the performance patterns within an individual or between individuals in one or different measures on one or different occasions. This difference in performance patterns might be systematic and following a set pattern or it might be unsystematic and random.

4 Operationalization of variability

Just as there is no standard definition of variability, there is no common approach to operationalizing variability. While some researchers *identify* whether a behavior is variable, other researchers *quantify* how much variability there is in the behavior. In this chapter, both ways of determining variable behavior are introduced.

Among the researchers that *identify* the presence of variability are McNeil and his research group (Hageman et al., 1982; Hula & McNeil, 2008; McNeil et al., 2005; McNeil et al., 2015) and Caplan and his research group (Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan et al., 1997). McNeil et al. (2005) use a cut-off value to identify intra-individual variability. They describe their approach as follows: "A difference score of .20 or more between items or subtests was required in order for items or subtests to be judged as different (e.g., 12.00 would be judged as different from 12.20 but not 12.10)." (McNeil et al., 2005, p.180). Thus, McNeil et al. operationalize variability as the difference between two raw values exceeding a minimum value. Caplan et al. (2006, 2007, 2013a) speak of intra-individual variability when performance is "poor" (Caplan et al., 2013a, p. 24) in one task and "good" (ibid.) in another. More specifically, the authors consider the performance between tasks as variable if three criteria are met. First, the performance in the poor task must be at the chance level; second, the performance in the poor task must be significantly below a language-unimpaired control group and the performance of the good task must be within the range of the language-unimpaired control group; and third, performance in the poor task must be significantly lower than in the good task. Thus, Caplan et al. use a more complex measure to identify variability than McNeil et al. since they consider the average performance level and the chance level. However, in principle, both approaches identify variability based on the difference between the scores of two items, subtests, or tasks.

The simplest way to *quantify* variability is to calculate the standard deviation based on the raw-score responses of the participants (Dykiert et al., 2012; Hultsch et al., 2011). The standard deviation of the sample's mean scores quantifies between-participant variability, and the standard deviation of the participants' mean scores quantifies within-participant variability. However, this method has disadvantages: The mean response time, for example, is often positively correlated with the standard deviation (Dykiert et al., 2012). Therefore, differences in variability may be confounded by systematic differences in mean response time between groups, e.g., a language-impaired and a language-unimpaired group (Hultsch et al., 2011). To account for the association between mean and standard deviation, some studies investigating variability in IWA have used a slightly modified measure called the coefficient of variation (COV, Mack et al., 2016; Villard & Kiran, 2015, 2018). The COV is obtained by dividing the standard deviation of the sample's mean scores by the sample's mean scores (Mack et al., 2016; Villard

& Kiran, 2015, 2018). However, another disadvantage of using the standard deviation as a measure of variability is not eliminated with the COV. Namely, the standard deviation only provides a single value for variability. Therefore, it is impossible to distinguish between different sources of variability (Hultsch et al., 2011). For example, it is not possible to discriminate between variability due to changes over time (e.g., practice effects), task demands, or language impairment. Additionally, the latter types of systematic variability cannot be separated from unsystematic variability, i.e., random fluctuation in the performance.

One option to model unaggregated data and investigate different sources of variance at the same time is offered by linear modeling. This way of quantifying variability will be illustrated here using a hierarchical linear model with uncorrelated intercept and slope adjustment for participants where participants are indexed as *i*, and items are indexed as *j*. The notation follows Vasishth et al. (2022, chap. 3.5.3), and the subsequent explanations of the model are based on Gelman and Hill (2007), Navarro (2013), Nicenboim et al. (2022), and Vasishth et al. (2022). First, the formula will be described from left to right without the different variance components, which will be explained afterwards.

$$y_{ij} = \beta_0 + u_{0i} + (\beta_1 + u_{1i}) \times x_{ij} + \epsilon_{ij}$$

In this formula, y_{ij} is the value of the dependent variable for the ith participant for the jth item, e.g., a participant's response time for an item. The parameter β_0 represents the intercept, i.e., the value for the dependent variable when the value of the independent variable is zero. The intercept could be, e.g., the grand mean response time. The parameter β_1 represents the slope, i.e., the value by which the dependent variable increases when the value of the independent variable increases by 1. The slope could be, e.g., the difference in response time between subject and object relative clauses. The independent variable is represented by x_{ij} and could be, e.g., the relative clause type that participant *i* saw for item *j*, .

In this model, three sources of variance can be distinguished. The two variances, u_{0i} and u_{1i}, are adjustments to the mean intercept and mean slope for each participant. That is, u_{0i} and u_{1i} reflect how much each participant deviates from the mean intercept and slope. For example, the difference between subject and object relative clauses might be larger or smaller than the average for some participants. Thus, the two variances are measures of between-participant variability in the effects. The variances u_{0i} and u_{1i} are estimated considering the data of all participants. As a result, the estimates for each participant's intercept and slope move closer to the mean of the intercept and slope, which is called shrinkage. The advantage of shrinkage is that it compensates for extreme values, especially in case of missing or sparse data, reducing errors in estimating the magnitude of the effects. Finally, the residual variance ϵ_{ij} describes the difference between each actual data point and the model predictions for that data point.

That is, ϵ_{ij} reflects the trial-to-trial variability within participants.

How are the different dimensions of variability (persons, measures, occasions) analyzed at the same time in a linear model? Each manipulated variable can be included as a separate slope β with a slope adjustment for participants u_i. In this thesis, systematic variability between *persons* is estimated by a slope for PARTICIPANT GROUP, which is the difference between language-impaired and language-unimpaired individuals. Furthermore, there could be systematic differences between persons within the group of IWA. Therefore, additional slopes for AGE, YEARS OF EDUCATION, YEARS POST STROKE, WORKING MEMORY SCORES, and APHASIA SYNDROME will be estimated. The systematic variability in *measures* and *occasions* are estimated by slopes for TASK (object manipulation, sentence-picture matching) and for PHASE (test, retest). The slope adjustments of each participant for the tasks' slope and phases' slope will be used to examine the between-participant variability in the differences between test phases and tasks. Finally, the unsystematic variability within each participant will be included in ϵ_{ij}.

In sum, there are different methods to identify and quantify variability. This thesis uses a quantifying method to gain a differentiated picture of the variability in sentence processing in aphasia. Among the quantifying methods, linear modeling offers several advantages over the COV: The data do not have to be adjusted, and therefore no variability is lost, the different variance components are calculated simultaneously and are therefore not mixed in one value, and finally, the average performance is taken into account in the estimation of the adjustments of the effects, resulting in robust estimates for each participant. Due to its advantages, this thesis uses linear modeling to quantify variability. In particular, emphasis will be placed on the slopes and slope adjustments to investigate the different dimensions of variability (persons, measures, occasions).

Based on the work of Caplan et al. (Caplan, 2012; Caplan et al., 2013a; Caplan et al., 2007; Varkanitsa & Caplan, 2018), this thesis focuses on the variability in syntactic processing rather than on the variability in raw scores. As mentioned in Chapter 2, Caplan et al. study syntactic processing by comparing structures that do (complex) or do not (simple) require a specific syntactic operation. For example, Caplan et al. (2007) considered object relative clauses to be syntactically complex because the syntactic operation of co-indexing traces and moved arguments is necessary to interpret them correctly. In comparison, Caplan et al. (2007) considered subject relative clauses to be syntactically simple because co-indexing is not required to interpret them correctly. The authors regard the comparison of syntactically simple versus complex sentences a prerequisite to studying syntactic processing in IWA (Varkanitsa & Caplan, 2018). For this comparison, the authors use minimal sentence pairs which have the same lexical material but differ syntactically, as in (1).

(1) a. *subject relative clause*

Hier ist der Esel, der$_{nom}$ den$_{acc}$ Tiger wäscht.
here is the donkey who$_{nom}$ the$_{acc}$ tiger washes
'Here is the donkey who washes the tiger.'

b. *object relative clause*

Hier ist der Esel, den$_{acc}$ der$_{nom}$ Tiger wäscht.
here is the donkey who$_{acc}$ the$_{nom}$ tiger washes
'Here is the donkey who the tiger washes.'

The two sentences in (1a) and (1b) differ only in the sequence of arguments in the relative clause. Differences in response measures (e.g., reaction time or accuracy) between (1a) and (1b) therefore cannot reflect difficulties in word processing. Instead, these differences must be attributed to difficulties in syntactic processing (Varkanitsa & Caplan, 2018). Based on this reasoning, this thesis focuses on the difference in response measures between syntactically simple and complex sentences in order to determine the variability in syntactic processing. Following Caplan et al. (Caplan et al., 2015, 2013a; Caplan et al., 2007), this difference is called *syntactic complexity effect*.

In the present work, the syntactic complexity effect will be represented by a slope β with a slope adjustment for participants u_i for simple versus complex sentences in the statistical model. To study variability in the syntactic complexity effect, two different approaches will be used. In a first model, the factor COMPLEXITY will be nested under PARTICIPANT GROUP, TASK and PHASE in order to gain separate estimates of the complexity effect for each group (control participants and IWA), task (object manipulation and sentence-picture matching) and test phase (test and retest). In a second model, the variability in complexity effects between test phases or tasks is represented in the interaction between COMPLEXITY and TASK or PHASE. These interactions will be estimated separately for control participants and IWA by nesting COMPLEXITY under PARTICIPANT GROUP.

An interaction between COMPLEXITY and TASK or PHASE would be a first indication of systematic between-task or between-session variability in sentence comprehension in the IWA or the control group. However, a single interaction would not be regarded as systematic variability since four different sentence structures will be analyzed at the same time (declaratives, relative clauses, control structures with a pronoun, and control structures with PRO). Therefore, an interaction in one sentence structure may occur accidentally due to the large number of predictors in the model. Thus, to be counted as systematic variability in sentence comprehension, interactions between COMPLEXITY and TASK or PHASE have to be present in at least two of the four sentence structures. Additionally, these interactions have to have the same sign (e.g., both interactions are positive). If they have a different sign, complexity effects both decrease and increase which would be an unsystematic pattern. Hence, the following outcomes would be regarded as unsystematic variability in the IWA or the control group: 1) no in-

teraction between COMPLEXITY and TASK or PHASE, 2) an interaction in only one sentence structure, or 3) a positive interaction in one sentence structure and a negative interaction in another sentence structure.

5 Research questions

The previous chapters lead to the following two broad research questions:

1. How does sentence comprehension performance in aphasia vary on the three dimensions, persons, measures, and occasions?

2. Can the resource reduction hypothesis explain the sentence comprehension performance in aphasia and especially the variability in the performance?

The first research question is motivated by the fact that no study on sentence comprehension in aphasia has simultaneously investigated the different dimensions of variability, i.e., persons, measures, and occasions, yet. Such a simultaneous investigation of the different dimensions of variability has the advantage that the participant group and sentence material remain the same across the different dimensions of variability, and thus, differences in variability across dimensions cannot be due to changes in participants or materials. Additionally, a comprehensive study of variability provides a large amount of data that leads to a more accurate statistical estimate and comprehensive overview of the variability in performance. The second research question is motivated by the fact that there is no consensus among researchers about the source of the variability in sentence comprehension in aphasia yet. The resource reduction hypothesis (Caplan, 2012), ascribes variability to random fluctuations in sentence processing resources due to noise. However, this assumption has not been tested by researchers other than Caplan and his colleagues yet. Furthermore, Caplan and his colleagues did not investigate all dimensions of variability. Therefore, the thesis puts the resource reduction hypothesis to the test.

To approach the two research questions, two studies were designed. Study 1 was dedicated to research question one. To investigate the three dimensions of variability, three sub-questions were formulated:

- Are there differences in variability between language-impaired and language-unimpaired participants?

- To what extent does the sentence comprehension performance vary between response tasks within and between IWA?

- To what extent does the sentence comprehension performance vary between test phases within and between IWA?

Study 2 was dedicated to research question two. Three predictions of the resource reduction hypothesis of Caplan (2012) were addressed:

- The syntactic knowledge of IWA is intact.

- The syntactic complexity of a sentence influences the success of sentence processing in IWA.
- Variability in sentence processing in aphasia is caused by random fluctuations in processing resources.

The investigation of the research questions was based on six experiments that were carried out with a group of 21 German-speaking IWA and a control group of 50 German-speaking language-unimpaired individuals. Three tasks were used to probe sentence comprehension, each performed at two test points, resulting in the total of six experiments. The three sentence comprehension tasks were object manipulation, sentence-picture matching with regular listening, and sentence-picture matching with self-paced listening. In the object manipulation task, a sentence was presented auditorily and acted out by the participants with figurines. In the two sentence-picture-matching tasks, a spoken sentence and two pictures were presented simultaneously, and the participants had to select the picture that matched the sentence. In the regular sentence-picture matching task, sentences were presented as a whole. This task was performed in the visual world paradigm to collect eye-movement data. In the self-paced sentence-picture matching task, sentences were presented phrase-by-phrase, and the participants controlled the phrases' presentation themselves by pressing keys. The tasks are described in further detail in Study 1, Chapter 2.2. As materials, the same 120 sentences were used in all experiments. Sentences belonged to one of four different sentence structures, namely declarative sentences, relative clauses, control structures with a pronoun, and control structures with PRO. Per sentence structure, the complexity of the sentences was manipulated leading to structurally simple and complex versions of each sentence structure. Detailed explanations of the sentence materials as well as sentence examples are provided in Study 1, Chapter 2.3.

In order to answer research question one, the offline response data of all six experiments were analyzed together and the slopes and slope adjustments were used as estimates for the variability in sentence processing (cf. Chapter 4). More specifically, a first model analyzed the syntactic complexity effect in the pooled response accuracies of the six experiments. A second model evaluated the syntactic complexity effect in the pooled reaction times of the four sentence-picture-matching experiments. Both models included slopes for participant group, task, and test phase. To investigate the between-participant variability, the differences in syntactic complexity effects between the participant groups and between the individual participants were compared. To study the between-task variability, the differences in syntactic complexity effects between the tasks were focused in each participant group and within each individual participant. Finally, the differences in syntactic complexity effects between test phases in each participant group and within each individual participant were evaluated to examine between-

session variability.

In order to answer research question two, the eye-tracking data of the regular sentence-picture-matching task were analyzed. More specifically, the predictions of the resource reduction hypothesis were used to derive expectations about the fixation behavior in visual world eye-tracking, and these expectations were compared with the actual eye-tracking results. To investigate the prediction that the syntactic knowledge in IWA is intact, the eye-tracking data of the IWA and the control participants and the IWA's eye-tracking data in correct and incorrect trials were compared. To evaluate the prediction that syntactic complexity influences the success of processing, the eye-tracking data of the structurally simple and complex versions of each of the four sentence structures were compared. Finally, the eye-tracking data of the test and the retest were compared to test the prediction that sentence processing in IWA varies randomly.

6 Summary of the results of Study 1 and 2

In this chapter, the results concerning the questions on variability in sentence comprehension in aphasia that were posed in Chapter 5 are presented in an overview. The detailed results can be found in Study 1, Chapter 3 and Study 2, Chapter 3. First, the question of how variable sentence comprehension performance is in IWA is addressed. Then, the question of whether the resource reduction hypothesis can explain the processing patterns of IWA is considered.

6.1 Summary of Study 1

The first research question was answered with the help of the offline data from the three sentence comprehension tasks. These data included the response time for the picture selection in the two sentence-picture matching tasks and the response accuracy in picture or figurine selection in all three tasks. Overall, the control participants responded faster and more accurately than the IWA. Furthermore, both participant groups showed faster reaction times in the retest than in the test, and the number of correct responses increased in the retest compared to the test. There was no interaction between the participant group and the test phase in reaction times. In contrast, the increase in response accuracy was more prominent in the IWA than in the control group.

As explained in Chapter 4, the variability in syntactic processing was not assessed based on raw reaction time and response accuracy. Instead, the variability was assessed based on the syntactic complexity effect, i.e., the difference in reaction time or response accuracy between syntactically simple and complex sentences. As a starting point for examining variability, it was first shown that a complexity effect occurs in each of the four sentence structures. Then, based on this proof, the extent to which complexity effects are variable was evaluated. The results are presented below split up by the three dimensions of variability (persons, measures, and occasions).

6.1.1 Between-participant variability

There was variability in sentence processing between the IWA which manifested itself by varying complexity effects between IWA. While most variability occurred in the size of complexity effects, in some extreme cases, individuals exhibited a negative complexity effect, i.e., they had more difficulties understanding the structurally simpler sentence than the structurally complex sentence. The variability in complexity effects between IWA could not be attributed to differences in age, years of education, years post-onset, working memory scores, and aphasia type of the Aachen Aphasia Test (AAT, Huber et al., 1983). The variability in the complexity effect was more pronounced between IWA than between control participants. In the control group, negative complexity effects did

not occur with one exception.

6.1.2 Between-task variability

Regarding the variability of sentence processing between different tasks, the findings were as follows: Overall, complexity effects occurred in object manipulation and the two variants of sentence-picture matching. This result suggests that complexity effects in IWA are measurable across different tasks. There were signs of variability in sentence processing between tasks, with complexity effects being more pronounced in one task than in another. Task demands could not explain this variability. That is, complexity effects were not systematically stronger in one task than in the other tasks across the different sentence structures or test phases in IWA. Finally, the between-task variability in IWA was compared with the between-task variability in control participants. In contrast to the IWA, the control group showed a systematic difference in the complexity effect between the two variants of the sentence-picture-matching task. This systematic difference was evident from the interaction between task and complexity effect at the group level. Furthermore, the difference in complexity effect was equally evident across 33 of the 50 control participants. These participants showed a difference in complexity effect between the same tasks, i.e., all of them exhibited more pronounced complexity effects in sentence-picture matching under regular listening than under self-paced listening.

6.1.3 Between-session variability

The following observations were made with respect to the variability of sentence processing between two test phases: Complexity effects occurred in the IWA at each task's first and second performance. However, complexity effects did not systematically increase or decrease in the retest. Thus, it seems that complexity effects persist in IWA under repeated exposure (although it cannot be ruled out that complexity effects change in IWA with more repetitions). There were signs of variability in sentence processing between the two test phases, with complexity effects being more pronounced in IWA at one test phase than at the other. This variability was not due to the test phase, i.e., complexity effects were not systematically stronger at the test or the retest across different sentence structures or tasks in IWA. Finally, the between-session variability in IWA was compared with the between-session variability in control participants. Unlike the IWA, there was a systematic difference in the control group's complexity effects between the two test phases. This systematic difference was apparent from the interaction between the test phase and the complexity effect at the group level, which occurred in three of the four investigated sentence structures.

6.1.4 Conclusions

The following conclusions can be drawn based on the results on variability within and between IWA, between tasks and test phases, and in comparison to the control group. First, relatively stable performance patterns emerged concerning complexity effects: Complexity effects occurred in both participant groups, in all three tasks, and both test phases. At the individual participant level, negative complexity effects (i.e., more difficulties in understanding the syntactically easy versus complex structure) occurred only rarely. Thus, sentence processing seems to be stable in the sense that syntactic complexity affects all investigated dimensions, i.e., persons, measures, and occasions. While the *occurrence* of complexity effects is stable, the *strength* of complexity effects is variable. Variability in complexity effects between tasks and test phases occurred in both participant groups. However, there were differences in variability between the two participant groups. In the IWA, changes in the size of the complexity effect were not due to differences in task or test phase. In the control group, on the other hand, both the task and the test phase influenced the size of the complexity effects. As explained in Chapter 3, unsystematic and systematic variability can be distinguished. Comparing the results of the IWA and the control group shows that variability takes different forms in the two participant groups. Variability in sentence processing in language-unimpaired control participants appears to be systematic as it can be explained by differences in task or test phase. In contrast, variability in sentence processing in IWA appears to be unsystematic. These findings are presented and discussed in detail in Study 1, Chapter 3 and 4.

6.2 Summary of Study 2

The unsystematic variability in sentence processing in aphasia observed in Study 1 is consistent with Caplan's (2012) resource reduction hypothesis. Therefore, this hypothesis was chosen to investigate sentence processing and its variability further. For this purpose, visual world eye-tracking data were analyzed. The proportions of fixations to a target picture were used as an index for sentence processing of the IWA. More specifically, three predictions were derived from the resource reduction hypothesis and tested using data from visual world eye-tracking. The predictions and results are summarized below.

6.2.1 Prediction 1: Normal-like processing

According to the resource reduction hypothesis, syntactic knowledge is intact in IWA. Based on this assumption, it was predicted that IWA should show a normal-like fixation pattern in correctly answered trials. The results were consistent with this prediction. First, the normal-like fixation pattern was evident in the steady increase in looks to the target picture throughout each trial, indicating that IWA preferred the target picture over

the distractor picture. Furthermore, the normal-like fixation pattern was evident in the difference in looks to the target picture between correct and incorrect trials, indicating that processing was different in the correct and incorrect trials. Finally, comparing the looks to the target picture in the IWA and the control group revealed a slower increase and lower maximum amplitude of target fixations in IWA. A slowed increase in target fixations is consistent with the resource reduction hypothesis, given the assumption that the reduced resource in IWA is processing speed. The lower maximum amplitude of target fixations does not contradict the assumption that syntactic knowledge is intact but could suggest that there are differences in sentence comprehension between IWA and control participants. The lower maximum amplitude could reflect that IWA are less certain in their final sentence interpretation than control participants. One reason for the uncertainty could be, e.g., that IWA are aware of their occasional incorrect sentence comprehension thus making them hesitant to decide for a sentence interpretation.

6.2.2 Prediction 2: Structural complexity effect

The resource reduction hypothesis further states that processing should be more difficult in syntactically complex versus simple sentences. Based on this assumption, it was predicted that IWA should fixate the target picture less in complex sentences than in simple sentences. This prediction was not confirmed. Instead, IWA gave more incorrect responses in complex versus simple sentences, but incorrect trials were excluded in the eye-tracking analysis. In the correct trials, IWA did not show fewer looks to the target picture in complex versus simple sentences. This result can be explained with the resource reduction hypothesis since processing should be intact in trials that are answered correctly (regardless of syntactic complexity). However, the IWA's pattern of target fixations differed from the control group. The control participants looked at the target picture less frequently in complex versus simple sentences. One possible explanation for the differences in fixation patterns in the two participant groups is as follows. Although syntactic complexity initially leads to processing difficulties in both participant groups, the groups deal differently with the difficulties. Control participants can overcome the processing difficulties, which is why they show fewer target fixations in complex versus simple sentences overall, but ultimately arrive at a correct sentence interpretation. However, IWA cannot overcome the processing difficulties and arrive at an incorrect sentence interpretation.

6.2.3 Prediction 3: Random variability in the performance

Finally, the resource reduction hypothesis assumes that variability in sentence processing in aphasia reflects random noise. Under this assumption, the looks to the target picture of the IWA should not differ between test and retest. However, if there were

systematic differences, e.g., a practice effect, target fixations should differ between test and retest. The results favor the resource reduction hypothesis because the target fixations of the IWA did not change systematically between the test and retest in all but one sentence structure. In complex declarative sentences, target fixations increased slower in the retest than in the test. A slowed increase in target fixations in the retest suggests persistent sentence comprehension difficulties in IWA. No notable changes in looks to the target picture occurred between the test and retest in the control group.

6.2.4 Conclusions

Overall, the eye-tracking data of the IWA and the comparison with the control group lead to the following conclusions. The predictions of the resource reduction hypothesis were mostly confirmed. The data are consistent with the predictions that syntactic knowledge is intact in aphasia, that syntactic complexity leads to processing difficulties, and that variability in performance originates from random noise. However, contrary to the predicted normal-like processing, the IWA's sentence processing differed from the control group's processing in the correct trials. First, the maximum amplitude of target fixations in correct trials was reduced in the IWA as compared to the control group. This reduced amplitude suggests that IWA are less sure about the picture selection than control participants. Second, in contrast to control participants, IWA did not show a difference in target picture fixations between complex and simple sentences in correct trials, and an increased number of errors in syntactically complex sentences. This pattern might suggest that IWA struggle with revising sentence interpretations. Under this assumption, the lack of differences in target fixations between complex and simple sentences and the increased number of errors in complex sentences could be explained in the following way. Possibly, IWA can sometimes form a prediction about sentence structure before the actual syntactic structure is disambiguated in the input. If this happens, IWA cannot revise their prediction once they get the structural information in the input. However, due to slow processing, IWA do not always form a prediction before hearing the relevant structural information in the input. If the input comes in before the prediction is ready, IWA can use the structural information to form a correct sentence interpretation. The slowdown in making predictions and the impairment in revising predictions could explain why the performance is near chance in complex sentences. A final striking feature in the fixation pattern of the IWA is the slower increase in target fixations in complex declarative sentences in the retest versus the test phase. This slowdown in the retest is surprising if one assumes that processing becomes more efficient with repetition, which should lead to faster target fixations in the retest. The slowdown in the retest could therefore suggest that IWA have difficulties adapting processing to the syntactic structures in the input. These findings are presented and discussed in detail in Study 2, Chapter 3 and 4.

6.3 Variability in Studies 1 and 2

The data presented in Studies 1 and 2 lead to the following conclusions with respect to variability. First, both studies suggest that IWA exhibit unsystematic variability in sentence processing. No differences in complexity effects are apparent between tasks or test phases in the reaction times and accuracy, and between test phases in the eye-tracking data. In contrast to the IWA, the control group shows systematic variability in sentence processing in response times and accuracy, since complexity effects systematically differ between tasks and test phases. However, the control group does not show systematic variability in the eye-tracking data since there is no difference in the target fixations between the test and retest. Thus, there is a discrepancy between the online and offline data of the control group.

The following section presents an analysis of a different part of the data set that has not been considered in Studies 1 and 2. The analysis was carried out to clarify more precisely which processing step varies systematically in the control group and to replicate the finding of unsystematic variability in IWA.

7 Study 3: Variability in self-paced listening

This chapter analyzes data obtained from the self-paced sentence-picture-matching task to gain further insights into variability in aphasia. In the self-paced listening experiment, sentences were presented auditorily phrase-by-phrase, and the participants controlled the duration of the presentation of the phrases via key press. In the remainder of this text, the presentation time for the phrases, i.e., the time from the auditory onset of a phrase until a button was pressed by the participant, will be called *listening time* to distinguish it from the reaction times for the picture selection at the end of the sentence. The analysis presented here will be limited to the listening time. The focus is set on this measure because the response time and accuracy for picture selection in sentence-picture matching has already been evaluated in Study 1. There are two ways in which variability will be investigated. First, within-session variability is examined by analyzing how listening times vary between trials. Second, between-session variability is examined by analyzing how listening times vary between the two test phases. Before coming to the analysis, the following section explains why it is worthwhile to evaluate the listening time in addition to the offline data (Study 1) and the eye-tracking data (Study 2).

7.1 Motivation Study 3

The results of the IWA obtained in Studies 1 and 2 led to the conclusion that variability in sentence processing in aphasia is unsystematic. If this conclusion is correct, then the variability in the listening times of the IWA should also be unsystematic. Thus, the listening times may help to support the conclusions of Studies 1 and 2. The control group's results obtained in Studies 1 and 2 are mixed. There was systematic variability in reaction times and response accuracy between tasks and test phases. However, the target fixations of the control participants in eye-tracking varied unsystematically between the test and retest. Therefore, it is an open question what form the variability in the listening times of the control participants will take. Answering this question is essential to determine which sentence comprehension processes (e.g., online or offline processes) vary systematically or unsystematically in control participants. Understanding which processes vary systematically or unsystematically in control participants is the basis for determining how sentence processing differs in individuals with and without aphasia. The following section explains what conclusions may be drawn about the variability in sentence processing when the listening times of control participants vary systematically or unsystematically.

The first possibility is that there is systematic variability in the listening times of the control group. In this case, the variability in listening times would be similar to the offline data (which also varied systematically) and different from the eye-tracking data (which varied unsystematically). In this case, the difference in variability could

be attributed to the way the response is given: Reaction times and response accuracy are based on a conscious decision to press a key. Listening times also involve a conscious decision to press a key. In contrast, target fixations in visual world eye-tracking are not subject to a conscious decision process and occur automatically (Dickey et al., 2007). Thus, the response in self-paced listening and the comprehension task occurs with a higher degree of awareness than the response in eye-tracking. If the listening times vary systematically, then the difference in variability between measures could be attributed to unconscious processes varying unsystematically and conscious processes varying systematically.

The second possibility is that there is unsystematic variability in the listening times of the control group. In this case, the variability in listening times would be similar to the eye-tracking data (which also varied unsystematically) and different from the offline data (which varied systematically). In this case, the difference in variability could be attributed to the timing of measurement collection: Reaction times and response accuracy are collected at the end of the sentence. Therefore, they reflect late processing stages, such as decision-making and response preparation (Caplan et al., 2013, Stowe et al., 2018). In contrast, listening times and eye-tracking data are collected during sentence presentation. Therefore, they provide insights into online sentence processing. If the listening times vary unsystematically, then the difference in variability between measures could be attributed to offline processes varying systematically and online processes varying unsystematically.

To summarize, the self-paced listening data of the control group can provide the following insights. If the listening times vary systematically, this could indicate that conscious and unconscious processing stages differ in variability. If the listening times vary unsystematically, this could indicate that online and offline processing differs in variability. Depending on the control group's results, conclusions can be drawn about which which sentence comprehension processes differ between IWA and control participants.

7.2 Methods Study 3

Participant group, sentence material, and details regarding the experiment's procedure are explained in detail in Study 1, Chapter 2. Therefore, only a summary is given. The analysis includes data from 21 IWA (mean age = 60.2, range = 38-78 years, 1-26 years post onset) and 50 control participants (mean age = 48, range = 19-83 years), all native speakers of German. Sentence comprehension was assessed by auditory sentence-picture matching with two pictures (target and foil). The material included four sentence structures: declarative sentences, relative clauses, and control structures with a pronoun or PRO. Half of the sentences in each sentence structure were syntactically simple and

half complex. A total of 120 sentences were tested, i.e., 60 syntactically simple and 60 syntactically complex sentences. The sentences were presented phrase-by-phrase, and the participants controlled the presentation rate. The task was performed at two time points spaced approximately two months apart.

Data analysis was performed on the listening times in the correct trials. Listening times exceeding a duration of 10 seconds were discarded, resulting in a loss of 26 observations (0.15% of the data). The analysis focused on listening times in the critical sentence region. The critical region was the part of the sentence in which syntactic processing should be more difficult for syntactically complex than simple sentences. This region was studied because it also was analyzed in the eye-tracking experiment. Furthermore, this region was selected to study the variability in complexity effects rather than variability in listening times per se. As explained in Chapter 4, variability in complexity effects was evaluated to gain insights into syntactic processing. The critical region of each sentence structure is highlighted in bold in Table 1. The critical region in declarative sentences and relative clauses is the part of the sentence where canonical and non-canonical word orders can be distinguished. Therefore, the first nominal phrase was examined in declarative sentences, and the relative pronoun was studied in relative clauses. The critical region in control structures with a pronoun or PRO is the anaphor in the subordinate clause, where participants have to establish a coreference with the main clause subject or object. Since this anaphor is very short in the case of an overt pronoun and is not overt in the case of PRO, the listening times at the nominal phrase directly following the anaphor were evaluated.

Data were analyzed using Bayesian hierarchical linear models with correlated random intercepts and slopes for participants and items using R (Version 3.6.3; R Core Team, 2020) and the R-package brms (Version 2.13.0; Bürkner, 2017, 2018). Listening times were log-transformed and estimates were backtransformed to milliseconds for the ease of interpretation. The predictors included test phase, trial number, participant group, and sentence types nested under participant group. Further predictors were the interaction of test phase and participant group and the sentence types nested under participant group and the interaction of trial number and participant group and the sentence types nested under participant group. For almost all predictors, sum contrasts were used. The only exceptions were the relative clause subtypes, for which sliding contrasts were used, and trial number, which was included in the model as a continuous predictor centered around the mean trial number. The priors of the model were mildly uninformative. The priors for the fixed effects intercepts were set to $\mathit{Normal}(0, 10)$, the priors for the fixed effects slopes to $\mathit{Normal}(0, 1)$, the priors for the correlations to $\mathit{LKJ}(2)$, and the prior standard deviations of the random effects and the residual error to $\mathit{Normal}(0, 1)$ truncated in zero. The mean and the 95% CrI of the estimated effects are reported below.

Table 1: Example of the declaratives, relative clauses, and control structures with PRO or an overt pronoun used in the experiment.

Sentence type	Condition	Sentence
Declaratives	SO	Hier tröstet **der**$_{NOM}$ **Tiger** gerade den$_{ACC}$ Esel Here the$_{NOM}$ tiger just comforts the$_{ACC}$ donkey
	OS	Hier tröstet **den**$_{ACC}$ **Tiger** gerade der$_{NOM}$ Esel Here the$_{ACC}$ tiger just comforts the$_{NOM}$ donkey
Relative clause	SRC	Hier ist der Tiger **der**$_{NOM}$ den$_{ACC}$ Esel gerade tröstet Here is the tiger who$_{NOM}$ comforts the$_{ACC}$ donkey
	ORC	Hier ist der Tiger **den**$_{ACC}$ der$_{NOM}$ Esel gerade tröstet Here is the tiger who$_{ACC}$ the$_{NOM}$ donkey comforts
Control, PRO	s-ctrl	Peter$_i$ verspricht nun Lisa$_j$ PRO$_i$ **das kleine Lamm** zu streicheln und zu kraulen. Peter$_i$ promises now Lisa$_j$ PRO$_i$ to pet and to ruffle the little lamb.
	o-ctrl	Peter$_i$ erlaubt nun Lisa$_j$ PRO$_j$ **das kleine Lamm** zu streicheln und zu kraulen. Peter$_i$ allows now Lisa$_j$ PRO$_j$ to pet and to ruffle the little lamb.
Control, pronoun	match	Peter$_i$ verspricht nun Thomas$_{MASC}$, dass er$_i$ **das kleine Lamm** streichelt und krault. Peter$_i$ promises now Thomas$_{MASC}$ that he$_i$ will pet and ruffle the little lamb.
	mismatch	Peter$_i$ verspricht nun Lisa$_{FEM}$, dass er$_i$ **das kleine Lamm** streichelt und krault. Peter$_i$ promises now Lisa$_{FEM}$ that he$_i$ will pet and ruffle the little lamb.

Note. S = subject O = object, SRC/ORC = subject/object relative clause. s-ctrl/o-ctrl = subject/object control, match/mismatch = gender match or mismatch of the main clause nouns. Critical region in bold.

7.3 Results Study 3

Listening times for the sentences are shown in Figure 2. Control participants had shorter listening times than IWA (-502 ms CrI: [-677, -339]). Listening times were shorter in the retest than in the test (-55 ms CrI: [-108, -3]), with no interaction between test phase and participant group (11 ms CrI: [-40, 63]). Similarly, listening times became shorter across trials. Between two adjacent trials, reaction times decreased by an average of 2ms (between the middle trial and the trial before it -2 ms CrI: [-2, -1]), with no interaction between trial number and participant group (1 ms CrI: [-1, 2]).

The estimates for the complexity effects in the four sentence structures are shown in Figure 3 A for the control group and in Figure 3 B for the IWA. In the control group, complexity effects occurred in declarative sentences and relative clauses (declaratives: -116 ms CrI: [-164, -72], relative clauses: -44 ms CrI: [-61, -28], control structures

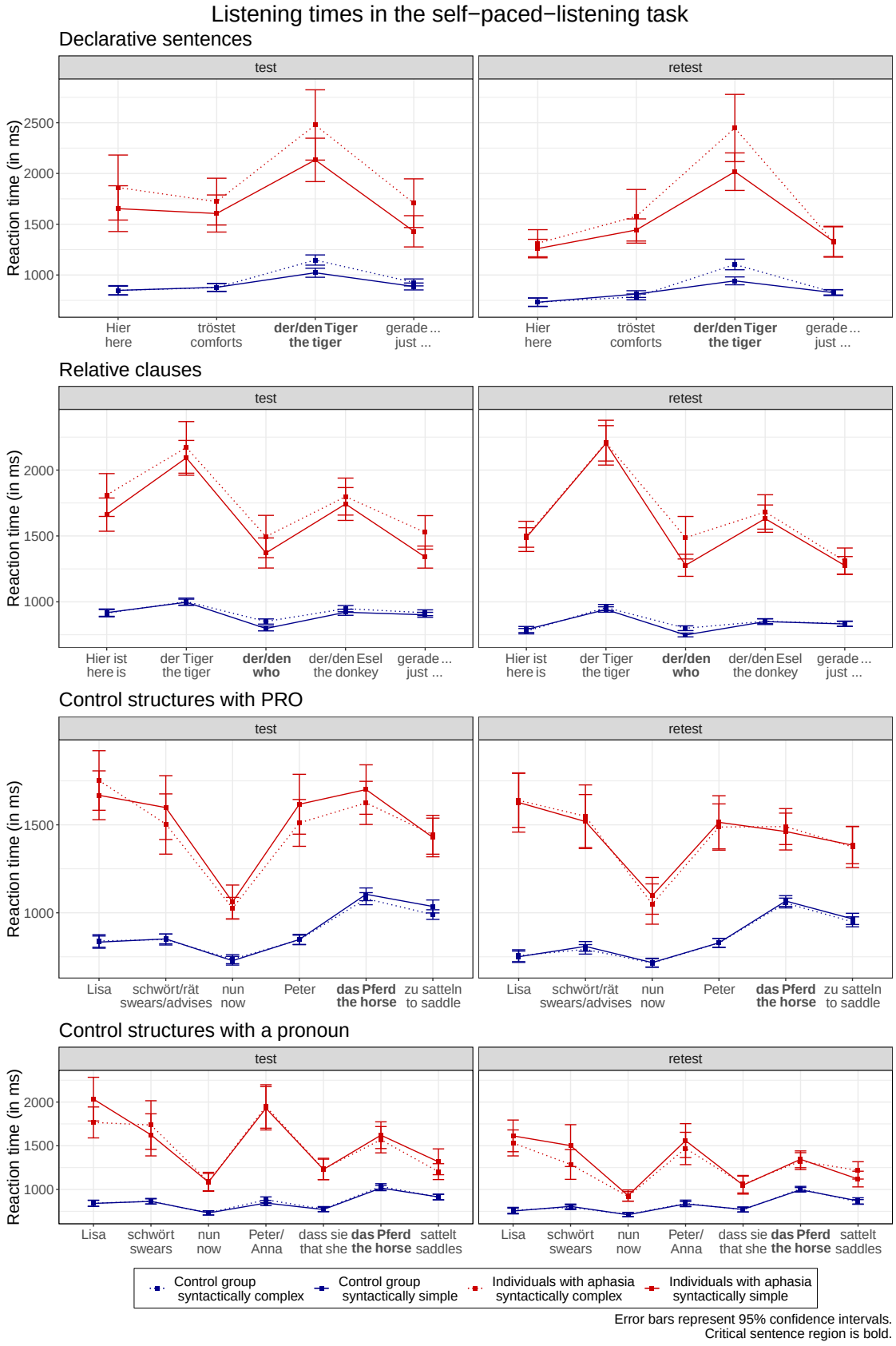

Figure 2: Mean listening times and 95% CI in syntactically simple (solid) and complex (dotted) versions of the four investigated sentence structures for the control group (blue) and the individuals with aphasia (red). The critical sentence region is highlighted in bold.

with PRO: 1 ms CrI: [-32, 34], control structures with a pronoun: 12 ms CrI: [-18, 42]). In the IWA, complexity effects occurred in declarative sentences, relative clauses, and control structures with a pronoun (declaratives: -248 ms CrI: [-561, 58], relative clauses: -63 ms CrI: [-139, 9], control structures with PRO: 56 ms CrI: [-84, 196], control structures with a pronoun: -73 ms CrI: [-163, 14]).

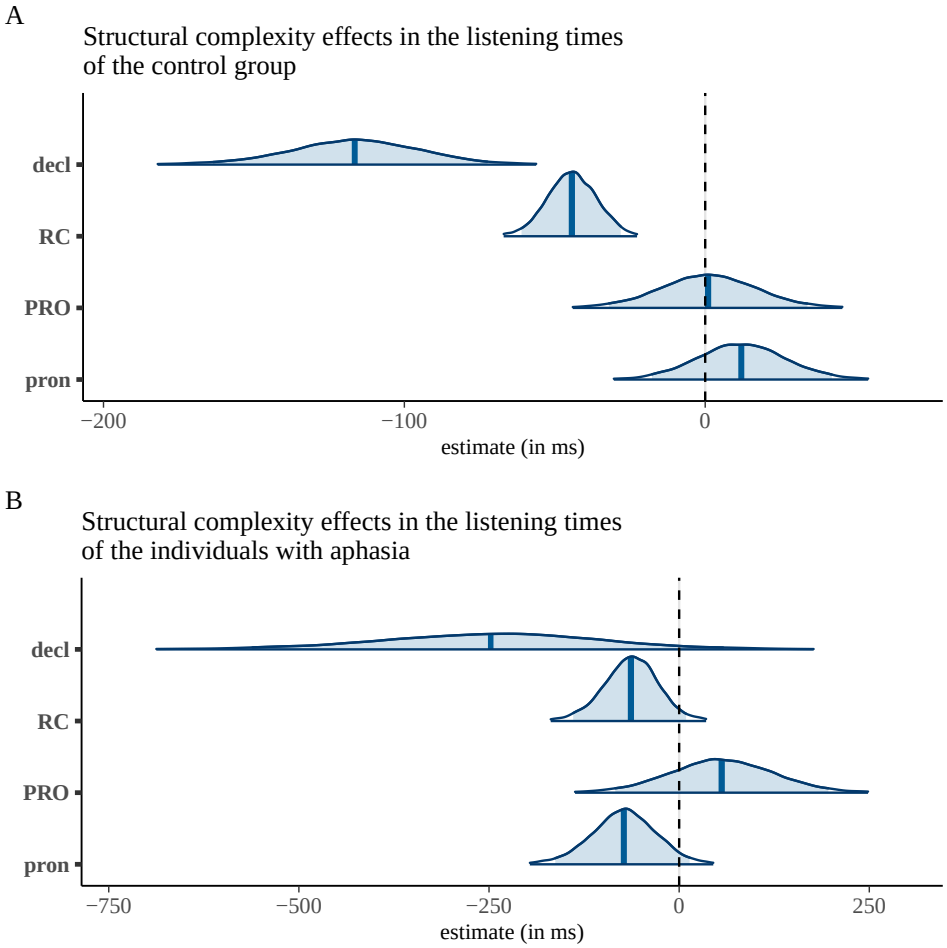

Figure 3: Structural complexity effects in declarative sentences (decl), relative clauses (RC) and control structures with a pronoun (pron) or PRO (PRO) in the control group (A) and the in dividuals with aphasia (B) for the pooled data of both test phases. Plots display the posterior estimates of the effects with 95% CrIs. The dashed line represents an effect size of zero. Distributions that are left-shifted denote faster listening times in the syntactically simple version of the sentence structure.

The estimates for the within-session changes in complexity effects between adjacent trials are shown in Figure 4 A for the control group and Figure 4 B for the IWA. In the control group, no changes in complexity effects occurred between trials (declaratives: 0 ms CrI: [-2, 1], relative clauses: 0 ms CrI: [0, 1], control structures with PRO: 1 ms CrI: [-1, 3], control structures with a pronoun: 0 ms CrI: [-1, 2]). In the IWA, changes in the complexity effect occurred in one of the four sentence structures. More precisely, the complexity effect increased by an average of 5 ms between two adjacent trials in the

control structures with a pronoun (-5 ms CrI: [-10, -1]). In the remaining three sentence structures, no changes in complexity effects occurred between trials (declaratives: -2 ms CrI: [-5, 1], relative clauses: -1 ms CrI: [-2, 1], control structures with PRO: 3 ms CrI: [-1, 7]).

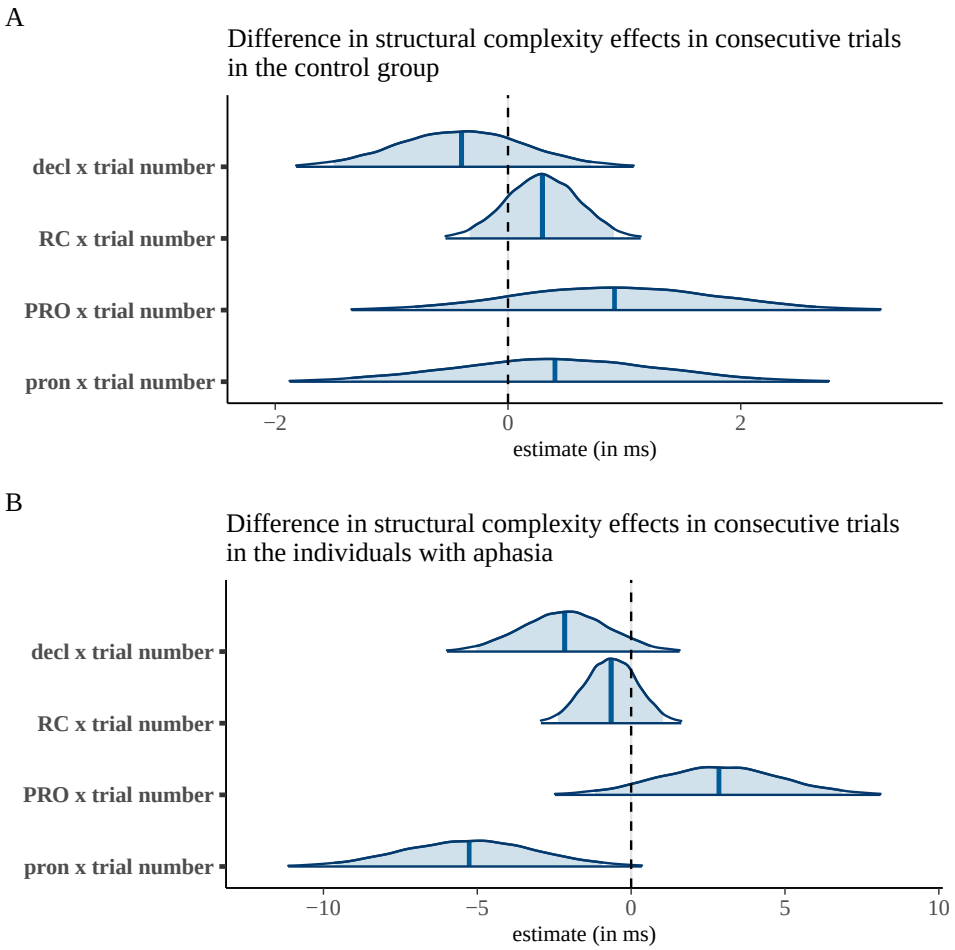

Figure 4: Within-session changes in structural complexity effects between consecutive trials in declarative sentences (decl), relative clauses (RC) and control structures with a pronoun (pron) or PRO (PRO) in the control group (A) and the in dividuals with aphasia (B). Plots display the posterior probabilities of the effects with 95% CrIs. The dashed line represents an effect size of zero. Distributions that are left-shifted denote larger complexity effects in a trial compared to the previous trial.

The estimates for the between-session changes in complexity effects between test and retest are shown in Figure 5 A for the control group and in Figure 5 B for the IWA. In the control group, no changes occurred in complexity effects between test and retest (declaratives: -14 ms CrI: [-39, 11], relative clauses: 3 ms CrI: [-11, 17], control structures with PRO: 9 ms CrI: [-17, 34], control structures with a pronoun: -7 ms CrI: [-32, 19]). In the IWA, changes in the complexity effect occurred in one of the four sentence structures. More precisely, in the relative clauses the complexity effect increased by an average of 54 ms in the retest (-54 ms CrI: [-95, -14]). In the remaining three sentence structures, no

changes in complexity effects occurred between test and retest in the IWA (declaratives: -23 ms CrI: [-91, 42], control structures with PRO: 13 ms CrI: [-46, 72], control structures with a pronoun: 26 ms CrI: [-33, 85]).

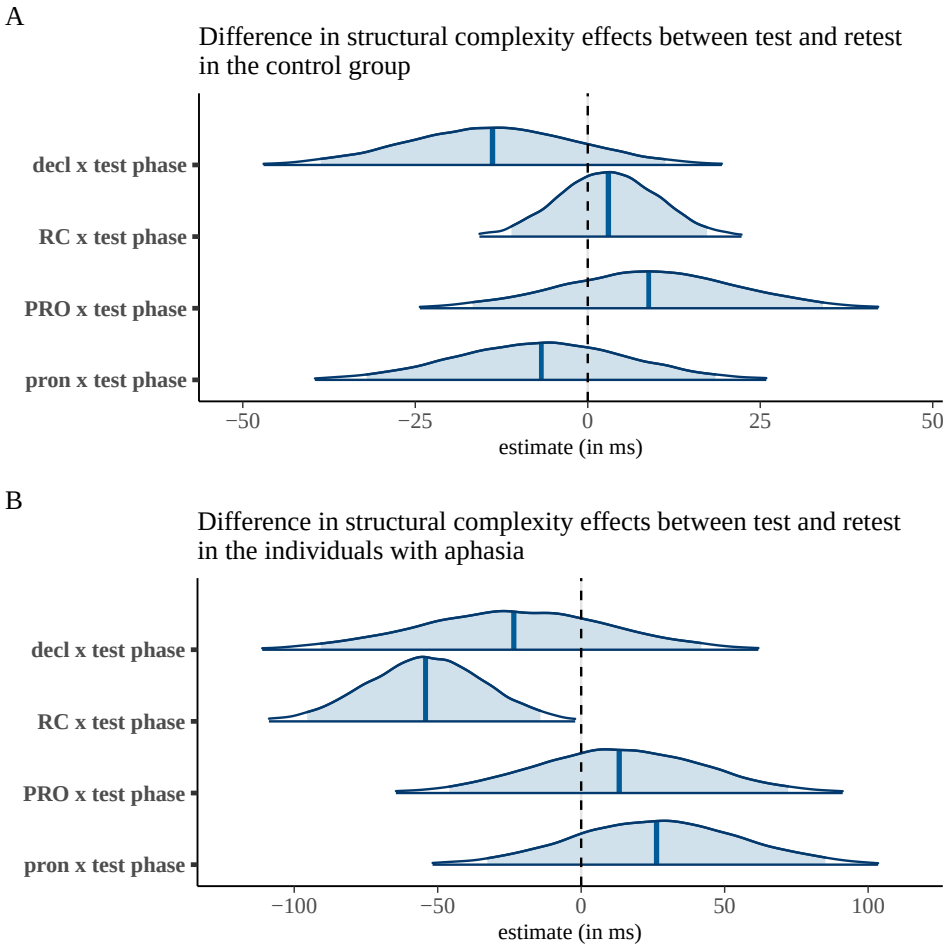

Figure 5: Between-session changes in structural complexity effects between test and retest in declarative sentences (decl), relative clauses (RC) and control structures with a pronoun (pron) or PRO (PRO) in the control group (A) and the in dividuals with aphasia (B). Plots display the posterior estimates of the effects with 95% CrIs. The dashed line represents an effect size of zero. Distributions that are left-shifted denote larger complexity effects in the retest in comparison to the test.

The previous results suggest hardly any signs for systematic differences within and between the test sessions. To further strengthen this result, a Bayes factor analysis was performed. The probability of an alternative model (M1) was compared with the probability of a null model (M0). To investigate whether systematic differences occur within and between sessions, the interaction between the complexity effect and trial number or the interaction between the complexity effect and test phase was included in M1 and omitted in M0. Because the priors can affect the magnitude of the Bayes factors, a range of Bayes factors was calculated with increasingly informative priors for the interaction. The following reasoning was used to determine the prior SD of the interaction:

In the self-paced listening experiment of Caplan et al. (2015), the SD for the listening times' main effects and interactions at the critical region of the sentence is between 0.18 and 0.54. That is, SDs between 0.1 and 0.5 are well-calibrated, i.e., agnostic regarding the direction of the effect but relatively informative regarding the size of the possible effect (Nicenboim et al., 2020). Based on this reasoning, the following increasingly informative priors were used for the interaction: *Normal*(0, *SD*) with *SD* 2, 1.5, 1, 0.5, 0.3, 0.1. The Bayes factor was calculated using bridge sampling with four chains and 40,000 iterations, 2000 of which were the warm-up phase. The calculated Bayes factor (BF_{01}) indicates how much evidence there is in favor of the null model compared to the alternative model. The following guideline from Jeffreys (1961), as cited in Lee & Wagenmakers (2014), was used to interpret the Bayes factor:

- BF01 > 100: Extreme evidence for M0
- BF01 = 30-100: Very strong evidence for M0
- BF01 = 10-30: Strong evidence for M0
- BF01 = 3-10: Moderate evidence for M0
- BF01 = 1-3: Anecdotal evidence for M0

Figure 6 shows the Bayes factors for the control group (upper part) and the IWA (lower part). Bayes factors were calculated only for sentence structures in which a complexity effect had occurred (i.e., control group: declarative sentences and relative clauses, IWA: declarative sentences, relative clauses, and control structures with a pronoun). No Bayes factors were calculated for the remaining sentence structures because, without a complexity effect, no systematic change in the complexity effect is possible. Figure 6 shows that in the control group for well-calibrated priors (*Normal*(0, *SD*) with *SD* between 0.1 and 0.5) for both declarative sentences and relative clauses the evidence against systematic change in complexity effect is strong to extreme. For IWA, the evidence for the null model is somewhat less strong than for the control group. The evidence against systematic changes in the complexity effect is between moderate and extreme for the IWA for well-calibrated priors (*Normal*(0, *SD*) with *SD* between 0.1 and 0.5).

7.4 Discussion Study 3

The listening times from self-paced listening were evaluated in order to test for systematic variability in syntactic processing between adjacent trials within a session and between the test phases. First, the raw listening times were considered. Across participant groups, listening times decreased both within and between sessions. Furthermore, the effect did not differ between control participants and IWA as there was no indication of an interaction between the decrease in listening times and participant group. Therefore, the results suggest systematic variability in both controls and IWA. However, it

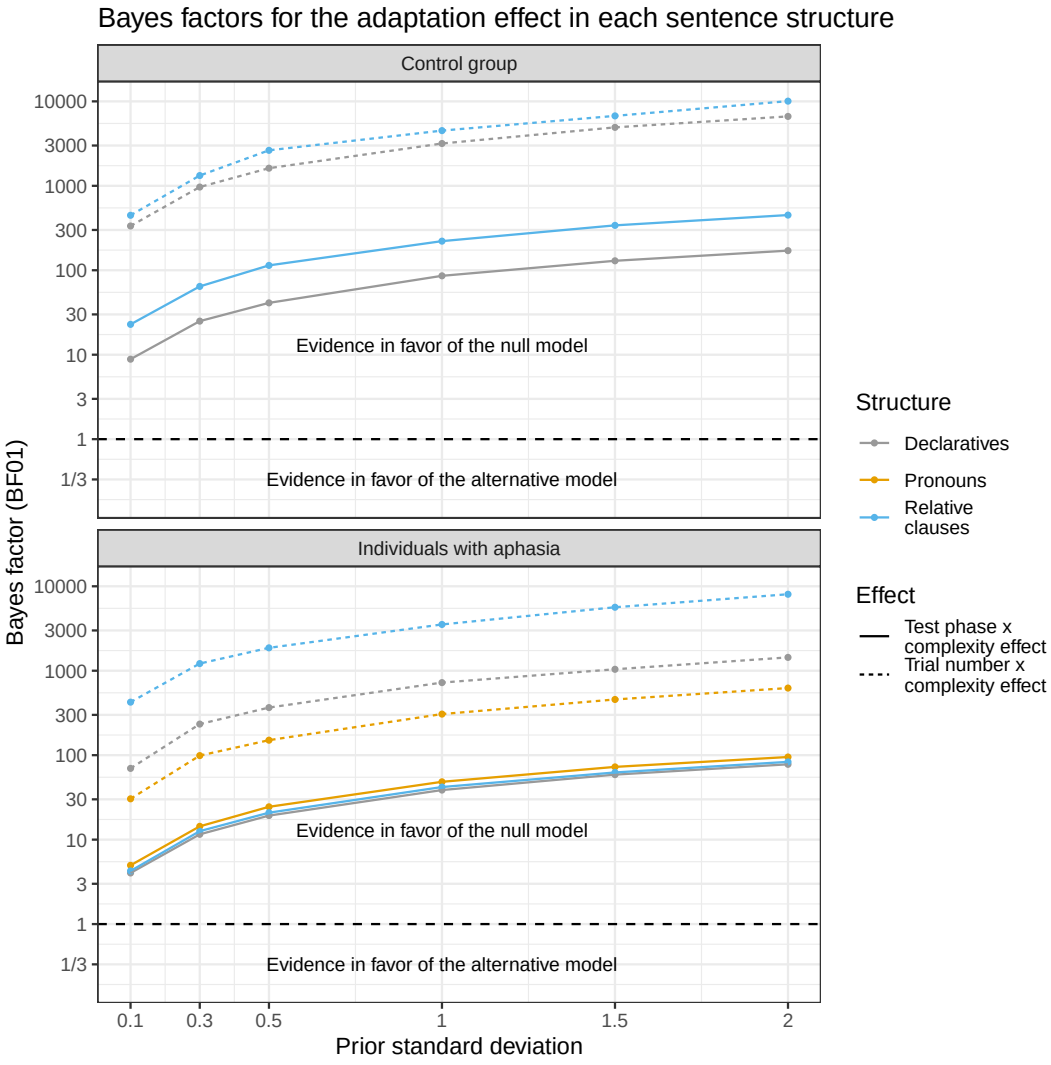

Figure 6: Bayes factors in favor of the null model compared to the alternative model (BF01) for different standard deviations for the prior for the adaptation effect across test phases and trials respectively.

is unclear whether this systematic variability was caused by changes in syntactic processing or improvements in task performance. Therefore, the listening time differences between syntactically complex and simple sentences were considered to clarify whether any systematic variability in syntactic processing occurred.

The control group displayed faster responses in declarative sentences and relative clauses with canonical versus non-canonical word order and no additional complexity effects in the control structures. For the structures in which complexity effects occurred, it was investigated whether syntactic processing varies systematically. The results speak against systematic variability in complexity effects. First, there was no indication of an interaction between complexity effect and trial number or test phase in the critical sentence region. Second, a Bayes factor analysis also revealed strong to extreme evidence against systematic differences in complexity effects in the control group, both within a session and between test phases.

The self-paced listening results of the control group add to the results of Studies 1 and 2 and might help uncover which language processes vary systematically or unsystematically in language-unimpaired adults: Systematic variability in complexity effects was present in the accuracy and response times but not in the eye-tracking and self-paced listening data. As explained in the motivation section above, eye-tracking and self-paced listening measure online sentence processing, while accuracy and response times measure offline processing. The difference in variability between eye-tracking and self-paced listening, on the one hand, and accuracy and response times, on the other hand, could therefore be attributed to differences in variability between online and offline processing.

Similar to the results of the control group, the results of the IWA also speak against the assumption that complexity effects vary systematically. The IWA displayed structural complexity effects in declarative sentences, relative clauses, and control structures with pronouns. Increases in complexity effects occurred in control structures with pronouns and relative clauses. However, the increase occurred *within* a session in control structures and *between* sessions in relative clauses. Thus, the differences in complexity effects are not consistent across sentence structure and session. This conclusion is corroborated by the Bayes factor analysis. This analysis revealed moderate to extreme evidence against systematic differences in complexity effects in the IWA, both within a session and between test phases.

The results of the IWA on variability in complexity effects in self-paced listening are consistent with the results obtained in Studies 1 and 2. All studies showed little evidence for interactions between complexity effects and test phase. Furthermore, in contrast to the control group, complexity effects rather increased than decreased within or between sessions in the IWA. However, these increases in complexity effects occurred only for single sentence structures and for different sentence structures across the dif-

ferent measures (accuracy and target fixations: declarative sentences, listening times: control structures with pronouns and relative clauses). Thus, these occasional increases in complexity effects rather do not speak for systematic between-session variability in IWA and might be accidental.

Overall, the listening time analysis yields the following picture regarding the variability in the control group and the IWA. In both groups, raw listening times decreased systematically over time, but complexity effects did not systematically change in either group. Thus, both groups seem to vary systematically in task performance rather than in syntactic processing. Regarding the question of which processing step varies systematically in the control group, the joint consideration of Studies 1, 2 and 3 revealed that offline rather than online processing varies systematically. For the IWA, the assumption that syntactic processing varies unsystematically was strengthened by considering the data of all three studies.

8 General discussion

This thesis aimed to investigate variability in aphasia, which is frequently reported anecdotally but rarely examined systematically. More specifically, the present work investigated variability in sentence comprehension building on studies by the research groups of Caplan and McNeil that already attested for variability in syntactic processing in aphasia (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan et al., 1997; Hageman et al., 1982; Hula & McNeil, 2008; McNeil et al., 2005; McNeil et al., 2015). However, these studies lacked a comprehensive examination of variability in syntactic processing for the dimensions of persons, measures, and occasions (cf. Hultsch et al., 2011). Caplan and his colleagues did not consider the dimension of occasions, and McNeil and his colleagues did not study variability across measures. Both groups of researchers did not directly compare variability in language-impaired and language-unimpaired adults. By comprehensively comparing variability in persons, measures, and occasions, the present work gained an overview of variability in syntactic processing that may lead to a better understanding of sentence processing in aphasia.

The following sections summarize and discuss the findings from Studies 1, 2, and 3. The next section will address the variability in the raw data (i.e., response times and accuracy in sentence-picture matching and object manipulation, listening times in self-paced listening, and target fixations in visual world eye-tracking). It will be argued that the variability in raw data does not necessarily allow for conclusions regarding syntactic processing because this variability might be due to differences in non-linguistic processes.

8.1 Variability in raw data

The raw scores indicated intra-individual variability between sessions. Both participant groups showed a decrease in response times and listening times and a slight increase in response accuracy in the retest compared to the test. The target fixations did not increase in the retest, but this might be due to limitations in the statistical analysis[1]. Thus, the raw scores overall suggest that both language-impaired and language-unimpaired adults vary systematically in their performance over time. Since there was an increase in response accuracy and a decrease in response times and listening times, the systematic changes across time are interpreted as practice effects. That is, both participant groups improved slightly over time.

[1]Due to limited computing capacity, the target fixations could not be evaluated in a single statistical model, which was possible for the other dependent measures. Therefore, fewer data were available to measure changes in the target fixations. In a similar eye-tracking study with a higher number of items, Mack et al. (2016) found an increase in target fixations between a test and a retest. Therefore, although the eye-tracking measures did not vary systematically in the present study, an increase in target fixations might be expected if all sentence structures could be analyzed jointly.

In addition to intra-individual variability, the raw data also reflected variability between participant groups. The IWA had a lower response accuracy and longer response times during self-paced listening and picture selection than the control participants. Furthermore, in the eye-tracking experiment, the IWA showed less fixations to the target picture than the control participants.

What does the variability in the raw scores tell us about syntactic processing? The practice effect seen in both participant groups might speak for an improvement in *linguistic* processing over time. However, the practice effect could also be attributed to a change in *non-linguistic* processing. A non-linguistic improvement over time could, for example, result from habituation to the experimental setting, improved understanding of the instructions, or more efficient execution of the tasks (Fine et al., 2010). For example, participants might speed up in self-paced listening because they become more efficient in the procedure of pressing a key to request new phrases. Also the variability between participant groups is not necessarily related to differences in syntactic processing between individuals with and without aphasia. The difference in reaction times could also be due to differences in motor ability. For example, some of the IWA had to respond with their non-dominant hand which might have slowed their responses. Additionally, besides difficulties in linguistic processing, difficulties in task-related processing might also slow down response speed in IWA (Caplan et al., 2006). Thus, the variability in raw scores does not allow us to draw clear conclusions regarding syntactic processing in aphasia.

The present study demonstrated that it is methodologically unfavorable to consider raw data in order to investigate variability in syntactic processing since influences of non-linguistic factors cannot be ruled out. The remainder of the discussion will therefore focus on variability in syntactic complexity effects, i.e., difference scores in which differences in overall response speed are largely factored out.

8.2 Variability in structural complexity effects

As already pointed out in Chapter 6.1.4, variability in complexity effects can be viewed from two perspectives, namely their *occurrence* in general and their *strength*. Both perspectives will be addressed in the following. Each section will end with a conclusion as to whether the results are consistent with the predictions of the resource reduction hypothesis (Caplan, 2012). As introduced in Chapter 3, the resource reduction hypothesis assumes that variability in sentence comprehension in aphasia is caused by random noise. Therefore, there should be random fluctuations in the strength of complexity effects of IWA, but in general, complexity effects should occur across all tasks and test sessions in IWA.

The *occurrence* of syntactic complexity effects was stable because these effects

occurred across all three dimensions of variability. On the dimension of persons, complexity effects occurred in both the IWA and the control group. On the dimension of measures, complexity effects were found in response accuracy and reaction times in all tasks (object manipulation, self-paced sentence-picture matching, regular sentence-picture matching) in Study 1, as well as in target fixations in eye-tracking in Study 2 and in listening times in self-paced listening in Study 3. Finally, on the dimension of occasions, complexity effects were present in both test phases. An increased processing effort for syntactically complex compared to syntactically simple structures thus occurs across participant groups, tasks, and test phases.

The findings concerning the complexity effect in aphasia are consistent with the resource reduction hypothesis. As predicted by the hypothesis, complexity effects occur in a stable manner. The hypothesis explains the occurrence of complexity effects by the increased demand for resources for processing syntactically complex sentences compared to syntactically simple sentences. The stability in the complexity effect indicates that the resources required for syntactic processing in IWA are permanently below the required level. This is because complexity effects should not have occurred across tasks and test sessions if the resources were occasionally high enough. In addition, the results of the IWA are similar to the control group's results, as complexity effects occurred across tasks and test sessions in both groups. The resource reduction hypothesis explains syntactic processing in aphasia, but Caplan et al. (2007, p.148) draw a connection to sentence processing in language-unimpaired participants. The authors hypothesize that insufficient resources could also explain the processing difficulties of language-unimpaired participants. For example, the lack of resources could explain their processing difficulties in garden-path sentences. Under the assumption that an insufficient level of resources causes complexity effects in both participant groups, one can conclude that the processing mechanisms are similar individuals with and without aphasia.

The *strength* of complexity effects was variable because, across all three dimensions of variability, they occurred at varying degrees. On the dimension of persons, there was a striking difference between the participant groups in response accuracy, as IWA showed pronounced complexity effects, whereas the control participants showed hardly any complexity effects due to ceiling effects. In contrast, there was no systematic variability between the IWA within the aphasia group. More specifically, the differences in complexity effects between the IWA could not be attributed to demographic variables (age, years of education, years post-onset) and cognitive or language abilities (working memory, aphasia severity, and aphasia syndrome). The results regarding the variability in the dimension of persons are consistent with the resource reduction hypothesis. The hypothesis explains the differences between participant groups by the difference in the overall level of resources between language-impaired and language-unimpaired adults. Furthermore, the hypothesis can account for the unsystematic variability between the

IWA because the resources in each IWA vary independently from the demographic or cognitive variables studied.

On the dimension of the measures, differences in the strength of the complexity effect between the different tasks occurred in accuracy and response times. The control group showed unsystematic variability between object manipulation and sentence-picture matching and systematic variability in the form of a lower complexity effect in self-paced compared to regular sentence-picture matching. The IWA showed unsystematic variability in complexity effects between tasks. The results of the IWA are consistent with the resource reduction hypothesis because differences in the strength of complexity effects could not be attributed to differences in tasks. Instead, the unsystematic between-task variability in IWA may be due to random fluctuations in processing resources as predicted by the resource reduction hypothesis.

On the dimension of occasions, there were notable differences in the type of variability between the IWA and the control group and between the online and offline measures. The IWA showed unsystematic variability in the complexity effects in the offline measures (i.e., reaction times and accuracy) because complexity effects increased in one sentence structure and decreased in one other sentence structure between the test and retest. In contrast, the control participants showed systematic variability in the offline measures because complexity effects decreased across three sentence structures between the test and retest. In the online measures (i.e., target fixations and listening times), neither participant group showed systematic variability between the test phases. The results of the IWA are consistent with the resource reduction hypothesis because the between-session variability in the complexity effects was unsystematic. Thus, the between-session variability in IWA could be ascribed to random fluctuations in processing resources as predicted by the resource reduction hypothesis.

The findings on between-session variability suggest that IWA show unsystematic variability, whereas control participants show unsystematic variability online and systematic variability offline. That is, between-session variability appears to be similar between participant groups in online processing and different in offline sentence processing. These differences in the type of between-session variability may be indicative of differences in sentence processing between language-impaired and language-unimpaired adults. The following sections will discuss the implications of the differences in between-session variability between IWA and control participants for impaired and unimpaired sentence processing. First, however, attention will be drawn to caveats regarding the observed between-session variability. Three caveats relate to the difference in variability of complexity effects in the offline versus online measures. The caveats indicate that this difference may be explained in purely non-linguistic terms. Two caveats relate to the systematic variability in the control group's offline data. These caveats indicate that this variability only appears systematic but actually is unsystematic.

8.2.1 Caveats on the findings regarding between-session variability

One possible non-linguistic explanation for the difference in variability online versus offline would be that longer reaction times are generally associated with more variability (Hultsch et al., 2011). Indeed, offline responses were overall slower than online responses in both participant groups (see Table 4 in comparison to Figure 2). To address the problem that longer reaction times are associated with more variability, the thesis investigated variability in difference scores, i.e., complexity effects. Thus, this caveat is controlled for, although it cannot be completely ruled out.

Furthermore, it could be argued that complexity effects are larger offline than online, allowing for greater variation in the effect between test and retest offline. In the present work, the complexity effects in the control group and the IWA are about 20 ms to 100 ms and 50 ms to 400 ms larger in the offline versus online response times respectively (see Chapter 7.3 and Study 1, Chapter 3.1). Thus, the complexity effect is slightly larger online than offline. Therefore, it cannot be completely ruled out that the somewhat stronger complexity effects led to the larger variability in complexity effects offline.

Finally, a reason for the difference in between-session variability online versus offline might be that the complexity effects could be estimated with a higher precision offline than online because more data is available offline than online. A higher precision might allow for detecting small differences in complexity effects between test phases. However, the precision of the complexity effects in the online listening times is higher than in the offline response times (see Chapter 7.3 and Study 1, Chapter 3.1). Thus, this caveat does not explain the present findings.

Another criticism relates to the systematic variability in the offline responses of the control group. The decrease in the structural complexity effect could be due to a floor effect in the participants' reaction times (Fine et al., 2010). Possibly, the control participants responded faster in the retest due to increased motor efficiency. A decrease in reaction times in simple sentences could have been impossible because the participants had already reached their fastest possible response time in the test phase (floor effect). The complexity effect could then have decreased because the reaction times could only become faster in structurally complex sentences but not in structurally simple sentences. However, this argument can be refuted because the reaction times in the control group were slower in the simple sentences offline than online (see see Table 4 in comparison to Figure 2). Thus, the online data attest that the control group could have responded even faster in the simple sentences offline. Hence, it is difficult to explain the decrease in offline complexity effects by a floor effect.

Alternatively, one might argue that improving motor execution has a stronger effect on structurally complex sentences than on structurally simple sentences (Prasad

& Linzen, 2021). More specifically, complex sentences are read slower than simple sentences at the start of the experiment. Therefore, there is more room for a task-related speed-up in complex sentences. In the present study, complex sentences were read slower than simple sentences online and offline. Thus, there was more room for a speed-up in complex sentences online and offline. Hence, the complexity effects should have decreased online and offline if the task-related speed-up affected complex sentences more than simple sentence. However, the complexity effect decreased only offline. Thus, it is difficult to explain the decrease in offline complexity effects by the overall slower reaction times for complex versus simple sentences at the start of the experiment.

The above-mentioned caveats show that differences in between-session variability between online and offline measures or between language-impaired and language-unimpaired adults could be explained without recourse to syntactic processing. These caveats should be kept in mind when reading the following sections. However, most of the issues could be refuted, at least in part. Thus, it is still conceivable that the differences in between-session variability between individuals with and without aphasia are due to differences in linguistic processing. The following sections start from this premise, i.e., they are based on the assumption that language-unimpaired individuals show systematic between-session variability in linguistic processing, and IWA show unsystematic between-session variability in linguistic processing.

8.2.2 Which processing step varies differently in language-impaired and language-unimpaired adults?

The difference in variability in offline but not online processing suggests that sentence processing differs in the two participant groups at a late stage of sentence processing. However, it has yet to be clarified what exactly differs between the two groups at this late time point in sentence processing. Below, different processes that could be reflected in offline responses are introduced.

A late processing step thought to be reflected in offline responses is the sentence wrap-up. There are several suggestions as to what sentence wrap-up is (Stowe et al., 2018). The wrap-up could involve checking that the complete sentence has been processed before lower-level information about the sentence is deleted from memory (Stowe et al., 2018). The sentence wrap-up might also involve completing processes at the syntactic, semantic, or discourse level that could not be completed online (Just & Carpenter, 1980; Reichle et al., 2009). For example, such incomplete processes might arise during pronoun processing or in case of ambiguities (Just & Carpenter, 1980). In addition, the sentence wrap-up might involve establishing inter-clausal connections and integrating the sentence into the larger context (Just & Carpenter, 1980).

Furthermore, offline responses could reflect postinterpretive processing. More specifically, Caplan and Waters (1999) propose a distinction between *interpretive* and

postinterpretive processing. The authors use the term *interpretive processing* to refer to "highly automatic, unconscious processes associated with assignment of the preferred syntactic structure and meaning of sentences" (Caplan et al., 2011, p.449) and they use the term *postinterpretive processing* to refer to "conscious, controlled processes that require storage and manipulation of linguistic representations" (Caplan et al., 2011, p.449). Postinterpretive processing, for example, includes decision-making in acceptability or grammaticality judgment tasks and response selection in sentence comprehension tasks (Caplan et al., 2013a). This task-related processing could be reflected in offline responses.

Given the proposals what offline responses reflect, differences in between-session variability between participant groups might originate from differences in sentence wrap-up or postinterpretive processing. Thus, the control group's systematic decrease in complexity effects between the test and retest phase could reflect changes in this wrap-up or postinterpretive processing over time. Likewise, the unsystematic variability in complexity effects in the IWA could indicate that no change in this wrap-up or postinterpretive processing occurred. Automatic sentence processing, on the other hand, seems to proceed similarly across sessions, as there was little variability in complexity effects in both participant groups online.

Several previous studies fit these conclusions to the present data. For example, in a study by Bader and Meng (2018) on the processing of declarative sentences, German-speaking language-unimpaired participants only showed processing difficulties with non-canonical sentences when answering comprehension questions but not when judging the plausibility of the same sentences. Bader and Meng (2018) concluded that participants had difficulties after completing sentence processing, i.e., when retrieving the information about who is agent or patient from memory. The findings of Bader and Meng (2018) are consistent with the the present findings because they demonstrate that processing difficulties with complex sentences can occur in postinterpretive processing, i.e., independently of interpretive processing.

Furthermore, James et al. (2018) found that individual differences between language-unimpaired participants in language experience and general cognitive abilities correlated with canonicity effects in sentence comprehension accuracies but not with canonicity effects in online listening times. According to James et al. (2018), these results are consistent with the assumption of Caplan and Waters (1999) that sentence processing happens automatically, whereas postinterpretive processing depends on cognitive abilities. Like James et al. (2018), Caplan and Waters (2005) also found a correlation between cognitive abilities (working memory, age) and canonicity effects offline but not online. The results of these studies fit the present findings because, in all studies, systematic variability in complexity effects in language-unimpaired participants occurred in the offline data but not in the online data. Thus the results of James et al. (2018) and Caplan and Waters (2005) support the conclusion that wrap-up or postinterpretive processing

can vary independently of interpretive processing.

However, Caplan et al. (2011) mention that the distinction between interpretive and postinterpretive processing is not synonymous with online and offline processing. Instead, postinterpretive processing begins during online processing (Caplan et al., 2011). The assumption that people start task-related processing before the sentence end seems plausible. To reconcile the assumptions of Caplan et al. (2011) with the previously mentioned studies on postinterpretive processing (Bader & Meng, 2018; Caplan & Waters, 2005; James et al., 2018) and the conclusions of the present work, one has to assume that postinterpretive processing can sometimes take place online, but happens mainly at the end of the sentence. In this way, complexity effects can occasionally vary systematically before the end of the sentence, as in Caplan et al. (2011). However, the majority of systematic variability should occur at the end of the sentence, as is the case in the present work and the studies of Bader and Meng (2018), Caplan and Waters (2005), and James et al. (2018).

So far, only studies investigating postinterpretive processing in language-unimpaired adults have been discussed. For language-impaired adults, Caplan et al. (2007) argue that if IWA have a disorder in interpretive processing, errors in sentence comprehension should be associated with an online pattern that deviates from the control group. In contrast, if IWA have a disorder in postinterpretive processing, errors in sentence comprehension should be associated with a regular online pattern. Since the IWA examined by Caplan et al. (2007) showed an irregular online pattern in incorrectly answered trials, the authors concluded that IWA exhibit intermittent deficiencies in interpretive processing while postinterpretive processing is unimpaired. The present work suggests that variability in IWA and control participants differs in postinterpretive processing. However, given the findings of Caplan et al. (2007), the difference between participant groups might originate from a disturbance in interpretive processing.

This section revealed that language-impaired and language-unimpaired adults show differences in between-session variability in sentence wrap-up or postinterpretive processing. A subsequent question is what leads to the systematic variability in sentence wrap-up or postinterpretive processing in language-unimpaired participants. This question will be considered next.

8.2.3 What leads to systematic between-session variability?

A systematic decrease in complexity effects between sessions, such as the one observed in the control group of this thesis, has already been observed in previous studies with language-unimpaired adults. Due to its similarities to the present work, a study by Wells et al. (2009) is presented first.

In the study of Wells et al. (2009), language-unimpaired participants read subject and object relative clauses over four sessions and answered comprehension ques-

tions about the sentences. As in the present work, participants did not receive explanations regarding the sentence structures or feedback on their response accuracy. Participants showed a complexity effect in the form of longer response times for object versus subject relative clauses. As in the present work, there was also an interaction between sentence complexity and session, as the complexity effect decreased after the repeated presentation. Wells et al. (2009) ascribed the interaction between sentence complexity and session to an experience effect. The authors compared this experience effect with syntactic priming effects. In both effects, prior experience with a structure changes the processing of this structure in subsequent trials. The authors distinguished the effects by the time that elapses between the presentation of the prime and target sentences. While the priming effect occurs in successive trials within the same test session, the experience effect occurs between sessions that are several days apart. According to Wells et al. (2009), however, both the priming effect and the experience effect are based on the same mechanism, namely statistical learning.

Due to the similarity with the Wells et al. (2009) study, this thesis assumes that the decrease in complexity effects observed in the control group of the present work is an experience effect. Subsequent publications refer to the experience effect observed by Wells et al. (2009) as *syntactic adaptation* (e.g. Fine et al., 2010; Harrington Stack et al., 2018; Kaan & Chun, 2018). Therefore, this thesis adopts the term syntactic adaptation. Syntactic adaptation refers to an implicit change in syntactic processing that occurs simply through a repeated presentation with a sentence structure, leading to improved comprehension of syntactically complex sentences.

In addition to the study by Wells et al. (2009), there is a growing body of research on syntactic adaptation whose discussion is beyond the scope of this thesis. Syntactic adaptation has, e.g., been studied in German canonical and non-canonical declarative sentences. For example, Kroczek and Gunter (2017) found that after exposure to non-canonical sentences, German listeners increasingly chose non-canonical readings in comprehension questions about ambiguous sentences while they had preferred the canonical reading before the exposure. Furthermore, Henry et al. (2017) showed that the availability of unambiguous case cues and prosodic cues influenced the fixation behavior of German-speaking participants in visual world eye-tracking. Participants adapted their fixation behavior depending on whether both cues or only one cue was present (Henry et al., 2017). Further studies on syntactic adaptation in other languages can, e.g., be found in the review by Kaan and Chun (2018).

However, several studies failed to find syntactic adaptation. For example, the attempts to replicate the syntactic adaptation results of a study by Fine et al. (2013) were unsuccessful (Andrews, 2021; Dempsey et al., 2020; Harrington Stack et al., 2018), or syntactic adaptation was only observable in a sample of several hundred participants (Prasad & Linzen, 2021). One possible reason for the absence of syntactic adaptation in

these studies could be that the period of exposure to the critical sentence structure was too short (Harrington Stack et al., 2018). In these replication studies, the exposure and the test for adaptation took place in blocks in the same testing session. In contrast, in the Wells et al. (2009) study and the present work, participants were exposed to structurally complex sentences over multiple sessions. Thus, although sentence processing may adapt, adaptation may require many trials of the critical structure or repeated stimulation with the critical structure over a long period of time.

In summary, the systematic between-session variability in complexity effects in the language-unimpaired group likely reflects syntactic adaptation, i.e., a change in the processing of complex sentences through repeated exposure to these sentences. As explained in the previous section, this adaptation probably improves sentence wrap-up or postinterpretive processing for complex sentences. For example, the language-unimpaired participants may improve the efficiency of retrieving the agent and patient from memory to solve the sentence comprehension task. The following section will focus on the mechanism underlying syntactic adaptation and why adaptation might fail to occur in IWA. First, the thesis offers a proposal of how the resource reduction hypothesis (Caplan, 2012) could explain adaptation and its absence. Subsequently, existing explanations for syntactic adaptation are discussed.

8.2.4 Explanations for syntactic adaptation and for its absence

This thesis used the resource reduction hypothesis of Caplan (2012) to explain the variability in aphasia. Study 2 demonstrated that the resource reduction hypothesis could explain the results of the IWA from visual world eye-tracking. In addition, the resource reduction hypothesis also correctly predicted the findings regarding variability in self-paced listening and offline responses of the IWA. Thus, the present results on variability in sentence comprehension in aphasia are consistent with the predictions of the resource reduction hypothesis. Furthermore, as mentioned above, Caplan et al. (2007) suggest that the resource reduction hypothesis could also be applied to syntactic processing in language-unimpaired participants. Building on this idea, the following paragraphs discuss whether the resource reduction hypothesis can also explain syntactic adaptation in language-unimpaired participants and the lack of adaptation in IWA.

The resource reduction hypothesis explains variability in the strength of complexity effects *between participants* by the strength and frequency of fluctuations in the available resources. The noise level determines the strength and frequency of the fluctuations. This thesis assumes that, analogously, variability in the strength of the complexity effects *within participants* can also be explained by the strength and frequency of the fluctuations in the available resources. Figure 7 illustrates this proposal.

Based on Caplan et al. (2007), one can assume that the resources in language-unimpaired adults fluctuate over time. If the noise rate is constant, complexity effects

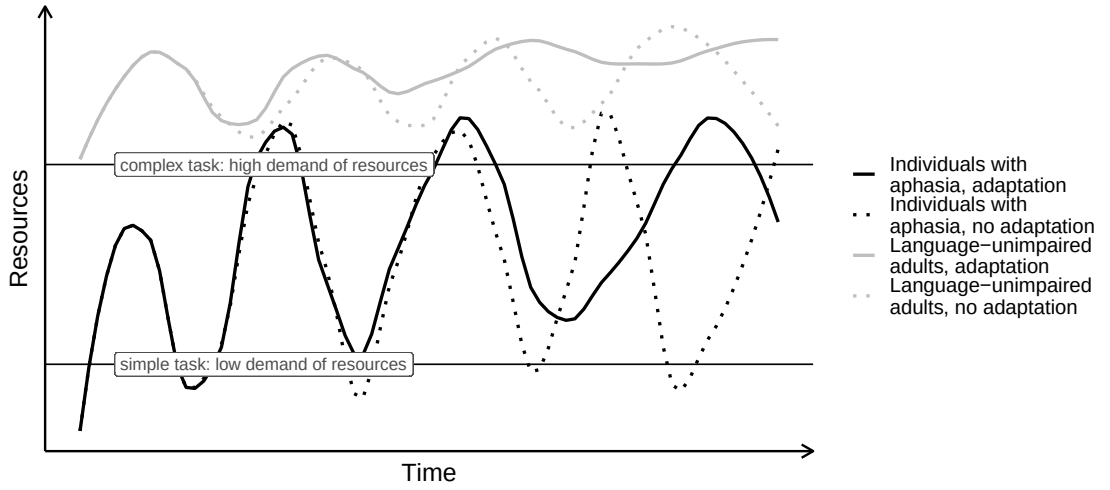

Figure 7: Proposal of how syntactic adaptation could be explained in the context of the resource reduction hypothesis. The solid lines represent the available resources in language-impaired (black) and language-unimpaired (grey) adults when adaptation is present. The dotted lines represent the resources when no adaptation is present.

remain the same over time (Figure 7, dotted gray line). In contrast, if the noise rate decreases, complexity effects decrease over time (Figure 7, solid gray line). More specifically, the frequency and amplitude of resource fluctuations decreases, and the occasions where resources fall below the level needed to process complex sentences get more rarely over time. Thus, the decrease in the noise rate offers a way to explain syntactic adaptation in language-unimpaired adults.

Two mechanisms could explain the lack of syntactic adaptation in IWA. First, one could assume that the overall reduction in resources is too large and the changes in noise rate over time are too small to lead to a measurable decrease in complexity effects, as illustrated by the solid black line in Figure 7. Although the frequency and strength of the fluctuations decrease, the available resources do not exceed the required resources for complex sentences long enough to lead to syntactic adaptation. Instead, as the solid black line shows, the changes in the noise rate lead to an increase in performance for structurally simpler sentences over time. More specifically, the resources fall less often below the level needed to process simple sentences over time (i.e., in Figure 7 the solid black line falls less often below the threshold of the simple task over time). Thus, according to this explanation, the complexity effect in IWA should increase over time because the increase in resources improves the processing in simple sentences rather than the processing in complex sentences.

A second explanation for the lack of syntactic adaptation in IWA would be that the noise rate is constant due to the language disorder. This is illustrated by the dotted black line in Figure 7. If this explanation is correct, there should be no change in the complexity effect in IWA under repeated exposure to the sentence structures.

The previous discussion demonstrates that the resource reduction hypothe-

sis could explain the occurrence of syntactic adaptation in language-unimpaired participants and the absence of syntactic adaptation in IWA via changes in the noise rate. However, Caplan (2012) does not consider that the noise rate can change. Therefore, it should be asked whether it is plausible to explain syntactic adaptation in terms of the noise rate. There is, in fact, already an approach in which noise plays a crucial role in explaining syntactic adaptation, namely, the rational inference approach (Gibson et al., 2013; Gibson et al., 2016; Warren et al., 2017). Next, this approach is presented and compared with the modified resource reduction hypothesis.

The rational inference approach explains effects in sentences in which plausibility is manipulated (e.g., *The mother gave the candle to the daughter* versus *The mother gave the candle the daughter*, Gibson et al., 2013). For this reason, this approach is described with an implausible sentence, listed in (2). In this sentence, the verb from the original sentence of the materials used in this thesis (*washes*) has been replaced by another verb (*bites*), resulting in an implausible sentence.

(2) *Implausible non-canonical declarative sentence*
Hier beißt den$_{acc}$ Tiger der$_{nom}$ Esel.
here bites the$_{acc}$ Tiger the$_{nom}$ donkey
'Here the donkey bites the tiger.'

According to the rational inference account, utterances are not always correctly transmitted from the speaker to the listener (Gibson et al., 2013; Gibson et al., 2016; Warren et al., 2017). Instead, the perceptual input may be altered by noise (e.g., slips of the tongue of the speaker, background noise, perceptual or processing difficulties of the listener). Therefore, the listener will not always rely on the utterance's literal meaning but may infer the speaker's intended meaning (ibid.). The probability that the listener will infer a new meaning depends on the likelihood that noise altered the utterance. This trade-off between literal and inferred meanings can be expressed using Bayes' rule. For example, one can determine the posterior probability that a speaker intended an implausible meaning in which the *donkey* is the agent and the *tiger* is the patient of the sentence in (2), given that the listener perceived the sentence in (2). Based on Bayes' rule, this posterior probability is given by the prior probability that the speaker intended the implausible meaning multiplied by the likelihood that the listener perceived (2) given that the speaker intended the implausible meaning:

$$P(\text{implaus}_{inteded}|\ (2)\ _{perceived}) \propto P(\text{implaus}_{intended}) \times P((2)\ _{perceived}|\text{implaus}_{inteded})$$

According to the rational inference account, the listener uses their world knowledge and knowledge of the frequency of sentence structures to determine the posterior probability of an utterance. Thus, based on world knowledge, the listener can determine that the prior probability is low that the speaker intended the implausible meaning. In ad-

dition, due to the low frequency of non-canonical declarative sentences in German, the likelihood that the speaker uses (2) to convey the implausible meaning is low. Therefore, the posterior probability that the speaker intended the implausible meaning when the listener perceived (2) is low, and the utterance was likely altered by noise. According to the rational interference account, the listener will try to infer which sentence was uttered and will edit the sentence by adding or removing single words (Gibson et al., 2013). For example, the listener could change the implausible sentence *The mother gave the candle the daughter.* to the plausible sentence *The mother gave the candle to the daughter.* by adding *to*. However, (2) would be difficult to edit by adding or removing single words. Therefore, one would have to assume a different kind of editing, namely swaps. Swaps are considered possible by Gibson et al. (2013) but not pursued further. If swaps are possible, the German sentence could be edited as in (3) by swapping the determiners *der* and *den*.

(3) *Plausible canonical declarative sentence*
Hier beißt der$_{nom}$ Tiger den$_{acc}$ Esel.
here bites the$_{nom}$ Tiger the$_{acc}$ donkey
'Here the tiger bites the donkey.'

The posterior probability that the speaker intended this plausible meaning, given the edited sentence in (3), is obtained as follows:

$$P(\text{plaus}_{inteded}|\ (3)\ _{edited}) \propto P(\text{plaus}_{intended}) \times P((3)\ _{edited}|\text{plaus}_{inteded})$$

Based on world knowledge, the prior probability that the speaker intended the plausible meaning is high. Furthermore, due to the high frequency of canonical declarative sentences in German, the likelihood that (3) is used to convey the plausible meaning is high. Therefore, the posterior probability that the speaker intended the plausible meaning given the edited utterance in (3) is high. Overall, the posterior probability of the edited sentence is higher than the posterior probability of the perceived sentence. Therefore, the probability that the listener adopts the inferred meaning instead of the literal meaning is high.

The probability that the listener adopts the inferred meaning does not only depend on the posterior probability of the perceived and edited sentence but also on additional factors. For example, the probability of a rational inference depends on the proximity between the perceived and edited structure, i.e., the more edits are necessary, the less likely an inference is (Gibson et al., 2013). Furthermore, the probability of rational inference depends on the noise rate because more noise makes an inference more likely (Gibson et al., 2013). Finally, the plausibility and frequency of sentence structures influence the occurrence of inferences (Gibson et al., 2013). For example, Gibson et al. (2013) showed in an experiment that language-unimpaired participants more of-

ten adopted the literal meaning when confronted with many implausible sentences than when confronted predominantly with plausible sentences. Thus, participants can adapt to the sentences in the input by increasing or decreasing the number of inferences they make.

The syntactic adaptation found in the control group in this thesis could be explained by the rational inference approach as follows: In the first test session of the study, the likelihood that the speaker used a non-canonical sentence to convey a meaning was low. Therefore, listeners sometimes inferred that the utterance was noisy and that the speaker intended a canonical sentence. Due to the high frequency of non-canonical sentences, the likelihood that the speaker used a non-canonical sentence to convey a meaning increased over the course of the test sessions. Therefore, the listeners adapted to the input by reducing the number of inferences they made for non-canonical sentences. This adjustment of rational inferences can explain the decrease in the complexity effect in accuracy[2].

Regarding sentence processing in aphasia, Gibson et al. (2016) and Warren et al. (2017) suggest that IWA have more noise in their processing system than language-unimpaired listeners due to the language disorder. IWA adapt their processing to the high noise level by relying more on structural frequency and plausibility of sentences than language-unimpaired listeners. Therefore, in the experiments of Gibson et al. (2016) and Warren et al. (2017), IWA made more inferences than language-unimpaired listeners, especially in implausible sentences.

The absence of syntactic adaptation in IWA in the present work could be explained by the rational inference approach as follows: As in the control group, the likelihood that the speaker used a non-canonical sentence to convey a meaning was low at the beginning of the study. In addition, due to the increased noise level in their language system, IWA frequently assumed that the utterance was distorted by noise. Therefore, IWA often inferred that the speaker intended a canonical sentence. However, despite the high frequency of non-canonical sentences, IWA did not adjust their assumption about the presence of noise over the course of the test sessions due to the increased noise level in their language system. Therefore, the IWA did not adapt to the input by making fewer inferences, which explains the lack of a decrease in the complexity effect in accuracy.

How do the rational inference account and the modified resource reduction hypothesis presented above differ? First, the accounts partly differ in their notion of noise. The resource reduction hypothesis regards noise as random fluctuations of resources in the listener's language system. The rational inference account also assumes noise in the listener's language system, but the speaker or the environment can also induce noise (e.g., slips of the tongue, ambient noise). Furthermore, the hypotheses differ in their ex-

[2]The elaboration of predictions regarding response times based on the rational inference account is beyond the scope of this thesis and could be addressed in future studies.

planations for incorrect responses. According to the resource reduction hypothesis, the comprehender lacks the resources to process the sentence correctly. According to the rational inference account, the comprehender assumes that they did not process the sentence correctly and infers that the speaker uttered a different, more plausible sentence. Finally, the rational inference account and the modified resource reduction hypothesis differ in their explanations of syntactic adaptation. According to the resource reduction hypothesis as proposed in this thesis, the noise in the listener's language system decreases over time, and thus, the resources more reliably exceed the level needed for sentence processing. According to the rational inference account, the noise in the listener's language system remains the same. Instead, the probability changes that the listener infers that a message has been distorted by noise.

Based on this comparison, a future study could determine which of the two accounts describes sentence processing more accurately. The modified resource reduction hypothesis would predict that the comprehender's noise in the language system leads to incorrect processing and that a change in noise rate leads to adaptation. The rational inference account would predict that the comprehender's assumption about the presence of noise leads to incorrect comprehension and that a change in the assumption about whether noise is present leads to adaptation.

Further explanations for syntactic adaptation are introduced, e.g., in the overview article on syntactic adaptation by Kaan and Chun (2018). Among others, error-based learning models (e.g., Chang et al., 2012; Dell & Chang, 2014) could explain syntactic adaptation. For example, Wells et al. (2009) described an error-based learning approach. Since the present work is similar to the study of Wells et al. (see Section 8.2.3), their error-based learning approach is presented to conclude this chapter.

Error-based learning approaches are based on connectionist models. In these models syntactic adaptation is regarded as implicit learning (e.g., Chang et al., 2012; Dell & Chang, 2014). The basis for implicit learning is that comprehenders predict what structure will occur in the input. This prediction is matched with the actual structure in the input. If a prediction error occurs, i.e., the prediction and the actual input do not match, the processing system adjusts its connections based on the prediction error. Due to this adjustment, it becomes more likely that the comprehenders will predict the target structure in the future. This process is implicit since the processing system unconsciously adjusts. For their findings, Wells et al. (2009) propose that the comprehenders' implicit knowledge about the distribution of word order in relative clauses changed due to the increased exposure to object relative clauses. This change in implicit knowledge is permanent, and comprehenders can generalize this knowledge to new sentences, resulting in syntactic adaptation.

According to this error-based learning approach, a cause for the lack of syntactic adaptation in IWA might be an impairment in implicit learning. Several studies

investigated implicit learning in IWA. Commonly, they measure reaction times during the repeated presentation of the same sequence of stimuli. A decrease in reaction times is regarded as implicit learning. In addition, they measure whether and how the reaction times increase when a new sequence appears after the repeated presentation of the same sequence. The sequences can consist of visual-spatial or auditory material. Several studies have shown that IWA can implicitly learn visuospatial sequences (Dominey et al., 2003; Goschke et al., 2001; Schuchard et al., 2017; Schuchard & Thompson, 2014; Vadinova et al., 2020; Zimmerer et al., 2014). In contrast, most studies suggest that IWA have difficulties in the implicit learning of auditory and linguistic sequences (Christiansen et al., 2010; Dominey et al., 2003; Goschke et al., 2001). Thus, impaired implicit learning could be a reason for impaired syntactic adaptation in IWA.

Based on the assumption that a disturbance in implicit learning causes a lack of syntactic adaptation in IWA, one may ask which step of implicit learning might be affected in IWA. According to error-based learning approaches, the basis for implicit learning in sentence processing is forming a prediction about the sentence structures in the input. Otherwise, if there is no prediction, it cannot be adjusted. Thus, one reason for the lack of syntactic adaptation would be that IWA do not predict sentence structures. However, if IWA would not have a preference for certain sentence structures, they should choose non-canonical and canonical readings equally often in sentence comprehension tasks. Therefore, since IWA have a preference for canonical structures, it seems likely that they are able to predict the canonical word order.

After forming a structural prediction, the next step for implicit learning is to recognize any prediction error. To this end, the input first needs to be perceived correctly. Next, the syntactic structure in the input needs to be matched to the syntactic prediction. Finally, the prediction needs to be updated if a mismatch between input and prediction is detected (Cope et al., 2017). IWA might have difficulties matching and updating their prediction about sentence structure. Such difficulties could cause IWA to be inflexible in revising incorrect predictions when they do not match the input. The impairment in prediction revision in each trial could eventually lead to reduced syntactic adaptation.

One way to test the hypothesis that IWA can form predictions but have difficulties revising incorrect predictions would be to conduct a combined EEG and visual world eye-tracking study. Eye-tracking would reveal whether IWA show predictive fixations in initially structurally ambiguous sentences. Simultaneously, the EEG would reveal whether IWA notice a prediction error when the sentence is resolved to a non-canonical structure. The absence of an ERP response in combination with continued fixations to the wrong interpretation would suggest that IWA can form but not revise predictions.

In summary, the findings regarding the differences in between-session variability between language-impaired and language-unimpaired adults suggest the following. First, language-unimpaired adults seem to adapt to syntactic structures, whereas IWA

have difficulties adapting to the input. Second, syntactic adaptation, or the lack of it, seems to affect offline performance, specifically, sentence wrap-up or postinterpretive processing. Third, the absence of syntactic adaptation in IWA might be explained by an increased and stable noise rate, by difficulty adjusting rational inferences, or by an impairment in implicit learning for linguistic material.

9 Overall conclusion

To conclude this thesis, the research questions outlined in Chapter 5 will be repeated and answered. The first research question aimed at a detailed description of the variability in sentence comprehension in aphasia. The question was:

1. How does sentence comprehension performance in aphasia vary on the three dimensions, persons, measures, and occasions?

The present work has shown that the occurrence of structural complexity effects is stable across the three dimensions persons, measures, and occasions. In contrast, the strength of structural complexity effects is variable across all three dimensions. An overview of the variability in the strength of complexity effects in aphasia is given in Table 2.

Table 2: Summary of the results concerning research question one.

Dimension	Variability in the size of complexity effects in aphasia offline (accuracy, response time)	 online (fixations, listening time)
Persons	**unsystematic** (not due to differences in demographic or cognitive measures)	**unsystematic** (not due to differences in aphasia severity or overall response accuracy)
Measures	**unsystematic** (not due to differences in object manipulation vs sentence-picture matching, or self-paced vs regular listening)	not applicable (only one task tested with the respective measure)
Occasions	**unsystematic** (not due to differences in test and retest)	**unsystematic** (not due to differences in test and retest)

As summarized in Table 2, there is unsystematic variability in the strength of complexity effects in aphasia on all three dimensions, i.e., persons, measures, and occasions. That is, variability in complexity effects in IWA cannot be attributed to the demographic or cognitive factors studied, differences in the tasks, or changes between test phases. Furthermore, the variability of the IWA was compared with the variability in a language-unimpaired control group. Complexity effects also occurred stably in the control group, whereas the strength of the complexity effects varied across the three dimensions. However, different from the IWA, the control group showed systematic variability in measures and occasions in offline performance, i.e., variability in sentence processing differed between language-impaired and language-unimpaired adults.

Having gained an overview of the variability in sentence comprehension in aphasia, the second research question examined an explanatory approach for the variability. The question was:

2. Can the resource reduction hypothesis explain the sentence comprehension performance in aphasia and especially the variability in the performance?

The answer to this question is yes. The resource reduction hypothesis (Caplan, 2012) correctly predicts the occurrence of structural complexity effects in sentence comprehension performance. Furthermore, the assumption of the resource reduction hypothesis that the variability in sentence comprehension in aphasia is caused by noise in the language system and is thus unsystematic is confirmed. Moreover, the resource reduction hypothesis might also explain the differences in variability between language-impaired and language-unimpaired participants. More specifically, it was suggested that the noise rate might decrease over time, causing structural complexity effects to diminish in language-unimpaired adults.

The findings on variability in sentence comprehension in aphasia open up further questions. For example: How does variability in aphasia affect other linguistic levels? Although the resource reduction hypothesis has been proposed for sentence comprehension, the explanation for variability, i.e., noise in the language system, is very general and thus not limited to the sentence level. If variability in aphasia is indeed caused by noise in the language system, then variability at other linguistic levels should also be unsystematic. To investigate this question, linguistically complex and simple conditions should be compared since linguistic and non-linguistic variability are difficult to separate in raw data, as illustrated in the present work. For example, word frequency or semantic typicality could be manipulated when examining variability in word comprehension.

The differences in the variability between IWA and language-unimpaired adults provide many opportunities for further research. In this work, the absence of systematic between-session variability in aphasia was attributed to problems in syntactic adaptation. Subsequently, it could be asked whether syntactic adaptation is absent altogether in IWA or whether it develops more slowly than in language-unimpaired adults. This would require stimulation with the sentence material over an extended period. Furthermore, it could be investigated what causes syntactic adaptation and an impairment in adaptation. For example, it might be worthwhile to investigate the mechanisms that lead to implicit learning (e.g., prediction, prediction error).

Overall, this thesis successfully replicates the earlier findings on variability in sentence comprehension (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan et al., 1997; Hageman et al., 1982; Hula & McNeil, 2008; McNeil et al., 2005; McNeil et al., 2015). As in the previous studies, variability in sentence comprehension in IWA emerged between test sessions and between tasks. Thus, the results previously available for English are confirmed with a sizable sample of German speakers. Furthermore, the present thesis extends the previous findings on variability in sentence comprehension (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan

et al., 1997; Hageman et al., 1982; Hula & McNeil, 2008; McNeil et al., 2005; McNeil et al., 2015). Not only was variability in sentence comprehension in aphasia identified, but this variability was further characterized by using hierarchical Bayesian modeling. This statistical approach allowed showing that variability in sentence comprehension in aphasia is unsystematic. In order to investigate the extent to which variability in sentence comprehension is a specific phenomenon in aphasia, for the first time, data from the control group and IWA were evaluated jointly. This joint evaluation made it possible to establish that variability differs between language-impaired and language-unimpaired adults because systematic variability occurred only in the control group. Finally, different online and offline methods were considered to describe variability. This combination of online and offline methods allowed for narrowing down the differences in variability between IWA and language-unimpaired adults to the offline data. In conclusion, the systematic investigation of variability contributes to a better understanding of language processing in aphasia and thus enriches aphasia research.

Study 1

Variability in sentence comprehension in aphasia in German

Article published in
Brain & Language 222 (2021) 105008
Dorothea Pregla, Paula Lissón, Shravan Vasishth,
Frank Burchert, and Nicole Stadie

Abstract: An important aspect of aphasia is the observation of behavioral variability between and within individual participants. Our study addresses variability in sentence comprehension in German, by testing 21 individuals with aphasia and a control group and involving (a) several constructions (declarative sentences, relative clauses and control structures with an overt pronoun or PRO), (b) three response tasks (object manipulation, sentence-picture matching with/without self-paced listening), and (c) two test phases (to investigate test-retest performance). With this systematic, large-scale study we gained insights into variability in sentence comprehension. We found that the size of syntactic effects varied both in aphasia and in control participants. Whereas variability in control participants led to systematic changes, variability in individuals with aphasia was unsystematic across test phases or response tasks. The persistent occurrence of canonicity and interference effects across response tasks and test phases, however, shows that the performance is systematically influenced by syntactic complexity.

1 Introduction Study 1

In the millennium issue of *Brain and Language* authors were invited to forecast the research issues of the next century with respect to the relationship of language and the brain (Joanette & Small, 2000). As one of these issues, Nespoulous (2000) identified the variability in performance of individuals with aphasia (IWA). The author lists five kinds of variability that research on aphasia should account for: 1) cross-linguistic variation, i.e., the variable characteristic of aphasia in different languages, 2) between-participant variability, i.e., the spread of performance in a group of participants (Shammi et al., 1998) , 3) between-task variability, i.e., the variation in performance depending on the task, 4) within-participant and within-task variability, i.e., the differences in performance *between* sessions or *within* sessions on successive trials of homogeneous tasks (McNeil, 1983), and 5) the variability in lesion sites among IWA (Nespoulous, 2000). Our research targets the variability in the area of auditory sentence comprehension in aphasia: We investigate the between-task variability in three sentence comprehension tasks focusing on specific syntactic effects (i.e., canonicity and interference effects) and the variability of the performance in each task between two test phases (i.e., test–retest variability). These types of variability will be investigated within and between language impaired and unimpaired participants.

In the next sections, we will outline the research on between-task and between-session variability in sentence comprehension in aphasia including a discussion of within- and between-participant variability.

1.1 Between-task variability in sentence comprehension

Differences in behavioral responses of participants between sentence conditions are generally ascribed to the manipulation of experimental variables but these differences could also depend on the response task that is carried out. In fact, various linguistic effects measured in brain responses (Caplan, 2010), listening and reading times (Hahn & Keller, 2018; Weiss et al., 2018), or fixation proportions (Salverda et al., 2011) in language unimpaired participants are affected by the response task. In what follows, we refer to the differences in performance that arise when the same linguistic stimuli are tested in different response tasks (e.g., object manipulation vs. sentence-picture matching) as task effects. Given the influence of task effects on the dependent variables commonly studied in psycholinguistic research, the question arises how to interpret differences in performance: as effects of linguistic manipulations or as effects imposed by the response task (Caplan et al., 2008).

The issue of task effects over and above linguistic effects is also important in the field of aphasia: Theoretical accounts of sentence comprehension in aphasia should consider that sentence comprehension difficulties are not solely induced by the sentence

structure but could rather be induced by the response task or both. Thus, if it is the response task itself that causes comprehension difficulties, this would hint at a processing deficit rather than a structural deficit (Caplan et al., 2013a; Caplan et al., 2007). To date, studies investigating task effects in sentence comprehension in aphasia are still sparse.

However, one group of researchers investigated task effects on sentence comprehension performance in more than 150 IWA and several response tasks (Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 2007; Caplan et al., 1997). Their results indicated correlations between response tasks, i.e., as accuracy scores in one response task increased, accuracy scores in the other response task also tended to increase. In addition, Caplan et al. (2006, 2007a, 2013a) analyzed the comprehension performance of a critical sentence (e.g., passive *The man was scratched by the boy*) in comparison to its syntactically less complex baseline sentence (e.g., active *The man scratched the boy*) within each IWA. These analyses revealed that despite the correlations individual participants mostly do show task dependent deficits for specific sentence constructions, i.e., in that difficulties in critical constructions (as compared to the baseline) were mostly observable in one but not in the other response tasks. Therefore, the authors concluded: "what appear to be specific deficits in individual pwa [people with aphasia] … are the result of differential demand made by different sentence types in different tasks and different levels of ability in different pwa…" (Caplan et al., 2013a, p.4).

In sum, it does not seem that there is a particular response task that is equally difficult to all IWA (Caplan et al., 2006; Caplan et al., 2013a). However, specific aspects of response task might pose problems in general: The availability of different options, e.g., in sentence-picture matching, could be difficult for IWA because inputs with opposing meanings need to be compared (Cupples & Inglis, 1993) or because distractors could interfere with the sentence interpretation of a participant (Caplan et al., 2013a). On the other hand, pictures often display the action mentioned in the sentence, which could facilitate comprehension in comparison to object manipulation tasks where the action of the sentence has to be enacted by the participant (Caplan et al., 2007; Caplan et al., 2013a; Des Roches et al., 2016; Kiran et al., 2012; Salis & Edwards, 2009). Additionally, object manipulation tasks require planning and executing a motor response and these executive processes might interfere with syntactic processing (Salis & Edwards, 2009). Consequently, each response task seems to have complicating and facilitating aspects for solving the response task that may affect IWA to a different extent making it difficult to determine whether a response task is generally easy or hard.

In addition, syntactic demands and response task demands might interact rendering it even more difficult to judge whether a response task is in general easy or hard to perform, e.g., a simple response task can become difficult when a syntactically complex sentence has to be processed. This means that only certain combinations of response task and sentence types induce impaired performance (Caplan et al., 2006). In most cases,

more comprehension errors can be observed in the syntactically complex sentences than in the baseline sentences. However, the reversed pattern with more errors in the baseline sentences can also occur (Caplan et al., 2006).

In order to account for this variability during sentence comprehension, Caplan (2012) proposes two essential features: resource demands and noise[1]. Considering the first feature of resource demands, the amount of resource demands associated with a given sentence type and response task can be estimated on the basis of the average accuracy and response time of language impaired and unimpaired participants, with slower and more incorrect responses reflecting higher resource demands (Caplan, 2012). With respect to the second feature of noise, the amount of noise seems to be inherent to the individual and can therefore be viewed as random error in the participant's performance[2]. Furthermore, Caplan (2012) suggests that noise could modulate the amount of resources available during sentence processing. Thus, the availability of sufficient resources leads to correct sentence processing, whereas a resource reduction results in incorrect sentence processing. Note that resource reduction is merely a descriptive phrase expressing that particular processing mechanisms are limited in IWA (Caplan et al., 2015). These processing mechanisms could be related to one or a combination of the following concepts: short-term or working memory, speed of parsing and interpretation or processing speed in general, operations needed to perform a response task such as action planning, or the ability to carry out multiple operations (Caplan, 2012; Caplan et al., 2007a; Caplan et al., 2013; Caplan et al., 2015). With the help of the two features resource demands and noise, between-task variability could be modeled as follows: Higher resource demands systematically result in more incorrect responses in syntactically complex as opposed to baseline sentences. In addition, noise randomly affects the available resources causing variable performance, e.g., occasional incorrect processing of baseline sentences and successful processing of complex sentences. In addition to fluctuations in the available resources, Caplan (2012) hypothesizes that a third feature could be necessary to explain the performance patterns, namely the general amount of available resources. This general amount of resources could be overall reduced in individual IWA. Consequently, IWA with greater resource reductions should produce more errors across sentence types than

[1]Caplan's (2012) concept of noise is different from noise in the rational inference or noisy channel approach to sentence processing in aphasia (Gibson et al., 2016; Warren et al., 2017). In the latter account, noise refers to errors of the language producer, environmental disturbance, misperceptions or sentence processing errors (Gibson et al., 2016), while in the former account noise refers to the random error in the comprehender (Caplan, 2012). In the rational inference approach, noise can lead to sentence distortions during communication making comprehenders adopt the most likely sentence interpretation. In Caplan (2012), noise affects the available resources in sentence processing and resource reductions lead to a higher variability in the performance.

[2]Note that the notion of noise is very abstract and that noise should be understood as a random error term in a cognitive model (Mätzig et al., 2018; Patil et al., 2016). As our reviewers pointed out, the noise parameter is not linked to a measurable physiological or psychological construct and therefore the construct is currently not very suitable to explain variability in IWA.

IWA with less resource reductions[3]. To conclude, the existence of between-task variability could be explained by demands imposed by the response task and the syntactic structure tested over and above the random noise inherent to the participant.

1.2 Test-retest variability in sentence comprehension

In this section, we will examine studies that investigate the performance within the same participants and the same response task but between different test sessions[4] (Shammi et al., 1998). The relationship of performance patterns between test and retest phases is usually measured by a correlation coefficient or an intraclass correlation coefficient. Several sentence comprehension studies investigated the correlation in language unimpaired participants in order to assess the stability in measurements, i.e., whether the same participant shows the same effect in a test and a retest. They reported only moderate correlations with respect to brain responses (Martín-Loeches et al., 2017), fixation proportions (Farris-Trimble & McMurray, 2013; Mack et al., 2016), or response accuracies (Flanagan & Jackson, 1997). The conclusion to be drawn from these studies is that these measurements are *not* stable within language unimpaired participants.

Instead of focusing on stability within participants between sessions, it could also be valuable to focus on variability within participants between sessions. Especially for IWA, investigating within-participant variability could shed light on the nature of the underlying sentence comprehension deficit: If a participant can understand given sentences at one test point but not at the other, one can assume that comprehension of the underlying linguistic structure is in principle spared. Therefore, within-participant variability between sessions can be interpreted as a processing deficit rather than loss of linguistic knowledge (McNeil & Doyle, 2000). Moreover, variable performance within IWA across sessions has been proposed to be an indicator for the potential of improvement after language treatment, i.e., higher variability prior to treatment should result in better treatment outcomes (Duncan et al., 2016; Porch, 1971).

Nevertheless, within the literature on sentence comprehension performance in IWA the issue of test-retest performance has rarely been considered. Test-retest performance has been investigated with the Revised Token Test using the noncomputerized 100-item variant of the test (McNeil & Prescott, 1978), the 50-item test (Park et al.,

[3]Note that while a permanent resource reduction can account for within-participant variability between different syntactic structures, it cannot account for within-participant variability on successive trials of the same syntactic structure or within homogeneous tasks.

[4]Note that we do not consider within-participant variability in one test session, i.e. moment-to-moment variability that has been investigated by McNeil and his colleagues. With respect to variability within a single test session, these authors have shown that the performance also fluctuates within IWA. Interestingly, the presence of this moment-to-moment variability is independent from the difficulty of the task while the frequency of variability increases with increasing task difficulty, and the frequency of variability is reliable between test sessions (e.g., Hageman et al., 1982; McNeil, 1983, 1988; McNeil et al., 2005).

2000) and the 100-item computerized test (McNeil et al., 2015) and these studies reported reliable test-retest scores. In another study, Mack et al. (2016) investigated test-retest performance in a sentence-picture matching task and found stable accuracy scores and response times in IWA. Thus, these few studies indicate that auditory sentence comprehension performance in IWA is stable between test sessions.

Despite of the above mentioned stability of overall scores between test sessions, the performance on each individual sentence over different test points, however, was found to be substantially variable within individual participants (Connor et al., 2000). In fact, Mack et al. (2016) observed a greater within-participant variability in sentence comprehension accuracy in IWA than in control participants. However, the within-participant variability in reaction times was actually greater in the control group. In contrast to the above mentioned stable performance, these results rather speak for a variable test-retest performance in individual IWA in sentence comprehension.

Regarding the interpretation of test-retest variability, Mack et al. (2016) and McNeil et al. (2015) hypothesize that at least parts of the observed variability can be ascribed to practice effects resulting from a higher familiarity with the general procedure and the task in the second test phase. Thus, McNeil et al. (2015) conclude that practice effects in a test-retest design in IWA do not originate from an improvement in language processing per se.

In their theoretical account for within-participant and within-task variability in sentence comprehension in IWA, McNeil and his colleagues propose that language mechanisms are preserved in aphasia (e.g., Hula & McNeil, 2008; McNeil et al., 1991). However, the central processing mechanism required to translate a stimulus into a response is slowed. The slowdown is caused by an inefficient allocation or reduction of resources in attention to tasks that require these mechanisms (Hula & McNeil, 2008). Consequently, if the demands exceed the allocated resources, the performance is intermittently impaired. The proposal that IWA have difficulties in attention allocation rather than linguistic processing per se is supported by studies on dual-task performance and experiments investigating non-linguistic abilities (Hula et al., 2007; Murray, 2000; Villard & Kiran, 2015). For example, Villard and Kiran (2015) found that IWA exhibited more within-participant variability between sessions than control participants in reaction times during non-linguistic attention tasks. This suggests that the variability is higher in the domain-general attention system for IWA relative to language unimpaired participants. In a related study, Villard and Kiran (2018) furthermore observed that the within-participant variability increased with higher task demands, confirming earlier results of McNeil (1983). These results are in line with Hula and McNeil (2008) and McNeil et al. (1991).

In the previous two sections, we presented the literature showing that sentence comprehension performance within IWA can be variable between response task and test

sessions. Accounts dealing with this variability agree in that the linguistic knowledge is preserved and that the difficulties in aphasia originate from fluctuations in the availability of processing resources (Caplan, 2012; Hula & McNeil, 2008). These fluctuations become visible when the demands imposed by the response task or the sentence structure exceed the available resources. The accounts, however, differ with respect to the hypothesized cause of the within-participant variability which either could arise due to random noise (Caplan, 2012) or to insufficient resource allocations by the control system (McNeil et al., 1991). Furthermore, the accounts differ with respect to what the resources are. Hula and McNeil (2008) ascribe the resources to the attentional system, whereas Caplan (2012) does not commit himself to one concept of resources and proposes different cognitive mechanisms such as processing speed or working memory.

In sum, the few studies investigating within-participant variability between response tasks and test points have shown both stable performance patterns in the overall accuracy and response times as well as variability at the individual level (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Caplan et al., 1997; Mack et al., 2016; McNeil et al., 2015). However, the number of studies that systematically investigated the variability in sentence processing in aphasia is still low. The current study seeks to further elucidate the between- and within-participant variability by exploring performance across different response tasks, different test points and focusing on the effects of different syntactic structures.

1.3 The present study

The overall aim of the current study is to better understand variability in sentence comprehension in aphasia. Furthermore, we intend to explore the extent of variable performance by investigating its limits. Our motivation for this investigation is to obtain a more detailed picture about the behavior of IWA in different sentence comprehension tasks, insights that could inform theoretical accounts of sentence comprehension deficits in aphasia. Furthermore, such research could guide assessment tools for detecting sentence comprehension deficits. Importantly, the current study will set the basis for a comprehensive cross-linguistic database of variability in sentence comprehension in aphasia by extending the existing dataset in English (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007) to German. In a future study, the German data presented here will be used to evaluate competing computational models of sentence comprehension in aphasia as done in Lissón et al. (2021) for English.

The extent of variability will be investigated by comparing performances in complex critical and simple baseline structures, similarly to what has been done in Caplan et al. (e.g., 2006; 2007; 2013a). A sentence structure is considered as complex if its processing is more demanding in language impaired and unimpaired participants at the

group level as expressed by longer reaction times and lower accuracies (Caplan, 2012). The amount of processing demand has been investigated by using sentences with different word orders. Therefore, we study canonicity effects which have been extensively investigated and are frequently attested in both participant groups (e.g., for language unimpaired participants: Grodner & Gibson, 2005; Vogelzang et al., 2019; e.g., for IWA: *English*: Caramazza & Zurif, 1976; *Greek*: Varlokosta et al., 2014; *Hebrew*: Friedmann, 2008; *Italian*: Garraffa & Grillo, 2008; *Russian*: Friedmann et al., 2010; *Turkish*: Yarbay Duman et al., 2011). In addition to canonicity effects, we investigate the amount of processing demand on the basis of interference effects. Interference effects arise during dependency formation in sentence processing when memory representations are similar as for example in number morphology (cf. Jäger et al., 2017). In the following sections, we will explain canonicity and interference effects in more detail.

1.3.1 Canonicity effects in sentence comprehension

Canonicity effects were investigated in declarative sentences (1) and relative clauses (2) with a non-canonical as opposed to canonical word order. These sentence structures will also be used in the present study.

(1) declarative sentence

a. Hier füttert **der**$_{nom}$ Igel **den**$_{acc}$ Hamster. (*canonical*)
here feeds **the**$_{nom}$ hedgehog **the**$_{acc}$ hamster

b. Hier füttert **den**$_{acc}$ Igel **der**$_{nom}$ Hamster. (*non-canonical*)
here feeds **the**$_{acc}$ hedgehog **the**$_{nom}$ hamster

(2) relative clause

a. Hier ist der Igel, **der**$_{nom}$ **den**$_{acc}$ Hamster füttert. (*canonical*)
here is the hedgehog **who**$_{nom}$ **the**$_{acc}$ hamster feeds

b. Hier ist der Igel, **den**$_{acc}$ **der**$_{nom}$ Hamster füttert.
here is the hedgehog **who**$_{acc}$ **the**$_{nom}$ hamster feeds
(*non-canonical*)

In German, the subject and object are distinguishable by case marking of the determiners (bold-faced). (1a) and (2a) are canonical, since the subject precedes the object. (1b) and (2b) are non-canonical, since the subject follows the object. In the processing of declaratives and relative clauses, both language unimpaired participants and IWA show canonicity effects in that they have more difficulties in processing non-canonical as compared to canonical sentences (*relative clauses:* e.g., Adelt et al., 2017; *declarative sentences:* e.g., Hanne et al., 2011). Two of the major accounts explaining canonicity effects are expectation-based accounts (e.g., surprisal, Hale, 2001; Levy, 2008) and memory-based accounts (e.g., dependency locality theory, Gibson, 2000). Expectation-based accounts

assume that non-canonical sentences pose more difficulties because they are less expected due to their lower frequency than canonical sentences. Memory-based accounts postulate that non-canonical sentences are harder to process because the object needs to be kept longer in memory than in canonical sentences (cf. Schlesewsky et al., 2003). Syntactically based accounts (e.g., intervention hypothesis) assume that canonicity effects occur because in non-canonical sentences the subject intervenes the dependency chain (Adelt et al., 2017; Engel et al., 2018; Sheppard et al., 2015; Sullivan et al., 2017). According to previous literature and the above mentioned theoretical accounts, we define non-canonical declarative sentences and object relative clauses as critical sentences because they are more complex than their canonical counterparts.

1.3.2 Interference effects in sentence comprehension

Interference effects are predicted to arise when memory representations overlap in features. One such feature is gender, which can either mismatch (3a) or match (3b) between nouns. In pronoun resolution, interference should be higher when the interfering noun (bold-faced) matches in gender with the target noun (3b).

(3) sentences with pronoun

a. Peter$_i$ verspricht **Lisa**$_j$, dass er$_i$ das Lamm streichelt. (*gender mismatch*)
Peter$_i$ promises **Lisa**$_j$ that he$_i$ the lamb pets

b. Peter$_i$ verspricht **Thomas**$_j$, dass er$_i$ das Lamm streichelt. (*gender match*)
Peter$_i$ promises **Thomas**$_j$ that he$_i$ the lamb pets

Furthermore, interference effects can vary with dependency length. In (4), a dependency has to be established between a covert pronoun called PRO and a noun of the matrix clause which controls the meaning of PRO. Interference should be higher if a noun (bold-faced) intervenes in the control relation (4b) than if the noun precedes the dependency (4a).

(4) sentences with PRO

a. **Peter**$_i$ erlaubt Lisa$_j$, PRO$_j$ das Lamm zu streicheln. (*short distance*)
Peter$_i$ allows Lisa$_j$ PRO$_j$ the lamb to pet

b. Peter$_i$ verspricht **Lisa**$_j$, PRO$_i$ das Lamm zu streicheln. (*long distance*)
Peter$_i$ promises **Lisa**$_j$ PRO$_i$ the lamb to pet

Interference effects are predicted under cue-based retrieval accounts (e.g., Lewis & Vasishth, 2005) and were found for language unimpaired participants in pronoun resolution (e.g., Badecker & Straub, 2002) and in sentences with control (e.g., Kwon & Sturt, 2016). In IWA, interference has been studied under the intervener hypothesis according to which IWA have difficulties when an element similar to the target of the dependency structurally intervenes in a dependency chain (e.g, Engel et al., 2018; Sheppard et al.,

2015; Sullivan et al., 2017). In control structures, IWA had higher comprehension accuracies when the distance between PRO and the controlling noun was short (Caplan & Hildebrandt, 1988, chap. 5). All in all, sentences where the controlling noun is distant or more similar to a second noun in the matrix clause should be more complex than the low-interference conditions (3a) and (4a).

1.3.3 Research questions and hypotheses of the current study

In order to investigate variability in sentence comprehension in language impaired and unimpaired participants, we investigate canonicity and interference effects in different response tasks and test points by measuring response times and accuracy scores. Specifically, we address the following research questions: 1) Can we observe canonicity and interference effects in sentence comprehension performance both in IWA and control participants at the group level considering all response tasks and test phases? 2) To what extent do canonicity and interference effects vary between response tasks and test points in IWA and control participants? 3) Do we observe a correlation in canonicity and interference effects between test phases and response tasks and how variable are these effects between test points and response tasks in the individual participants? In addition to these research questions, we explore the relationship between the variability in these linguistic effects and non-linguistic participant characteristics (e.g., age, years of education) in order to unveil the influence of these factors on sentence comprehension in aphasia.

In order to investigate our research questions, we study our syntactic manipulations (i.e., canonical versus non-canonical sentences, sentences with high versus low interference) in three different sentence comprehension tasks, which we will refer to as response tasks. These response tasks are object manipulation, and two variants of sentence-picture matching that differ in the presentation mode, namely sentence-picture matching at a normal speech rate, and sentence-picture matching at a self-paced speed. As discussed in the section on task variability above, both object manipulation and sentence-picture matching require syntactic processing as well as interpretation and both response tasks impose different extra-linguistic demands. With respect to the presentation mode of sentence picture matching, Caplan et al. (2007; 2015) speculate that in the self-paced presentation mode some IWA profit from the extra time for incremental processing. On the other hand, other IWA suffer from the working memory load that the extra time causes. As a result, self-paced sentence-picture matching and regular sentence-picture matching do not differ with respect to accuracy (Caplan et al., 2007). In conclusion, we do not expect systematic differences between the three response tasks at the group level as task demands are individually different and therefore level each other. Regardless of task effects, we expect canonicity and interference effects to occur in each response task. More specifically, we expect longer reaction times and lower accuracies

in the critical sentences, namely non-canonical and high-interference sentences, across all response tasks at the group level. Within individual participants in comparison to the respective group, we predict high correlations in canonicity and interference effects between response tasks for IWA but lower correlations for the control participants due to an overall lower variability in this group (Caplan et al. 1997; 2006; 2007; 2013a). Within individual participants analyzed separately, we predict variable response patterns, i.e., varying sizes of canonicity and interference effects across response tasks (Caplan et al. 2006; 2007; 2013a).

In order to study test-retest variability in canonicity and interference effects, each response tasks was carried out at two different test points. We hypothesize a decrease in response times and an increase in accuracy in the retest phase due to practice effects as reported for language unimpaired participants by Farris-Trimble and McMurray (2013), Mack et al. (2016), and Palmer et al. (2018) and for IWA by Mack et al. (2016) and McNeil et al. (2015). The correlation of canonicity and interference effects between test phases should be high in IWA and lower in the control participants because of the overall lower variability in this group (Mack et al., 2016). Within individual participants analyzed separately, we expect higher variability across test phases in IWA than in control participants for accuracy, but lower variability across test phases in IWA than in control participants for response times (Mack et al., 2016).

To summarize, our research question is whether canonicity and interference effects are observed in IWA and control participants in all tasks and test phases. These effects will be estimated within a Bayesian statistical framework. The output of Bayesian models consists of the posterior distributions of model parameters. In the current study, we consider an effect of canonicity or interference to be present if the posterior distribution is shifted in the predicted direction. This means that the difference between baseline and critical sentences is positive for accuracies (i.e., higher for the baseline) and negative for response times (i.e., faster for the baseline).

2 Methods and Material Study 1

This section begins with a description of the participants, followed by the illustration of the applied response tasks, sentences structures, and materials, that were designed to test for canonicity and interference effects. The effects were examined in two separate experiments, which will be called *canonicity experiment* and *interference experiment*.

2.1 Participants

A total of 71 adults, all native speakers of German participated in the study: 21 IWA (9 females, mean age = 60.2 years, $SD = 11.4$, range = 38–78; mean education = 15.2 years,

$SD = 3.2$, range = 8–21.50). Furthermore, 50 control participants were included that reported no history of neurological or language impairment (32 females, mean age = 47.7 years, $SD = 19.6$, range = 19–83; mean education = 18.1 years, $SD = 4.0$, range = 6–26). All participants had normal or corrected-to-normal hearing and vision as assessed with a self-report questionnaire[5]. Participants gave written consent in accordance with the ethics committee of the University of Potsdam and were paid for participation.

Control participants were recruited from the University of Potsdam and from a church parish. According to the Edinburgh Handedness Inventory (Oldfield et al., 1971), all but 2 control participants were right-handed (1 left-handed, 1 ambidexter). Control participants were screened for dementia using the Montreal Cognitive Assessment (MoCA, Nasreddine et al., 2005).

IWA were recruited from a database of the University of Potsdam and from aphasia self help groups in Potsdam and Berlin. A summary of the demographic and neurological information about the IWA is given in Table 3. In all but one participant the aphasia had been caused by a single stroke that occurred at least one year prior to participation in the study. Except from three participants, the IWA were pre-morbidly right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield et al., 1971). The Aachen Aphasia Test (Huber et al., 1983) was administered for syndrome classification of aphasia, estimation of the severity and assessment of the comprehension. The AAT comprehension score is a composite score of both auditory and visual comprehension that includes 10 items per modality on the word level and on the sentence level.

All IWA showed good auditory processing abilities at the word level, assessed with an auditory word-picture matching task (all scores at least 90% correct) and a lexical decision task (all scores at least 88% correct) of the German psycholinguistic test battery LEMO 2.0 (Stadie et al., 2013). Although IWA were less accurate (estimated effect of participant group 4%, CrI [1.6, 6.5]) and displayed longer response times than the control group (estimated effect of participant group 2120 ms, CrI [1571, 2739]) in the lexical decision task, IWA were similar to the control group with respect to the influence of psycholinguistic variables: Taking both groups together, we found lexicality effects (482 ms faster responses for words than for non-words, CrI [294, 679]), frequency effects (236 ms faster responses for high-frequency than for low-frequency words, CrI [69, 411]), and an effect of abstractness (216 ms faster responses for concrete than for abstract words, CrI [46, 387]). Frequency and abstractness did not interact with participant group, while the effect of lexicality was 334 ms bigger in the control group (CrI [190, 485]).

In total, five control participants were excluded prior to data analyses because they did not complete all experiments (2 participants) or because of a history of psychological or neurological disorder (3 participants). Furthermore, six IWA were excluded

[5]For 19 IWA, information on the intactness of hearing and vision was additionally available from the database from which they were recruited.

Table 3: Demographic and neurological data of the individuals with aphasia.

							LEMO[1] (raw scores)		AAT[2]		
IWA	Gender	Years Age	Years Education	Years P.O.	Etiology	Locali-zation	T3 (n=80)	T11 (n=20)	Aphasia type	Severity (standard nine)	Comprehension score (percentile)
2	F	72	8	7	IMI	L	77	19	Anomic	6.8 (mild)	101 (86)
3	M	76	20	17	IMI	L/R	61	20	Not-classifiable	7 (mild)	110 (97)
4	F	47	13	21	IMI	L	78	20	Anomic	7.8 (mild)	112 (98)
6	M	55	14	10	IMI	L	67	20	Anomic	6.8 (mild)	113 (99)
8	F	51	19	7	MA	L	74	20	Anomic	7.4 (mild)	100 (85)
9	M	64	15	2	IMI	L	73	20	Anomic	7.4 (mild)	109 (96)
10	M	58	18	1	IMI	L	52	20	Broca	5 (moderate)	82 (55)
11	F	63	12	1	IMI	L	73	20	Broca	6.8 (mild)	113 (99)
12	F	46	12	13	IMI	L	65	20	Broca	4.2 (moderate)	68 (36)
13	M	74	13	8	IMI	L	57	20	Broca	4.4 (moderate)	86 (61)
14	M	66	13	17	IMI	L	75	20	Anomic	6.4 (mild)	95 (75)
15	F	59	21	4	I	L	77	20	Broca	5.2 (moderate)	84 (58)
16	M	67	17	26	VH	R	72	19	Broca	5.4 (moderate)	99 (83)
17	F	43	14	10	IMI	L	65	20	Broca	6.6 (mild)	110 (97)
18	M	57	13	1	I	L	67	18	Wernicke	not available	not available
19	F	52	19	8	IMI	L	76	20	Broca	5.8 (moderate)	91 (68)
20	M	38	13	3	IMI	L	73	19	Broca	4.2 (moderate)	98 (81)
21	M	57	18	2	IMI	L	66	18	Broca	6 (mild)	104 (91)
22	F	67	16	5	IMI	L	76	20	Anomic	6.2 (mild)	106 (93)
23	M	74	15	7	IMI	L	67	20	Anomic	6.6 (mild)	106 (93)

Note. [1] LEMO 2.0 (Stadie et al., 2013) T3 = auditory lexical decision, T11 = auditory word-picture matching, [2] Aachen Aphasia Test (Huber et al., 1983), IWA = individual with aphasia, P.O. = post onset, F = female, M = male, L = left, R = right, IMI = ischemic arteria cerebri media infarct, I = infarct, MA = arteria cerebri media aneurysm, VH = vertebrobasilar hemorrhage.

because they had no apparent aphasia according to the Aachen Aphasia Test (3 participants), they scored less than 90% correct in auditory word-picture matching in LEMO 2.0 (2 participants), and one IWA stopped participation on her own.

2.2 Tasks and Procedure

We will first describe the response tasks and registration of the dependent variables for each of the three administered response tasks followed by a description of the general procedure of the current study.

2.2.1 Object manipulation (OM)

The general aim of this task was to enact the meaning of a sentence with figurines. Figurines relevant for the subsequently presented sentence were placed in front of the participant and introduced (e.g., *Hier sind Lisa und Peter.* 'Here are Lisa and Peter.'). Next, the target sentence was presented orally. In the interference experiment, which tested the comprehension of sentences with control verbs (e.g., *Peter promises Lisa to pet and ruffle the little lamb.*), participants were instructed to move the figurine (e.g., *Peter*) that "does something with the animal". In the canonicity experiment, that tested the comprehension of declaratives and relative clauses (e.g., *Here is the tiger that the donkey just comforts.*), participants were instructed to move the figurine (e.g., *donkey*) that "does something". It was not required to act out the specific action of the mentioned verbs (e.g., *tröstet* 'comforts'). Responses were scored correct if the figurine representing the agent of the target sentence (canonicity experiment) or the subject of the subclause (interference experiment) was selected. We will report the accuracy of figurine selection.

2.2.2 Sentence-picture matching, regular listening (SPM-regular)

The general aim of this task was to select one of two pictures that represented the meaning of the auditorily presented target sentence. Sentences were presented with a computer at a regular speech rate. Each trial began with a preview phase of 4000 ms during which the pictures were introduced. Following this, the target sentence was presented. Pictures were displayed until a picture was selected by the participant by button press or for maximally 30 seconds. In the interference experiment that tested the comprehension of sentences with control verbs, participants were instructed to select the picture with the referent that "does something with the animal" (an example is given in Figure 9). In the canonicity experiment that tested the comprehension of declarative sentences and relative clauses, the instruction was to select the picture "that fits with the sentence" (an example is given Figure 9). We measured the response time and accuracy of picture selection.

2.2.3 Sentence-picture matching, self-paced listening (SPM-SPL)

Aim and procedure of the task were the same as in the regular sentence-picture matching task except for the presentation of the target sentence that proceeded phrase-by-phrase (e.g., *Hier ist* | *der Tiger* | *den* | *der Esel* | *gerade* | *tröstet.* 'Here is | the tiger | that | the donkey | just | comforts'). After the preview phase, participants were prompted to press the space bar to start the target sentence. Sentence chunks were played back one by one triggered by space bar presses of the participant. Pictures stayed on the screen during sentence presentation and until the target picture was selected. We will report the response times and accuracy of the picture selection. The self-paced listening procedure was implemented with Linger Version 2.94 (Rohde, 2003).

2.2.4 General procedure

The general procedure of the study is visualized in Figure 8. We administered an object manipulation task, a regular and a self-paced sentence-picture matching task. Task administration was randomized with one response tasks per session (max. 90 minutes) and per week (mean = 8 days, $SD = 12$ days). All three response tasks were administered a second time in a retest phase after a pause of approximately 2 months between the same response task ($SD = 1$ month; similar in the two participant groups: $\Delta M = -13.61$, 95% CI $[-28.45, 1.23]$, $t(39.04) = -1.85$, $p = .071$). Similar to the first test phase, the investigation of weekly response tasks (retest: mean = 8 days, $SD = 9$ days) was randomized.

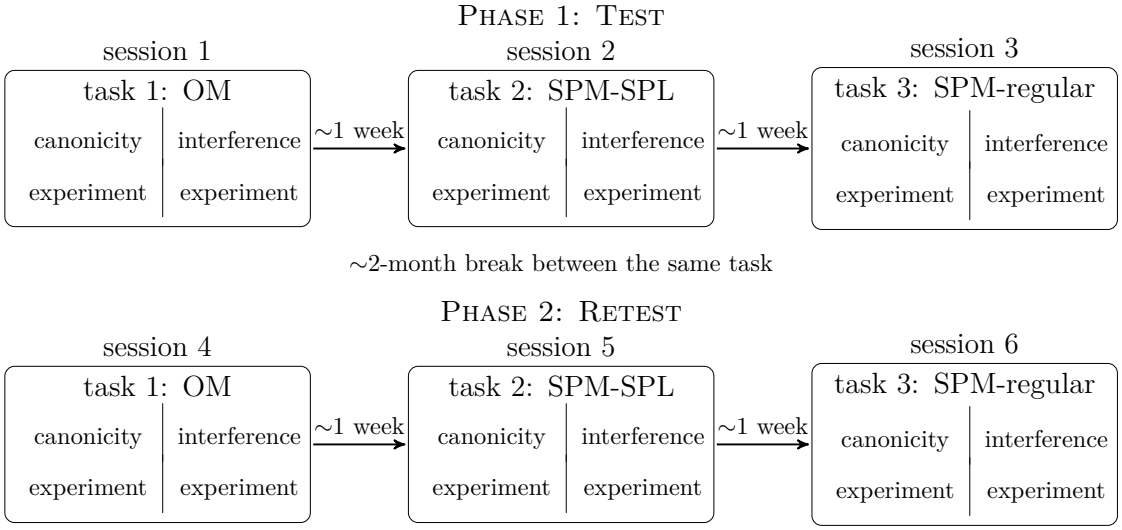

Figure 8: General procedure of the study. All participants completed an object manipulation task (OM) and two versions of a sentence-picture matching task, in which the sentences were presented at a self-paced speed (SPM-SPL) or at a normal speech rate (SPM-regular). The three response tasks were completed twice (test phase, retest phase). In all response tasks, two experiments (canonicity and interference experiment) were carried out. The order of the response tasks and experiments was randomized.

All response tasks aimed at investigating the comprehension of sentences with control verbs in order to identify interference effects, and the comprehension of declarative sentences and relative clauses in order to identify canonicity effects. Canonicity and interference effects were investigated blockwise. Within each response task, both experiments were conducted successively in randomized order, including each five practice items with feedback about response accuracy, followed by the test items without feedback. Each experiment lasted approximately 15 minutes in control participants and 30 minutes in IWA. The remaining time in each session was used for setting up and explaining the response tasks. In addition, we investigated working memory performance by administering the digit span task (forward and backward recall) of the Wechsler Memory Scale–Revised (Härting et al., 2000).

We conducted two screenings to ensure that the participants understood the items of the experiments. First, we tested that the participants were able to match the nouns of the target sentences to the pictures or figurines used in the response tasks. In case of misassignmets, participants were trained until they could correctly assign 100% of the nouns. Second, we made sure that the participants were able to auditorily discriminate the morphological endings of the verbs and determiners used in the target sentences. In an auditory discrimination task with a total of 28 items, participants heard either two identical verbs / determiners (e.g., *streichel-t – streichel-t* "pet-3SG" or *der – der* "the.nom") or minimal pairs (e.g., *streichel-t – streichel-n* "pet-3PL – pet-3PL" or *der – den* "the.nom – the.acc")) that were presented as sound files. Mean performance of the participants was 26 correct items ($SD = 2$, range = $20 - 28$).

2.3 Material

We will present the sentence structures used in the canonicity experiment, followed by the structures of the interference experiment.

2.3.1 Sentence stimuli for the canonicity experiment

Examples for the sentences of the canonicity experiment were given in (1) and (2), all items are given in the appendix. In total, the experiment had 80 sentences. We included 20 declarative sentences: 10 baseline sentences with canonical order (1a) and 10 critical sentences with non-canonical order (1b). Furthermore, we included 60 sentences which contained a relative clause, namely 30 baseline sentences with a subject relative clause (2a) and 30 critical sentences with an object relative clause (2b). These were subdivided in 10 subject and 10 object modifying relative clauses, and 10 relative causes with a plural noun in the subclause. Sentences were pseudo-randomized: Each condition appeared at most three times in a row and the same item never appeared twice in a row.

Sentences were constructed using 10 transitive depictable action verbs with

two syllables and a mean lemma frequency of 85.22 ($SD = 211.28$) per million tokens in dlexDB. The arguments of the verb consisted of two masculine two-syllable animals that had a similar mean lemma frequency in dlexDB. Twenty-three students rated that the animals of each action were equally plausible as agent or patient of the action to ensure that sentences were pragmatically reversible.

2.3.2 Sentence stimuli for the interference experiment

Examples for the sentences of the interference experiment were given in (3) and (4), all items are given in the appendix. In total, the experiment had 50 sentences. We compared the comprehension of overt pronouns in 10 baseline sentences with a gender mismatch (3a), and in 10 critical sentences with a gender match (3b) of the two main clause nouns. Furthermore, we examined the comprehension of PRO in 10 baseline sentences with object control (4a) and 10 critical sentences with subject control (4b). Finally, we included 10 filler sentences. Sentences were pseudo-randomized: Each of the four conditions (subject control, object control, match, mismatch) and the fillers appeared at most three times in a row and the same item never appeared twice in a row.

Sentences consisted of a matrix clause with two nouns and a control verb (e.g., *versprechen* "promise") and a subclause with a noun phrase in neuter gender and two synonymous action verbs. The matrix nouns were common two-syllable German first names referring unambiguously to a male or female person. Each name appeared with equal probability as the first or second noun of the matrix clause. In the sentences with PRO, nouns were always of different gender. In the sentences with a pronoun, gender was manipulated and the two matrix nouns were of equal or different gender.

Control verbs were selected from the *ZAS Database of Clause-Embedding Predicates* (Stiebels et al., 2018) by the following criteria: 1) No particle verb, 2) argument structure with one propositional argument P and two individuals x and y, 3) x and y realized in nominative and dative case, and 4) controller corresponds to x or y. Five subject control and five object control verbs with similar mean lemma frequency in the dlexDB database (Heister et al., 2011) were extracted. Sentences with PRO included a subject or object control verb to manipulated the distance between the controlling noun and PRO. Sentences with a pronoun included subject control verbs. Fillers had the same structure as the sentences with a pronoun but included object control verbs.

2.3.3 Auditory stimuli

Sentences in the object manipulation task were presented by the experimenter or as audio files in regular and self-paced sentence-picture matching. Sentences were spoken with a neutral prosodic contour, which was kept constant in all sentences. The audio files were recorded in a sound-proof booth with a trained female native speaker of German.

Each sentence was recorded twice: 1) as a whole for regular sentence-picture matching, 2) in chunks for self-paced sentence-picture matching. In regular sentence-picture matching, sentences were spoken with a rate of 4.79 or 3.95 syllables per second in the canonicity and interference experiment respectively. These rates fall in the range of 3–6 syllables per second which is considered a normal speech rate (Levelt, 2001). Recordings were post-processed with Praat (Boersma & Weenink, 2018). We used the same sound file for pairs of baseline and critical sentences (i.e., canonical / non-canonical declaratives, subject / object relatives, subject / object control, match / mismatch). This was achieved by cutting out and exchanging the manipulated region in the sound files[6]. Auditory stimuli were presented at a comfortable volume for each participant.

2.3.4 Pictures

The pictures of the regular and self-paced sentence-picture matching tasks consisted of black-and-white drawings. Per item, two pictures were presented. In the canonicity experiment, the target picture displayed the agent acting on the patient, and in the foil picture the agent and patient roles were reversed (e.g., Figure 9). In the interference experiment, the target picture displayed the target referent interacting with the animal mentioned in the sentence, and the foil picture displayed the distractor referent in the same interaction (e.g., Figure 9). Referents had the same size, adopted the same postures, and were identifiable by a letter on their T-shirt (e.g., *L* for *Lisa*). The positions of the agent being either left or right of the patient within a single picture as well as the positions of the target and foil pictures were balanced throughout both experiments.

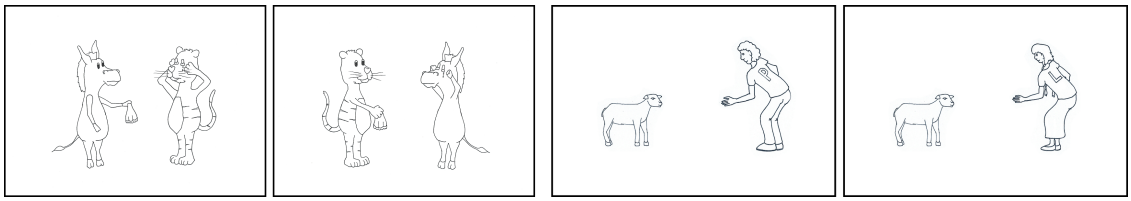

Figure 9: Sample pictures of sentence-picture matching tasks. For the canonicity experiment (left pair), the canonical sentence *Here comforts the*$_{nom}$ *tiger the*$_{acc}$ *donkey* matches the right picture and the left picture is the foil, and conversely, the non-canonical sentence *Here comforts the*$_{acc}$ *tiger the*$_{nom}$ *donkey* matches the left picture and the right picture is the foil. For the interference experiment (right pair), the object control sentence *Peter allows Lisa to pet the lamb* matches the right picture and the left picture is the foil, and conversely, the subject control sentence *Peter promises Lisa to pet the lamb* matches the left picture and the right picture is the foil.

[6]It was checked in a pilot with four students and four elderly control participants that the spliced stimuli sounded natural.

2.4 Data analysis

Data analysis was performed on accuracy scores and response times. Additionally, we evaluated the participant characteristics age, years of education, years post onset, severity (stanine) and comprehension score in the Aachen Aphasia Test, and working memory (in form of a composite score of the forward and backward digit span task). Accuracy was measured in all three response tasks (i.e., object manipulation, regular and self-paced sentence picture matching). Response times were only collected in regular and self-paced sentence-picture matching and were defined as the duration from the offset of the audio file until button press. Response times longer than 30 seconds or shorter than -1 second (i.e. when participants pressed a button more than 1 second before the trial ended) were discarded which resulted in a loss of 0.5% of the data.

The data were analysed with Bayesian methods. One major reason for choosing this approach instead of frequentist analyses was the complexity of the model structure. Frequentist models fit in *lme4* (Bates et al., 2015) did not converge when all sentence types, test phases and response tasks were included as fixed and random effects, while Bayesian models including all predictors converged. Because an important goal of our study was to evaluate the within- and between-participant variability, the inclusion of all predictors in the fixed and random effects was essential. Additionally, the credible interval of an effect in a Bayesian model can be interpreted and provides a measure of the uncertainty of the estimated effect given the data and the model. In contrast to that, the confidence interval of a frequentist model does not allow statements about the uncertainty of an effect (Kruschke & Liddell, 2018). The information about the uncertainty of the estimates is very important for the evaluation of the effects and they can be compared to predictions from computational models in future work.

We fit Bayesian hierarchical linear mixed models with correlated random intercepts and slopes for participants and items using R (Version 3.6.3; R Core Team, 2020) and the R-package *brms* (Version 2.13.0; Bürkner, 2017; Bürkner, 2018). Reaction times were log-transformed since they are skewed with a longer right tail and a left tail that is cut off at zero. Response accuracies are binary (0 and 1). Therefore, we used a logistic link function to fit a Bayesian generalized linear mixed model. We report model estimates that are back-transformed into milliseconds and proportions for the ease of interpretation. For our predictors, we used sum contrasts except for the relative clause subtypes, where we used a sliding contrast, and the continuous factors age and years of education, which were centered. In a first step, we pooled the data of the three response tasks and two test phases to estimate the overall canonicity and interference effects and added test phase and response task as separate predictors well as the factors age and years of education. To get estimates of the canonicity and interference effects for each participant group, the predictors for the sentence types were nested under participant group. Fur-

thermore, we nested the regular and self-paced sentence-picture matching tasks under sentence-picture matching. Finally, we included interactions of the sentence types with response tasks, test phases, age and education respectively. The nestings and contrast codings are illustrated in Figure 14 in the appendix. In a second step, we estimated the canonicity and interference effects separately for each repetition of the experiment. In this model, canonicity and interference effects were nested under participant group, test phase and response task. Apart from that, the contrast codings were the same as in the first model. The second model also included the factors age and years of education. In a third model, we separately evaluated the data of the IWA. In parallel to model one, this model included the predictors sentence type and the nested conditions, response task and test point. Additionally, the model contained the centered and scaled factors age, years of education, years post onset, severity (stanine) in the AAT, comprehension score in the AAT, a composite score of the forward and backward digit span task, and the sum coded predictor aphasia type (+1 anomic, -1 broca), as well as the interaction of these factors with the predictor sentence type and the nested conditions.

We specified our prior beliefs about the shape of the parameters for the Bayesian models. We used mildly uninformative priors. For the reaction time data, we set the prior of the fixed effects intercepts to $Normal(0, 10)$, the prior of the fixed effects slopes to $Normal(0, 1)$, and the prior standard deviations of the random effects and the residual error to $Normal_+(0, 1)$ which means that they are truncated in zero to only allow positive values. For the response accuracy, we set the prior of the fixed effects intercepts to $Normal(0, 1.5)$, the prior of the fixed effects slopes to $Normal(0, 1)$, and the prior standard deviations of the random effects to $Normal_+(0, 1)$ truncated in zero. The output of a Bayesian model consists of the posterior distributions of the parameters. We will report the mean and the 95% CrI of the estimated effects. The 95% CrI is the range for which we can be 95% certain that it includes the true effect, given the data and the model.

For the correlation analysis, we extracted the estimates of the correlations of the canonicity and interference effects between the test phases, between object manipulation and sentence-picture matching and between self-paced and regular sentence-picture matching from the random effects structure of the participants that are estimated together with the group level effects (cf. Kliegl et al., 2011). For this analysis, we fit separate models for each participant group and sentence type to simplify the random effects structure. Additionally, we calculated intraclass correlation coefficients for each participant group and sentence type to compare the results to earlier studies. To this end, we fit absolute-agreement two-way random effects models with the following formula (Streiner et al., 2015):

$$ICC2(A,1) = \frac{\sigma^2_{participants}}{\sigma^2_{participants} + \sigma^2_{observers} + \sigma^2_{error}}$$

where σ^2 are three sources of variance (participants, observers, error). Intraclass correlation coefficients were calculated with the R-package *irr* (Gamer et al., 2019) using the specifications (1) model "twoway", (2) type "agreement" , (3) unit "single". All data and code are accessible at https://osf.io/hb9gu.

3 Results Study 1

The mean response times and accuracies for the control group and the IWA in each sentence type across response tasks and test sessions are summarized in Table 4 in the appendix. Considering the full data set, control participants had 26% CrI: [19, 34.3] higher accuracies and responded -2082ms CrI: [-2761, -1491] faster than IWA. In both participant groups, the differences in accuracies between response tasks were close to zero for object manipulation in comparison to sentence-picture matching both in test and retest (control group, test phase: 0.2% CrI: [0, 0.5], retest phase: 0.2% CrI: [-0.1, 0.4]; IWA, test phase: 3% CrI: [-5.2, 11.3], retest phase: -2.4% CrI: [-11.1, 6]) and for regular in comparison to self-paced sentence-picture matching (control group, test phase: 0.3% CrI: [-0.1, 0.8], retest phase 0.2% CrI: [-0.2, 0.5]; IWA, test phase: -8.2% CrI: [-18.5, 1.6], retest phase: -7.8% CrI: [-17.3, 0.8]). Although participant groups showed no differences in accuracies between response tasks, they responded slower in regular than in self-paced sentence-picture matching (control group, test phase: 171ms CrI: [67, 279], retest phase: 302ms CrI: [228, 379], IWA, test phase: 504ms CrI: [-112, 1141], retest phase: 672ms CrI: [-145, 1499]). Both participant groups answered faster in the retest phase (control group: -146ms CrI: [-196, -98], IWA: -303ms CrI: [-734, 109]), but only the IWA exhibited considerable improvements in accuracy in the retest phase (control group: 0.1% CrI: [0, 0.3], IWA: 7.3% CrI: [1, 13.8]). In sum, the control group responded faster and more accurately than IWA, both participant groups had similar accuracies between response tasks but responded faster in self-paced listening than regular listening, and both groups responded faster in the retest. Additionally, accuracy scores in IWA increased in the retest.

3.1 Variability in canonicity and interference effects at the group level

The subsequent Figure 10 addresses research question one and two, namely whether we observe canonicity effects and interference effects overall across response tasks and test phases in IWA and control participants, and second whether these effects vary between response tasks and test points. Therefore, we compared the effects in the pooled data of

all sessions and tasks with the posterior estimates of the effects in each separate session.

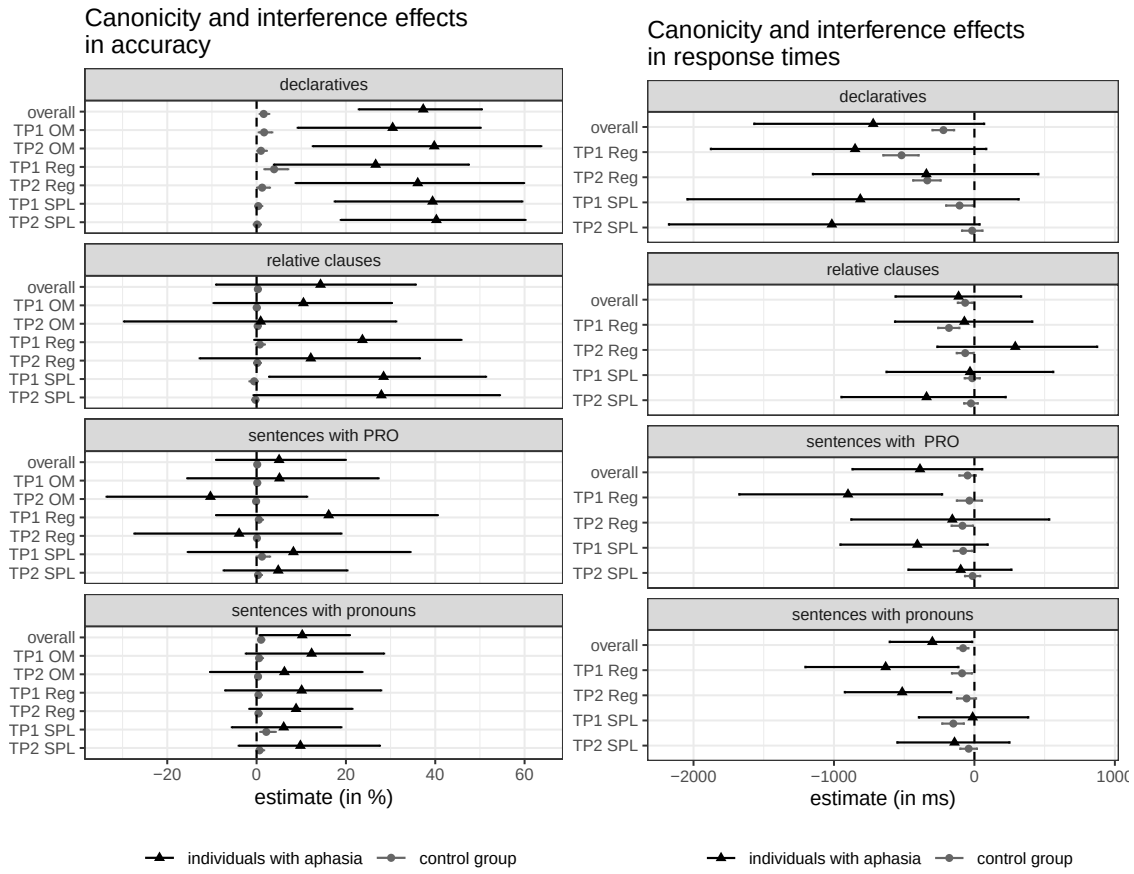

Figure 10: Canonicity effects in declaratives and relative clauses and interference effects in sentences with a pronoun or PRO in the control group (gray) and the individuals with aphasia (black). Overall effects aggregated across test phases and response tasks, and separate effects in two test phases (TP1, TP2) and three response tasks: object manipulation (OM), regular (Reg) and self-paced (SPL) sentence-picture matching. Plots display the posterior probabilities of the effects with 95% CrIs. The dashed line represents an effect size of zero. Distributions that are right-shifted denote higher accuracies and slower responses in the baseline structure (canonical or low-interference condition).

3.1.1 Canonicity and interference effects across test phases and response tasks

We will first address research question one and consider the canonicity and interference effects when pooling the data of all test phases and response tasks. In declarative sentences, both participant groups had higher accuracies and responded faster in canonical than in non-canonical sentences (control group: 1.6% CrI: [0.7, 2.8] and -220ms CrI: [-299, -144]; IWA: 37.3% CrI: [22.9, 50.4] and -721ms CrI: [-1568, 69]). Similarly, for relative clauses, both participant groups displayed higher accuracies and responded faster in canonical than in non-canonical sentences, however, the estimates were closer to zero than in declaratives and included both positive and negative values (control group: 0.3% CrI: [-0.3, 1] and -66ms CrI: [-118, -14]; IWA: 14.3% CrI: [-9.1, 35.6] and -113ms CrI: [-562, 333]). Also in sentences with PRO, accuracies were higher and response times were

faster in the baseline condition in both participant groups, however, the estimates were closer to zero than in declaratives and included both positive and negative values (control group: 0.1% CrI: [-0.2, 0.6] and -49ms CrI: [-107, 11]; IWA: 5% CrI: [-9.1, 19.9] and -388ms CrI: [-868, 56]). Also sentences with a pronoun were answered faster and more accurate in the baseline condition in both participant groups (control group: 1% CrI: [0.5, 1.7] and -81ms CrI: [-120, -43]; IWA: 10.2% CrI: [0.7, 20.9] and -300ms CrI: [-603, -15]).

3.1.2 Canonicity and interference effects in each test phase and response task

We will now address research question two and turn to the canonicity and interference effects of each single session of the experiment in IWA and the control group. We will first consider the variability in the effects between test phases followed by the variability between response tasks.

In the control group, effects were either very close to zero in both test phases or the distributions shifted closer zero in the retest phase. This decrease in effects was reflected in the interactions of test phase and baseline versus critical sentences. In the response times, these interactions occurred in all sentence types except for sentences with PRO. In accuracy scores, interactions occurred in declarative sentences (declaratives: 89ms CrI: [34, 144], -2.9% CrI: [-5.8, -0.7], relative clauses: 25ms CrI: [-7, 58], -0.7% CrI: [-1.5, 0.2], pronouns: 47ms CrI: [-10, 104], -2.3% CrI: [-8.5, 1.4], PRO: 8ms CrI: [-48, 64], -0.8% CrI: [-2.8, 1]). Considering IWA, we observed less interactions between baseline versus critical sentences and test phase. In response times, interference effects in sentences with PRO decreased in the retest. With respect to accuracies, canonicity effects in declaratives increased in the retest phase (declaratives: -33ms CrI: [-176, 108], 3.5% CrI: [0, 7.3], relative clauses: -8ms CrI: [-87, 71], 0.1% CrI: [-2, 2.2], pronouns: -108ms CrI: [-248, 27], 0.3% CrI: [-3.4, 4], PRO: 163ms CrI: [18, 314], -2.6% CrI: [-6.8, 1.3]). In sum, control participants showed decreasing effect sizes for most of the sentence types whereas IWA exhibited both increasing and decreasing effect size for only a few sentence types.

With respect to task differences, the effect sizes varied between object manipulation and sentence-picture matching in both participant groups. The control group showed more pronounced canonicity effects in declaratives in object manipulation as compared to sentence-picture matching (declaratives: 0.4% CrI: [0.2, 0.7], relative clauses: 0.1% CrI: [-0.1, 0.2], pronouns: -0.4% CrI: [-1.4, 0.2], PRO: -0.2% CrI: [-0.6, 0.1]). Conversely, the IWA showed more pronounced canonicity effects in relative clauses and more pronounced interference effects in sentences with PRO in the sentence-picture matching task as compared to object manipulation (declaratives: 1.3% CrI: [-2.6, 5.2], relative clauses: -6.2% CrI: [-8.6, -3.9], pronouns: 1.8% CrI: [-2, 5.8], PRO: -4.3% CrI: [-8.8, -0.3]). With respect to the presentation mode in the sentence-picture matching task, control participants exhibited more pronounced canonicity effects when presented in regular listening as opposed to self-paced listening. This holds true for declaratives and

relative clauses in both accuracy and response times (declaratives: -193ms CrI: [-256, -132], 0.6% CrI: [0.2, 1.1], relative clauses: -50ms CrI: [-84, -16], 0.2% CrI: [0.1, 0.5], pronouns: 19ms CrI: [-34, 74], -0.1% CrI: [-0.8, 0.4], PRO: -7ms CrI: [-62, 49], 0% CrI: [-0.2, 0.3]). In IWA, interactions were observed in accuracy, where the interference effect in sentences with PRO was more pronounced in self-paced listening. In the response times, the interference effect in sentences with a pronoun was more pronounced in the regular listening (declaratives: 52ms CrI: [-35, 140], -3.5% CrI: [-8.2, 0.9], relative clauses: 41ms CrI: [-10, 92], -0.3% CrI: [-2.7, 2], pronouns: -131ms CrI: [-223, -41], -1.1% CrI: [-5.6, 3.2], PRO: -55ms CrI: [-146, 33], -6% CrI: [-11.3, -1.1]). In sum, in both groups differences between object manipulation and sentence-picture matching were less observed than differences between regular and self-paced listening. The presentation mode influenced control participants more than IWA.

3.2 Variability at the individual participant level

In what follows, we will address research question three concerning canonicity and interference effects at an individual participant level. We will first investigate whether these effects are correlated in the participants between test phases and response tasks. Afterwards we will explore the variability in effects for each individual participant and the influence of participant characteristics on the effects.

3.2.1 Correlation in canonicity and interference effects between response tasks and test phases

In order to investigate whether sizes of canonicity and interference effects are stable in individual participants, we analyzed the correlation estimates of the random effect structure provided by the Bayesian model. Figure 11 shows the posterior estimates for the correlations of the canonicity effects in declarative sentences and relative clauses and of the interference effects in sentences with a pronoun or PRO.

With respect to the accuracy of the control group (see Figure 11A), the correlations of the canonicity and interference effects between test phases or between response tasks were close to zero in all sentence types. The IWA (see Figure 11B) displayed numerically higher correlations than the control participants. However, except for relative clauses estimates were uninformative with respect to the question whether the canonicity or interference effects are correlated between test phases or between response tasks. In relative clauses, IWA showed positive correlations in canonicity effects between test phases (0.58 CrI: [0.23, 0.82]), between object manipulation and sentence-picture matching (0.62 CrI: [0.28, 0.85]), and between regular and self-paced sentence-picture matching (0.78 CrI: [0.52, 0.93]). Thus, IWA showing greater canonicity effects in relative clauses in the test phase also showed greater canonicity effects in relative clauses in the retest

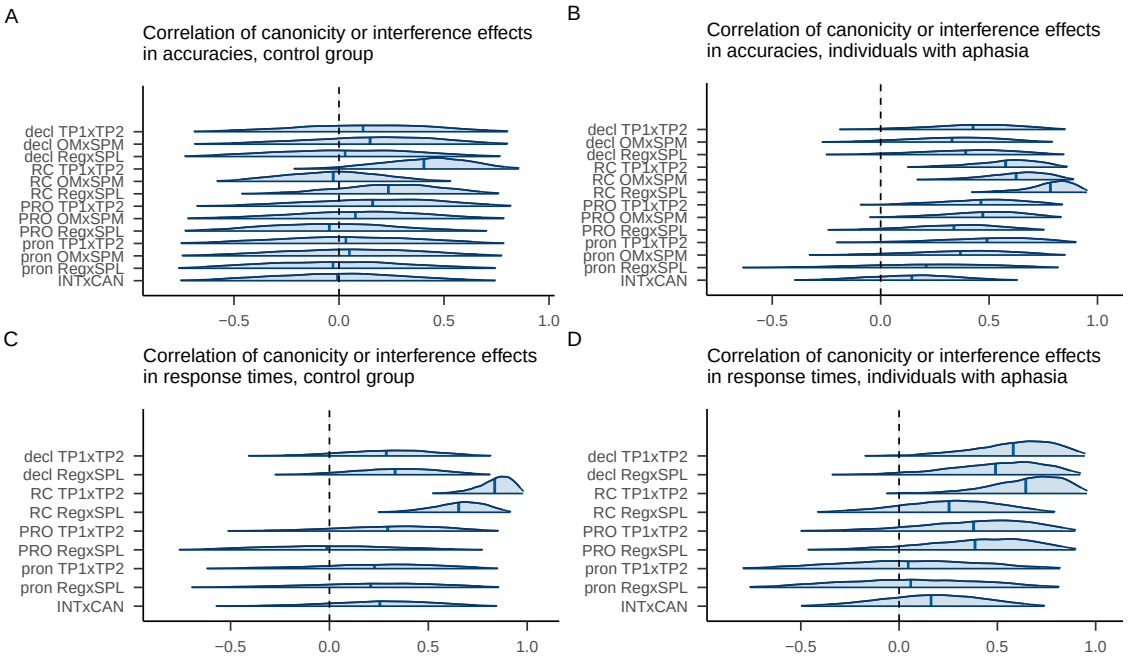

Figure 11: Correlation of canonicity effects in declarative sentences (decl) and relative clauses (RC) and correlation of interference effects in sentences with pronouns (pron) and PRO in the control group (A, C) and the individuals with aphasia (B, D). The distributions display the posterior estimates of the correlations. The shaded areas under the curves are the 95% CrIs and the solid lines mark the means. The plot depicts the correlations in accuracies (A, B) and in response times (C, D) between the test and the retest phase (TP1×TP2), between object manipulation and sentence-picture matching (OM×SPM), between regular and self-paced sentence picture matching (Reg×SPL) and between the interference and canonicity effects (INT×CAN).

phase. Likewise, greater canonicity effects in relative clauses in one task were associated with greater canonicity effects in relative clauses in the other response tasks. Additionally, we compared the size of canonicity and interference effects that each participant exhibited in the pooled data of all response tasks and both test phases. In both participant groups, the estimates were uninformative with respect to the question whether canonicity and interference effects are correlated.

Turning to the response times of the control group (see Figure 11C), participants displayed distributions close to zero or slightly positively-shifted distributions for most of the correlation estimates except for relative clauses. In this sentence type, the control group showed positive correlations in the canonicity effect between the test phases (0.84 CrI: [0.62, 0.96]) and between regular and self-paced sentence-picture matching (0.65 CrI: [0.35, 0.87]). This means that control participants showing greater canonicity effects in relative clauses in the test phase or in regular sentence-picture matching also showed greater canonicity effects in relative clauses in the retest phase or in self-paced sentence-picture matching. The IWA (see Figure 11D) displayed correlation estimates that were uninformative in all sentence types.

To sum up, only the correlation estimates of the the relative clauses in IWA (in accuracy) and control participants (in response times) were clearly positive. The distributions of the other sentence type were uninformative.

To be able to compare the results of the Bayesian analysis with earlier studies using intraclass correlation coefficients, we also calculated intraclass correlation coefficients for the correlations reported above. These are represented in Table 5 in the appendix. In our analysis, intraclass correlation coefficients around 0.8 and higher mostly corresponded with distributions in the Bayesian analysis that were situated in the positive space. However, intraclass correlation coefficients below 0.8 were associated with distributions with wide CrIs that were uninformative with respect to the question whether the effects are correlated.

3.2.2 Between- and within-participant variability in canonicity and interference effects

In order to investigate the variability in canonicity or interference effects in each individual participant, we analyzed the by-participant random effects of the Bayesian model. Figure 12 displays canonicity and interference effects (in accuracy and response times) for each single IWA with respect to all sentence types and response task for each test phase separately. Distributions with the same distance to the x-axes in each plot visualize the within-participant variability, whereas the spread of the distributions along the y-axes visualizes the between-participant variability. We assume that an effect is variable in a participant if the 95% CrIs of two distributions (e.g., test vs. retest) of this participant do not overlap.

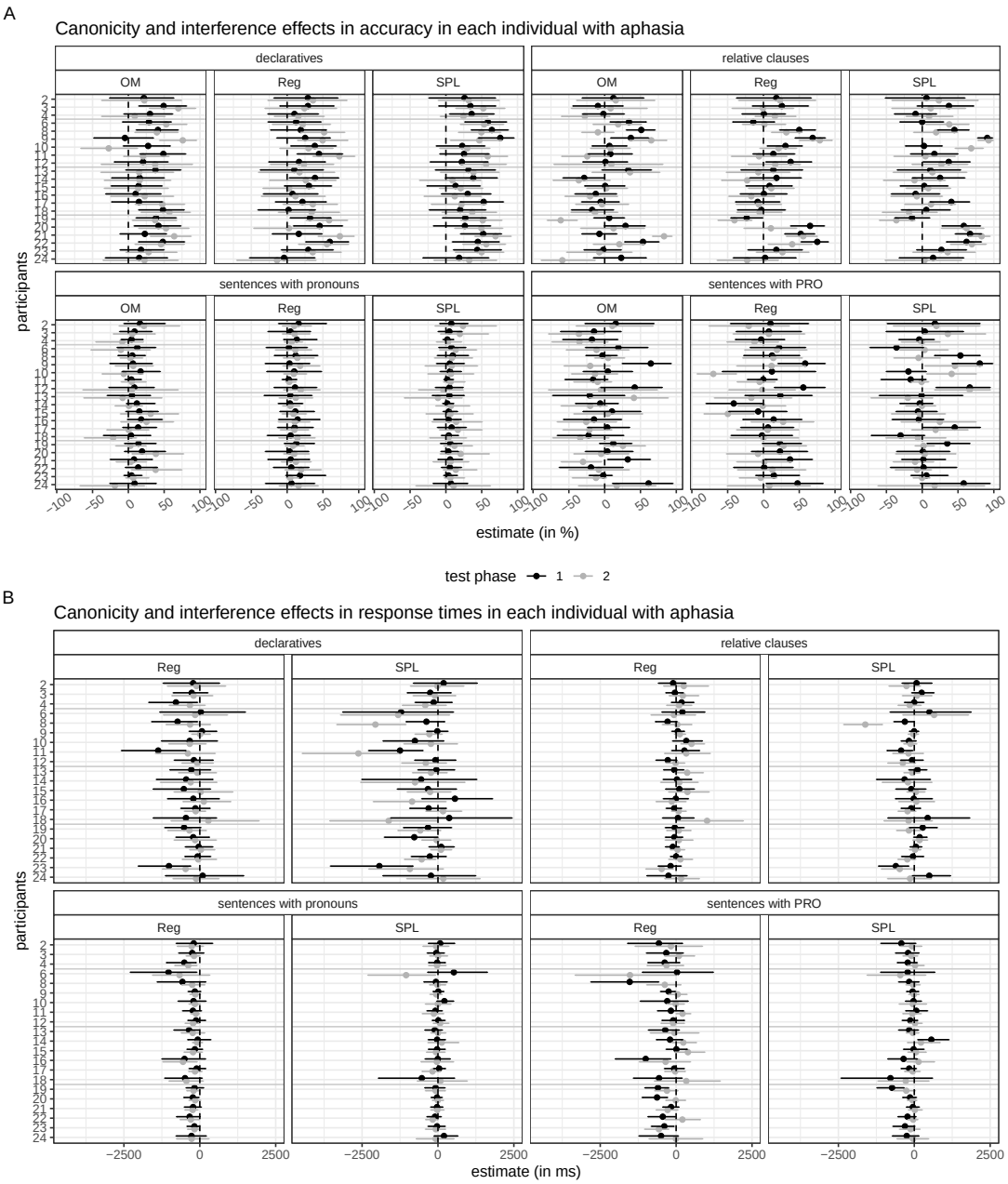

Figure 12: Canonicity effects in declaratives and relative clauses and interference effects in sentences with pronouns or PRO in accuracy (A) and response times (B) of each individual with aphasia. Each participant completed three response tasks, object manipulation (OM), regular (Reg) and self-paced (SPL) sentence-picture matching in two test phases. Plots depict mean estimates (dots) and 95% credible intervals (solid lines) of the effects. The dashed line marks an effect size of zero. Distributions that are right-shifted denote higher accuracies and slower responses in the baseline structure (canonical or low-interference condition).

We will first consider within-participant variability. In accuracy, all IWA showed comparable effect sizes between response tasks and test phases in sentences with a pronoun, and only one IWA (IWA 9) showed differences in effect sizes in declaratives, i.e., variability was low. In contrast to that, more participants showed differences in the effect sizes in relative clauses (IWA 8, 9, 10, 19 and 21) or in sentences with PRO (IWA 3, 9, 10 and 21). In response times, effects in pronouns were comparable across test phases and response tasks in all IWA. In declaratives, one participant (IWA 8), in relative clauses, two participants (IWA 8 and 10) and in sentences with PRO, two participants (IWA 8 and 20) showed differences in the effect sizes.

Only a small number of control participants (n = 3) exhibited variable effects in accuracy. In contrast, 33 participants showed larger canonicity effects in the regular as compared to the self-paced listening presentation mode in response times. Overall, differences in effect sizes only occurred in declarative sentences and in none of the other sentence structures.

We will now turn to the between-participant variability of canonicity and interference effects. In accuracy, all IWA showed either no or positive effects in sentences with a pronoun and in declarative sentences. In contrast, there were instances of negative effects in relative clauses (IWA 19 and 24) and sentences with PRO (IWA 3, 10, 14, 15 and 21). Similarly to accuracy, most of the IWA showed either no differences or faster response times in baseline sentences while occasionally participants showed negative effects (IWA 10 and 18 in relative clauses, IWA 11 and 14 in control structures).

In the control group, most of the participants showed either no or positive effects. There was only one case of negative effects in accuracy (in relative clauses) and one case with faster response times in subject control than in object control.

In sum, the within-participant variability in accuracy was larger in IWA than in controls. These differences in effect sizes in IWA, however, did not occur systematically between response tasks or test phases. The within-participant variability in response times was larger in controls. These differences in effect sizes did occur systematically, i.e., the effect sizes were larger in regular than in self-paced listening in all participants who exhibited variable effects. The between-participant variability was larger in IWA than in controls with occasionally less accurate performances and longer response times in the baseline than in the critical sentences.

3.2.3 Influence of participant characteristics on canonicity and interference effects

Finally, we explored whether differences in overall accuracy, response times and sizes of canonicity or interference effects were influenced by demographic variables (age, years of education, years post onset) and cognitive or language abilities (working memory, scores and aphasia type of the Aachen Aphasia Test). Figure 13 displays the interaction

of these different participant characteristics with the response measures and canonicity or interference effects. The overall accuracy decreased with increasing age (-12.6% CrI: [-23.3, -1.6]) and increased with higher digit span scores (15.4% CrI: [2.4, 28.1]). The remaining estimates of interactions with overall accuracy or response times were uninformative. Turning to the canonicity and interference effects, all interactions of the effects with the participant characteristics were inconclusive in accuracy. In response times, the size of the effects was influenced by two factors: Interference effects in pronouns decreased with a higher comprehension score in the Aachen Aphasia Test (-461ms CrI: [-862, -115]) and canonicity effects in declarative sentences decreased with higher digit span scores (-615ms CrI: [-1141, -134]). In sum, age and working memory influenced comprehension accuracy, whereas the interactions with response times and canonicity or interference effects were inconclusive in most cases.

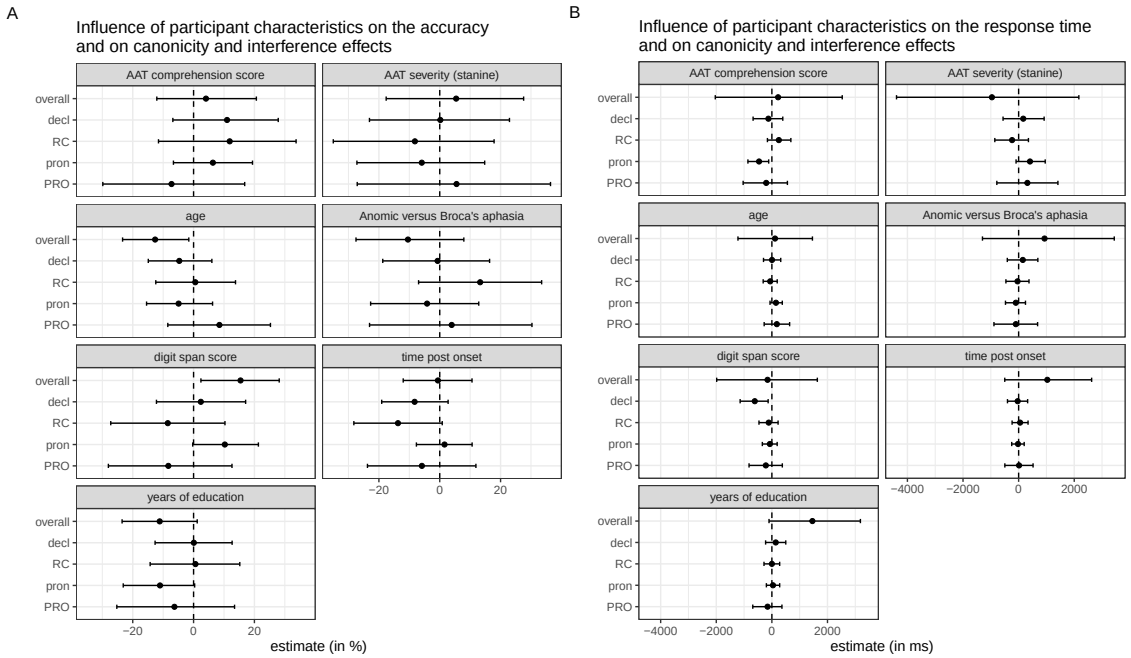

Figure 13: Mean estimates (dots) and 95% credible intervals (solid lines) of the interaction of different participant characteristics with the overall accuracy (A) and response times (B) and with canonicity effects in declarative sentences (decl) and relative clauses (RC) and interference effects in sentences with a pronoun (pron) or PRO. Distributions that are shifted to the right denote higher accuracies and slower response times between the mean value and one unit increase in the respective participant characteristic.

4 Discussion Study 1

In the current study, we investigated variability in sentence comprehension in language impaired and unimpaired participants. More specifically, we focused on the variability in the occurrence of canonicity and interference effects in three different response tasks (object manipulation, auditory sentence-picture matching with regular presenta-

tion speed, and auditory sentence-picture matching at a self-paced presentation speed). All response tasks were carried out twice, namely in a test and retest phase. Canonicity and interference effects were measured in accuracies and response times for declarative sentences and relative clauses, and for control structures with an overt pronoun or PRO. Similar to Caplan et al. (2006; 2007; 2013a), we investigated canonicity and interference effects by computing the difference in the dependent measures between a baseline sentence and its structurally more complex counterpart. Our research questions were whether canonicity and interference effects are observable in our two participant groups, and to what extent these effects vary between response tasks and test points. Furthermore, we investigated whether the size of the canonicity and interference effects correlates between test phases and response tasks and how variable these effects are in the individual participants.

4.1 Variability of canonicity and interference effects between response tasks and test phases

In line with previous studies (e.g., Hanne et al., 2011; Vogelzang et al., 2019), canonicity effects were observed in declarative sentences in both participant groups. Similarly, both groups showed interference effects in sentences with a pronoun. An interference effect in pronoun resolution in sentences with gender markings had not been attested for IWA before, thus providing additional support for the intervener hypothesis (Engel et al., 2018; Sheppard et al., 2015; Sullivan et al., 2017). In contrast to the clear canonicity and interference effects that we observed for declaratives and sentences with a pronoun, canonicity effects in relative clauses and interference effects sentences with PRO were less informative due to a lower magnitude and higher uncertainty in the effects. However, the means of the estimates of the canonicity and interference effects were shifted in the expected direction in both participant groups (i.e., better performance in the baseline compared to the critical sentences). Thus, performance patterns in the sentence structures under investigation indicated for both participant groups the occurrence of canonicity and interference effects.

With respect to the variable occurrence of the canonicity and interference effects across response tasks, previous studies hypothesized that object manipulation is a more demanding task than sentence-picture matching (Caplan et al., 2013a; Des Roches et al., 2016; Kiran et al., 2012; Salis & Edwards, 2009). Other authors assumed the reverse, namely sentence-picture matching being more demanding than object manipulation (Caplan et al., 2013a; Cupples & Inglis, 1993). In contrast, in our study the differences in the overall sentence comprehension performance between the two tasks object manipulation and sentence-picture matching were too low to support the assumption of different task demands. Therefore, we infer that task demands had no major influence on the

performance patterns. In support of this conclusion, we also did not observe systematic differences in the size of canonicity and interference effects between both response tasks. In sum, neither response task seemed to be more demanding than the other response task for the two groups of participants.

With respect to the presentation mode in the sentence-picture matching task, there were no clear differences in overall accuracy between self-paced and regular sentence-picture matching similar to previous results (Caplan et al., 2007). Unexpectedly, the control group systematically exhibited smaller canonicity effects in self-paced listening, a result that has actually been predicted for IWA (Caplan et al., 2007). This could mean that the control group profited from the extra time for incremental processing in self-paced listening. However, the IWA in our study did not show systematic differences in canonicity and interference effects between the two listening conditions. The reason why there were no systematic differences between presentation modes in the IWA could be that only some IWA profited from self-paced listening whereas others did not and as a result any potential differences were leveled. It could be speculated that it is the working memory capacity that determines whether an IWA can profit from the self-paced presentation or not.

With respect to test-retest variability, we observed varying performance patterns in both participant groups. In the retest phase, response latencies decreased in both language impaired and unimpaired groups whereas accuracy scores increased only in IWA. Increases in the overall performance of IWA were previously ascribed to a higher familiarity with the task and its execution (Mack et al., 2016; McNeil et al., 2015). However, it remained unclear whether these increases in performance can also be attributed to an improved sentence processing for complex sentences. In order to disentangle increases due to higher task familiarity from increases due to improved sentence processing we analyzed the difference between baseline and critical sentences. We focused on decreases in canonicity and interference effects as we assume that these decreases can only originate from improvements in sentence processing. In our group of IWA, the effects did not systematically decrease between test and retest, and the canonicity effect in declarative sentences even increased in the retest phase. This speaks for persistent sentence processing difficulties despite higher task familiarity as reflected by an overall higher accuracy. In the control group however, the canonicity and interference effects systematically decreased in the retest phase. Thus, it seems that the increase in performance reflects an increase in processing proficiency for complex sentences in controls whereas in IWA this increase in performance seems to reflect a higher task familiarity.

4.2 Correlations of canonicity and interference effects between response tasks and test phases

Up to this point we solely considered canonicity and interference effects at the group level and found stability in the occurrence of the effects. However, from this stability we cannot necessarily infer that the same stability holds true for each individual IWA. Therefore, we now turn to the individual level and investigate how stable canonicity and interference effects are between response tasks and between test phases within single participants. These analyses allow us to see whether the stability in the occurrence of the effects at the group level also holds true at the individual level or whether the stability at the group level originates from variability at the individual level (i.e., participants who show a large effect size in one session or response task for a given sentence structure might show a small effect in other sessions in the same sentences and other participants display the reverse). Variable performance within individual participants would corroborate theories assuming fluctuations in available resources in the processing system (Caplan, 2012; Hula & McNeil, 2008). Again, we will focus on the correlation of canonicity and interference effect sizes across response tasks and test points instead of analyzing performance with respect to accuracy or response times. Only the analysis of effect sizes can inform us about the consistency of syntactic processing in a single IWA. Studies analyzing accuracy and response times reported high correlations within IWA for various sentence types between response tasks (Caplan et al., 1997; 2007; 2013a) and between test phases (Mack et al., 2016; McNeil et al., 2015). Accordingly, we expected to observe the same consistency in canonicity and interference effects in our study. In our analyses, the estimates of the correlations in the effect sizes were only high and informative in relative clauses but not in declaratives, and sentences with pronoun or PRO where the estimates of the correlations were uninformative. However, the correlations in all sentence types were larger in IWA than in the control group and were positive, i.e., participants who showed a large effect in one session or response task also showed a large effect in another session or response task.

With respect to the high correlations we observed in relative clauses, we assume that this is due to the number of observations in relative clauses which was three times larger than in the other sentence types. The higher number of observations could have led to a higher precision in the correlation estimate in relative clauses. This higher precision could explain why IWA exhibited higher correlations in relative clauses as opposed to all other sentence types. Similarly, the control participants also displayed higher correlations in relative clauses than in the other sentence types. The high correlation in relative clauses together with the positive shift in the other sentence types lead us to conclude that the level of syntactic difficulties in each IWA is stable. This would speak for permanent reductions in available resources for syntactic processing (Caplan,

2012). While the degree of reduction remains stable within participants, the degree of reduction is different between participants. Noise, then, would play a second secondary role in syntactic processing within participants. This interpretation, though, should be confirmed with a study with a larger number of observations and a higher precision. Alternatively, the data of the current study could be used in a meta analysis.

In addition to the correlations between response tasks and test phases, we also analyzed whether there is a correlation in the sizes of the canonicity and the interference effect. Such a correlation would be expected under the assumption that a canonicity effect can be regarded as a form of an interference effect (Adelt et al., 2017; Sullivan et al., 2017). In both participant groups, we did not see a correlation between canonicity and interference effects. In addition to that, canonicity effects in declarative sentences were twice as large as interference effects in pronouns in the IWA as illustrated in Figure 4. These results, thus, do not support the intervener hypothesis which assumes that canonicity effects can be reduced to interference effects.

4.3 Within-participant variability

The correlation analyses informed us about the consistency in the size of canonicity and interference effects, in what follows, we examine the variability of the individual participants in more detail. With respect to within-participant variability, Mack et al. (2016) reported that IWA showed more variability in accuracy but less variability in response times than control participants in a sentence-picture matching task. The authors concluded from these results, that IWA are not always more variable than control participants, in contrast to the generally increased variability in IWA (e.g., Caplan et al., 2007; Villard & Kiran, 2015). Our results for canonicity and interference effects are similar to Mack et al. (2016), i.e., we observed more variability in the effect sizes in accuracy but less variability in effect sizes in response times in IWA than in controls. This corroborates the finding of Mack et al. (2016) that the variability is not always larger in IWA than in control participants.

With respect to the larger variability in control participants in response times, Mack et al. (2016) hypothesized that this variability could arise from practice effects since participants exhibited shorter response times in the retest phase. In our study, we also observed systematic changes in the control group in that each individual participant showed larger canonicity effects in regular listening compared to self-paced listening, similar to what we have seen at the group level. This means that each control participant showed a pattern similar to the group pattern. Considering the IWA, the within-participant variability in the effect sizes in accuracy were unsystematic in that each individual exhibited a unique pattern of changes in effect sizes. These unsystematic patterns of single IWA were also reflected by the pattern observed at the group level in

which systematic interactions between effect sizes in response tasks or test phases were not observed. To conclude, specific extra-linguistic task manipulations such as repetition of the experiment or presentation mode systematically influenced canonicity and interference effects in control participants but not in IWA. Thus, it seems that we are dealing with two different types of variability in syntactic processing, namely systematic versus unsystematic changes. The systematic changes in control participants can be explained by manipulated factors of the experiment whereas the changes in IWA cannot be explained by these factors. Instead, the major cause of variability in canonicity and interference effects in IWA seems to be inherent to the participant. According to theoretical accounts of variability in IWA, these factors inherent to the participant could be random fluctuations in processing resources (Caplan, 2012) or insufficient allocation of attention (Hula & McNeil, 2008).

One aspect of our findings is not in line with the concept of random fluctuations in processing resources. According to this concept, all sentence types should be affected by noise equally. However, we observed that the variability within and between participants was not of equal size across all sentence types. More specifically, we observed less variability in the effects in sentences with a pronoun than in the other sentence types, a finding that cannot be disregarded as an artifact, because it occurred across participants and across response tasks. If the variability in the effects was in fact solely due to random noise, we had to assume that the noise level systematically varies between different sentence types. A possible alternative explanation for the result can be derived from the observations of McNeil (1983) which was confirmed by Villard and Kiran (2018) that the intra-individual variability increases with higher demands. In our study, we observed that interference effects in sentences with pronouns were overall smaller than canonicity effects in declaratives and relative clauses. This difference in effect sizes could be interpreted in the sense that the increase in complexity between the baseline and critical sentences was smaller in sentences with pronoun than in the other sentences, i.e., we assume that the difference in effect sizes was due to differences in the increase of demands. Based on this assumption, we argue, in line with McNeil (1983) and Villard and Kiran (2018), that the variability in interference effects was smaller than in canonicity effects because the increase of demands was smaller in the complex pronoun sentences than in the complex declaratives and relative clauses.

Although not the main focus of the present study, we would like to turn briefly to the influence of individual participant characteristics (i.e., age, working memory, comprehension scores and aphasia type of the Aachen Aphasia Test, years of education, years post onset) on sentence comprehension performance in IWA. Our study revealed that age and working memory had an influence on the overall performance in that accuracy was higher in younger IWA and IWA with higher working memory scores, which is in line with previous studies (e.g., Caplan et al., 2011; Caplan et al., 2013b). Similarly, canon-

icity effects were smaller in IWA with a higher working memory score which would speak for an influence of working memory on syntactic processing according to Caplan et al. (2013b). However, from our study it is difficult to conclude that working memory has a general impact on syntactic processing as the interaction between syntactic effects and working memory was restricted to declarative sentences. Considering the results of the Aachen Aphasia Test (Huber et al., 1983), we did not find systematic influences of the measures severity, syndrome and comprehension score on overall accuracy and response times, with the exception of one interaction between interference effects in sentences with pronouns and the comprehension score of the Aachen Aphasia Test. The lack of interactions could be due to the high uncertainty in the estimates of the interactions. The uncertainty in turn may have resulted from the highly variable performance in our study of participants who displayed similar scores in the Aachen Aphasia Test. So far, it seems that there is no single factor that unequivocally influences the size of syntactic effects[7].

4.4 The limits of variability in aphasia

Variability in sentence comprehension in IWA can be explored from two perspectives, namely variability in overall accuracy scores and response times or variability in the size of syntactic effects. Our study focused on the variability in the size of syntactic effects because this allows us to investigate variability in *syntactic* processing. We could show that syntactic processing difficulties in IWA remain unchanged as canonicity and interference effects occurred constantly across test phases and response tasks although general accuracy increased. This leads us to hypothesize that the increase in general performance is not due to improvements in syntactic processing but rather due extralinguistic factors such as a higher task familiarity. Thus, one limit in variability in processing difficulties is their stability between sessions. In contrast, the performance of control participants in complex sentences increased which seems to be due to more efficient syntactic processing. This could be interpreted as an effect of adaptation which was absent in IWA. Furthermore, limits of variability were also seen in IWA across response tasks and modes of presentation as no systematic differences in canonicity and interference effects occurred. Again, this was different in the control group in which the variability in canonicity and interference effects was contingent upon the mode of presentation. The higher performance in self-paced as compared to regular sentence-picture matching could also be interpreted as an effect of adaptation which was absent in IWA. Yet another limit in the variability lies in differences in processing demands of

[7] An anonymous reviewer pointed out to us that the fact that we did not apply the test battery *Sätze Verstehen* (Burchert et al., 2011) is a potential limitation for the conclusions we drew in our study. In future studies suitable test procedures should be used a priori to tease apart a general sentence comprehension impairment from specific impairments such as for complex sentences.

different sentence structures. More specifically, within- and between-participant variability in syntactic effects varied depending on the type of the syntactic effect as interference effects in sentences with pronouns were smaller and less variable across IWA than canonicity effects. In sum, sentence comprehension performance in aphasia is both stable and variable. Stability can be seen in the persistent occurrence of syntactic effects and variability is observable in different sizes of these effects. However, this variability takes different forms in language impaired participants than in controls: Syntactic effects fluctuate unsystematically in IWA whereas they systematically decrease in control participants which possibly reflects adaptation to the sentence structure.

How can these limits in variability uncovered in our study inform the existing accounts of variability in aphasia by Caplan (2012) and Hula and McNeil (2008)? Both accounts can explain syntactic effects by differences in processing demands of different sentence structures and fluctuations in these effects by factors inherent to the participant such as random noise or insufficient attention allocation. However, in order to fully account for the limits of variability as reported in the current study the above mentioned processing accounts might need to take into account the adaptation to the sentence structure to explain systematic decreases in syntactic effects in control participants over time, as well as the absence of such decreases in IWA. In a processing model, adaptation could lead to a more efficient allocation of resources to process complex sentences. In control participants, adaptation increases the available resources such that difficulties in processing complex sentences decrease leading to smaller syntactic effects. Due to smaller adaptation or its absence in IWA, the available resources remain the same despite repeated exposure. This concept of adaptation should be studied more thoroughly in future studies.

With respect to practical implications for assessment and treatment in aphasia, our study revealed that despite possible differences in task demands both object manipulation and the two variants of sentence-picture matching were equally suitable to detect canonicity and interference effects in language impaired participants. However, a minimum of 60 baseline and 60 critical sentences was needed to gain a conclusive estimate of the size of syntactic effects in a single participant. With respect to the mode of presentation in the auditory input, self-paced presentation as opposed to normal speech rate did not lead to a decrease in syntactic effects, a finding which could be relevant for treatment in IWA. Finally, the mere repetition of sentences across sessions (six in our case) did not lead to a reduction in the difficulties with complex sentences in IWA. Thus, whether an even larger number of repetitions or a specific intervention focusing on structurally complex sentences leads to a decrease in syntactic effects remains an open issue.

4.5 Conclusion Study 1

This is the first data-set in German that provides a comprehensive evaluation of between- and within-participant variability in individuals with aphasia and a control group, spanning multiple syntactic constructions, and systematically evaluating the consistence of canonicity and interference effects between different response tasks and test phases. From a theoretical point of view our dataset is important in different respects. First, it provides important insights into the nature of variability in sentence comprehension and second, it fosters the development of computational models (e.g., Mätzig et al., 2018) and allows for quantitative evaluation of competing accounts of sentence processing in aphasia (e.g., Lissón et al., 2021). With respect to the nature of variability in sentence comprehension, our study demonstrated variability in the size of canonicity and interference effects both for language impaired and unimpaired participants. However, variability in control participants was systematic and led to a decrease in the effect sizes due to adaptation whereas in individuals with aphasia variability led to unsystematic changes in the size of the canonicity and interference effects over time or response tasks. The persistent appearance of canonicity and interference effects, however, shows that the performance is systematically influenced by syntactic complexity.

Acknowledgements We would like to thank all participants who volunteered to contribute to this study as well as three anonymous reviewers for their constructive comments. We are grateful to Andreas Schmidt for helpful discussions. We would also like to thank Silke Böttger, Sarah Düring and Therese Mayr for assisting with data collection. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317633480 – SFB 1287, project B02 (PIs: Shravan Vasishth, Frank Burchert, and Nicole Stadie).

A1 Appendix Study 1

A1.1 Sentence stimuli

Declarative sentences

The manipulated determiners (*nominative / accusative*) are presented in italics.

1. Hier badet *der / den* Esel gerade *den / der* Tiger.
   Here *the (nom /acc)* donkey just bathes *the (acc /nom)* tiger.

2. Hier zeichnet *der / den* Büffel gerade *den / der* Panther.
   Here *the (nom /acc)* buffalo just draws *the (acc /nom)* panther.

3. Hier kitzelt *der / den* Hamster gerade *den / der* Igel.
   Here *the (nom /acc)* hamster just tickles *the (acc /nom)* hedgehog.

4. Hier rettet *der / den* Pudel gerade *den / der* Kater.
   Here *the (nom /acc)* poodle just rescues *the (acc /nom)* tomcat.

5. Hier bürstet *der / den* Kater gerade *den / der* Pudel.
   Here *the (nom /acc)* tomcat just brushes *the (acc /nom)* poodle.

6. Hier tröstet *der / den* Tiger gerade *den / der* Esel.
   Here *the (nom /acc)* donkey just comforts *the (acc /nom)* tiger.

7. Hier leitet *der / den* Panther gerade *den / der* Büffel.
   Here *the (nom /acc)* panther just guides *the (acc /nom)* buffalo.

8. Hier füttert *der / den* Igel gerade *den / der* Hamster.
   Here *the (nom /acc)* hedgehog just feeds *the (acc /nom)* hamster.

9. Hier findet *der / den* Eber gerade *den / der* Otter.
   Here *the (nom /acc)* boar just finds *the (acc /nom)* otter.

10. Hier streichelt *der / den* Otter gerade *den / der* Eber.
    Here *the (nom /acc)* otter just pets *the (acc /nom)* boar.

Relative clauses

The manipulated sentence onsets to get subject and object modifying relative clauses (*here is the / I see the*) and determiners to get subject and object relative clauses (*nominative / accusative*) are presented in italics. In the plural condition, the noun in the subclause was plural.

1. *Hier ist der / Ich seh den* Esel, *der / den den / der* Tiger gerade badet.
   Here is the / I see the donkey *who (nom / acc) the (nom / acc)* tiger just bathes.

2. *Hier ist der / Ich seh den* Büffel, *der / den den / der* Panther gerade zeichnet.
   Here is the / I see the buffalo *who (nom / acc) the (nom / acc)* panther just draws.

3. *Hier ist der / Ich seh den* Hamster, *der / den den / der* Igel gerade kitzelt.
   Here is the / I see the hamster *who (nom / acc) the (nom / acc)* hedgehog just tickles.

4. *Hier ist der / Ich seh den* Pudel, *der / den den / der* Kater gerade rettet.
   Here is the / I see the poodle *who (nom / acc) the (nom / acc)* tomcat just rescues.

5. *Hier ist der / Ich seh den* Kater, *der / den den / der* Pudel gerade bürstet.
   Here is the / I see the tomcat *who (nom / acc) the (nom / acc)* poodle just brushes.

6. *Hier ist der / Ich seh den* Tiger, *der / den den / der* Esel gerade tröstet.
   Here is the / I see the tiger *who (nom / acc) the (nom / acc)* donkey just comforts.

7. *Hier ist der / Ich seh den* Panther, *der / den den / der* Büffel leitet.
   Here is the / I see the panther *who (nom / acc) the (nom / acc)* buffalo just guides.

8. *Hier ist der / Ich seh den* Igel, *der / den den / der* Hamster gerade füttert.
   Here is the / I see the hedgehog *who (nom / acc) the (nom / acc)* hamster just feeds.

9. *Hier ist der / Ich seh den* Eber, *der / den den / der* Otter gerade findet.
   Here is the / I see the boar *who (nom / acc) the (nom / acc)* otter just finds.

10. *Hier ist der / Ich seh den* Otter, *der / den den / der* Eber gerade streichelt.
    Here is the / I see the otter *who (nom / acc) the (nom / acc)* boar just pets.

Sentences with PRO

The manipulated verb (*subject control / object control*) is presented in italics.

1. Peter *verspricht / erlaubt* nun Lisa, das kleine Lamm zu streicheln und zu kraulen.
   Peter now *promises / allows* Lisa to pet and to ruffle the little lamb.

2. Thomas *versichert / gestattet* nun Anna, das dicke Rind zu melken und zu hüten.
   Thomas now *assures / allows* Anna to milk and to tend the thick cattle.

3. Thomas *droht / befielt* nun Lisa, das schnelle Huhn zu jagen und zu fangen.
   Thomas now *threatens / commands* Lisa to chase and to catch the fast chicken.

4. Peter *garantiert / empfiehlt* nun Anna, das stolze Ross zu bürsten und zu striegeln.
   Peter *guarantees / recommends* now Anna to brush and to comb the proud steed.

5. Thomas *schwört / rät* nun Anna, das süße Ferkel zu waschen und zu säubern.
   Thomas now *swears / advises* Anna to wash and to clean the sweet piglet.

6. Lisa *verspricht / erlaubt* nun Peter, das alte Schaf zu impfen und zu pflegen.
   Lisa now *promises / allows* Peter to vaccinate and to nurse the old sheep.

7. Anna *versichert / gestattet* nun Thomas, das junge Kalb zu malen und zu zeichnen.
   Anna now *assures / allows* Thomas to paint and to draw the young calf.

8. Anna *droht / befielt* nun Peter, das kluge Schwein zu füttern und zu mästen.
   Anna now *threatens / commands* Peter to feed and to fatten the clever pig.

9. Lisa *garantiert / empfiehlt* nun Thomas, das scheue Reh zu locken und zu suchen.
   Lisa now *guarantees / recommends* Thomas to lure and to search the shy deer.

10. Lisa *schwört / rät* nun Peter, das schöne Pferd zu satteln und zu zäumen.
    Lisa now *swears / advises* Peter to saddle and to bridle the nice horse.

Sentences with a pronoun

The manipulated noun (*same gender / different gender*) is presented in italics.

1. Peter verspricht nun *Thomas / Lisa*, dass er das kleine Lamm streichelt und krault.
   Peter now promises *Thomas / Lisa* that he will pet and ruffle the little lamb.

2. Thomas versichert nun *Peter / Anna*, dass er das dicke Rind melkt und hütet.
   Thomas now assures *Peter / Anna* that he will milk and tend the thick cattle.

3. Thomas droht nun *Peter / Lisa*, dass er das schnelle Huhn jagt und fängt
   Thomas now threatens *Peter / Lisa* that he will chase and catch the fast chicken.

4. Peter garantiert nun *Thomas / Anna*, dass er das stolze Ross bürstet und striegelt.
   Peter guarantees now *Thomas / Anna* that he will brush and comb the proud steed.

5. Thomas schwört nun *Peter / Anna*, dass er das süße Ferkel wäscht und säubert.
   Thomas now swears *Peter / Anna* that he will wash and clean the sweet piglet.

6. Lisa verspricht nun *Anna / Peter*, dass sie das alte Schaf impft und pflegt.
   Lisa now promises *Anna / Peter* that she will vaccinate and nurse the old sheep.

7. Anna versichert nun *Lisa / Thomas*, dass sie das junge Kalb malt und zeichnet.
   Anna now assures *Lisa / Thomas* that she will paint and draw the young calf.

8. Anna droht nun *Lisa / Peter*, dass sie das kluge Schwein füttert und mästet.
   Anna now threatens *Lisa / Peter* that she will feed and fatten the clever pig.

9. Lisa garantiert nun *Anna / Thomas*, dass sie das scheue Reh lockt und sucht.
   Lisa now guarantees *Anna / Thomas* that she will lure and search the shy deer.

10. Lisa schwört nun *Anna / Peter*, dass sie das schöne Pferd sattelt und zäumt.
    Lisa now swears *Anna / Peter* that she will saddle and bridle the nice horse.

A1.2 Contrast coding

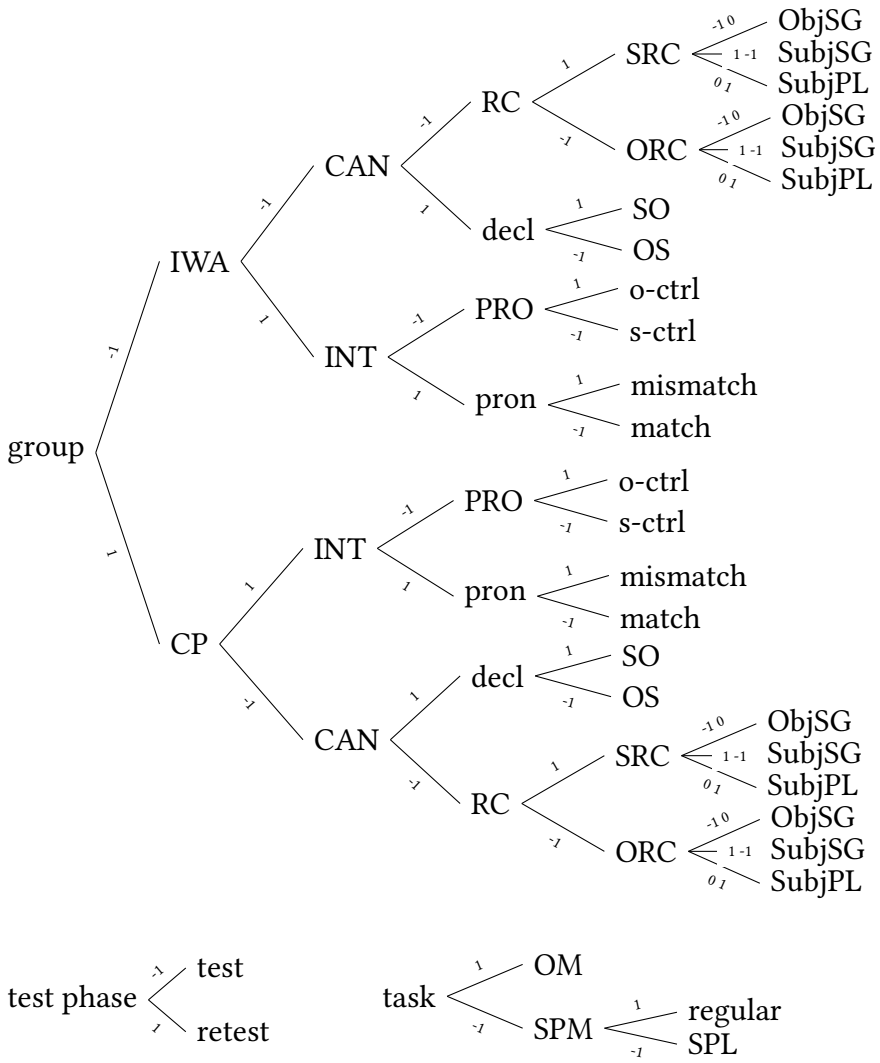

Figure 14: Visualization of the nestings and contrast codings of the model. CP = control particiapnts, IWA = individuals with aphasia, INT = interference, CAN = canonicity, pron = pronoun, decl = declarative, RC = relative clause, SRC/ORC = subject/object relative, SO/OS = subject-before-object/object-before-subject, s/o-ctrl = subject/object control, Subj/Obj = subject/object modifying relative clause, SG = singular, PL = plural, OM = object manipulation, SPM = sentence-picture matching, SPL = self-paced listening.

A1.3 Descriptive statistics

Table 4: Accuracy and response times across three tasks and two test sessions in individuals with aphasia and control participants.

		Canonicity Experiment				Interference experiment			
		SO	OS	SRC	ORC	mis-match	match	o-ctrl	s-ctrl
Accuracy									
IWA	Mean	75.0	43.3	66.9	46.6	70.3	60.4	75.5	60.3
	SE	1.2	1.4	0.8	0.8	1.3	1.4	1.2	1.4
CP	Mean	98.9	95.6	96.9	97.1	99.8	97.9	99.2	98.2
	SE	0.2	0.4	0.2	0.2	0.1	0.3	0.2	0.2
response time									
IWA	Mean	5192.0	6226.6	5039.4	5201.0	3144.6	3566.2	3073.5	3311.6
	SE	104.3	133.2	69.7	67.4	95.2	109.4	95.9	101.9
CP	Mean	1618.3	1906.7	1740.2	1805.7	1343.4	1449.5	1322.5	1337.9
	SE	16.6	22.8	13.6	13.4	16.3	17.3	18.1	13.1

Note. IWA = individuals with aphasia, CP = control participants, SO/OS = canonical/non-canonical declarative sentence, SRC/ORC = subject/object relative clause, match/mismatch = gender of the main clause nouns is the same/different, s-ctrl/o-ctrl = subject/object control.

A1.4 Correlation coefficients of the Bayesian models and intraclass correlation coefficients

Table 5: Bayesian correlation estimates and intraclass correlation coefficients of canonicty and interference effects in individuals with aphasia and control participants.

	Bayesian correlation estimate	intraclass correlation coefficient	F-value	df1	df2	p-value	lower bound	upper bound
Accuracy control group								
Decl RegxSPL	0.03 CrI: [-0.6, 0.64]	0.12	1.36	49	47	0.148	-0.10	0.36
Decl OMxSPM	0.15 CrI: [-0.51, 0.7]	0.46	2.71	49	50	0.000	0.21	0.65
Decl TP1xTP2	0.11 CrI: [-0.52, 0.7]	0.47	2.92	49	45	0.000	0.22	0.66
PRO RegxSPL	-0.05 CrI: [-0.62, 0.54]	0.08	1.18	49	49	0.285	-0.20	0.35
PRO OMxSPM	0.08 CrI: [-0.58, 0.68]	0.18	1.47	49	50	0.091	-0.09	0.43
PRO TP1xTP2	0.16 CrI: [-0.49, 0.71]	0.45	2.71	49	49	0.000	0.20	0.64
pron RegxSPL	-0.03 CrI: [-0.63, 0.61]	0.04	1.08	49	50	0.389	-0.21	0.29
pron OMxSPM	0.05 CrI: [-0.6, 0.66]	0.28	1.83	49	48	0.019	0.02	0.51
pron TP1xTP2	0.03 CrI: [-0.62, 0.63]	0.22	1.62	49	50	0.047	-0.04	0.46
RC RegxSPL	0.24 CrI: [-0.28, 0.67]	0.26	1.87	49	40	0.022	0.01	0.49
RC OMxSPM	-0.03 CrI: [-0.46, 0.42]	0.36	2.12	49	49	0.005	0.09	0.58
RC TP1xTP2	0.4 CrI: [-0.06, 0.78]	0.39	2.31	49	50	0.002	0.14	0.60
Accuracy IWA								
Decl RegxSPL	0.39 CrI: [-0.09, 0.77]	0.48	2.90	20	21	0.010	0.09	0.75
Decl OMxSPM	0.33 CrI: [-0.12, 0.72]	0.40	2.28	20	20	0.036	-0.04	0.71
Decl TP1xTP2	0.43 CrI: [-0.04, 0.79]	0.50	3.03	20	21	0.008	0.11	0.76
PRO RegxSPL	0.34 CrI: [-0.1, 0.69]	0.47	2.86	20	21	0.010	0.08	0.74
PRO OMxSPM	0.47 CrI: [0.07, 0.78]	0.59	4.02	20	20	0.001	0.23	0.81
PRO TP1xTP2	0.46 CrI: [0.04, 0.79]	0.65	4.68	20	21	0.000	0.32	0.84
pron RegxSPL	0.21 CrI: [-0.42, 0.72]	0.32	1.93	20	20	0.074	-0.12	0.66
pron OMxSPM	0.37 CrI: [-0.15, 0.77]	0.47	2.71	20	20	0.015	0.05	0.75
pron TP1xTP2	0.49 CrI: [-0.03, 0.85]	0.68	5.09	20	21	0.000	0.36	0.86
RC RegxSPL	0.78 CrI: [0.52, 0.93]	0.86	12.48	20	20	0.000	0.68	0.94
RC OMxSPM	0.62 CrI: [0.28, 0.85]	0.68	7.20	20	9	0.003	0.24	0.87
RC TP1xTP2	0.58 CrI: [0.23, 0.82]	0.71	5.59	20	20	0.000	0.40	0.87
Response times control group								
Decl RegxSPL	0.33 CrI: [-0.12, 0.72]	0.19	1.74	49	24	0.070	-0.06	0.43
Decl TP1xTP2	0.29 CrI: [-0.23, 0.72]	0.23	1.72	49	44	0.036	-0.02	0.47
PRO RegxSPL	-0.01 CrI: [-0.63, 0.62]	0.11	1.23	49	49	0.235	-0.18	0.37
PRO TP1xTP2	0.29 CrI: [-0.33, 0.77]	0.09	1.21	49	49	0.257	-0.19	0.36
pron RegxSPL	0.21 CrI: [-0.49, 0.77]	0.47	2.79	49	50	0.000	0.23	0.66
pron TP1xTP2	0.23 CrI: [-0.45, 0.75]	0.22	1.63	49	49	0.046	-0.04	0.46
RC RegxSPL	0.65 CrI: [0.35, 0.87]	0.68	5.40	49	46	0.000	0.49	0.80
RC TP1xTP2	0.84 CrI: [0.62, 0.96]	0.80	10.29	49	31	0.000	0.65	0.89
Response times IWA								
Decl RegxSPL	0.49 CrI: [-0.08, 0.87]	0.53	3.34	20	21	0.004	0.15	0.77
Decl TP1xTP2	0.58 CrI: [0.08, 0.91]	0.71	5.66	20	20	0.000	0.40	0.87
PRO RegxSPL	0.39 CrI: [-0.25, 0.83]	0.23	1.58	20	20	0.157	-0.22	0.59
PRO TP1xTP2	0.38 CrI: [-0.28, 0.83]	0.03	1.06	20	20	0.446	-0.42	0.46
pron RegxSPL	0.06 CrI: [-0.62, 0.71]	0.14	1.40	20	21	0.225	-0.20	0.49
pron TP1xTP2	0.05 CrI: [-0.64, 0.72]	0.04	1.07	20	20	0.438	-0.41	0.46
RC RegxSPL	0.25 CrI: [-0.26, 0.69]	0.04	1.09	20	20	0.427	-0.40	0.46
RC TP1xTP2	0.64 CrI: [0.17, 0.92]	0.67	5.26	20	21	0.000	0.36	0.85

Note. Decl = declarative, RC = relative clause, pron = pronoun, RegxSPL = correlation regular x self-paced sentence-picture matching, OMxSPM = correlation object manipulation x sentence-picture matching, TP1xTP2 = correlation test x retest, IWA = individuals with aphasia.

Study 2

Can the resource reduction hypothesis explain sentence processing in aphasia? A visual world study in German

Article published in
Brain & Language 235 (2022) 105204
Dorothea Pregla, Shravan Vasishth, Paula Lissón,
Nicole Stadie, and Frank Burchert

Abstract: Resource limitation has often been invoked as a key driver of sentence comprehension difficulty, in both theories of language-unimpaired and language-impaired populations. In the field of aphasia, one such influential theory is Caplan's resource reduction hypothesis (RRH). In this large investigation of online processing in aphasia in German, we evaluated three key predictions of the RRH in 21 individuals with aphasia and 22 control participants. Measures of online processing were obtained by combining a sentence-picture matching task with the visual world paradigm. Four sentence types were used to investigate the generality of the findings, and two test phases were used to investigate RRH's predictions regarding variability in aphasia. The processing patterns were consistent with two of the three predictions of the RRH. Overall, our investigation shows that the RRH can account for important aspects of sentence processing in aphasia.

1 Introduction Study 2

In sentence processing research, it is well-established that limitations in resource capacity can affect sentence comprehension (Just & Carpenter, 1992). The idea of a limited resource capacity has also been implemented to explain the performance of individuals with aphasia (IWA) in sentence comprehension tasks, e.g., sentence-picture matching (Caplan, 2012; Miyake et al., 1994). The resource reduction approach predicts the following performance pattern for IWA: Resource reduction should impair sentence comprehension across different types of sentence structures (e.g., relative clauses, or sentences with pronouns, Caplan & Hildebrandt, 1988; Caplan et al., 2015). Furthermore, resource reduction should generate a variable impairment in sentence comprehension depending on the amount of available resources. These predictions can be tested experimentally by comparing comprehension performance of the same IWA across different tasks and sentence structures. This approach has been taken by Caplan et al. (2006), Caplan et al. (2015, 2013a), Caplan et al. (2007) for English and more recently by Pregla et al. (2021) for German. The tasks consisted in different versions of sentence-picture matching (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Pregla et al., 2021), grammaticality judgement (Caplan et al., 2006; Caplan et al., 2007), and object manipulation (Caplan et al., 2006; Caplan et al., 2013a; Caplan et al., 2007; Pregla et al., 2021). These studies showed that IWA had a variable degree of difficulty comprehending the same sentence structures in different tasks (Caplan et al., 2006; Caplan et al., 2015, 2013a; Caplan et al., 2007; Pregla et al., 2021). Furthermore, comprehension difficulty was not restricted to a specific sentence structure but affected complex sentences in general. Both the variability in performance and the general impairment for complex sentences speak for the view that the sentence comprehension impairment seen in IWA is brought about by resource reduction.

This paper will examine the resource reduction approach more closely. More specifically, this paper will investigate one influential instantiation of this approach, the *resource reduction hypothesis* (RRH, e.g., Caplan, 2012; Caplan et al., 2006; Caplan et al., 2015; Caplan et al., 2007). Below, we introduce the RRH, and examine whether previous findings relating to online sentence processing in aphasia are consistent with this account.

1.1 The resource reduction hypothesis

According to the RRH, sentence comprehension depends both on the resource capacity of a participant and the amount of resources a particular sentence comprehension task demands from the available resources of that given participant (Caplan, 2012). Although the exact nature of the resource capacity is not defined in the RRH, Caplan et al. (2013a) enumerate a number of different resource types that might cause sentence comprehen-

sion impairments in case they do not function as expected. The authors suggest that sentence comprehension impairments might for example be caused by a reduced processing efficiency, by "slowed or otherwise disrupted activation of lexical representations", by "a more general disturbance affecting skilled performances, such as an across-the-board slowing of processing speed", by a reduction in "operations that are needed to perform a task to which comprehension is directed, such as planning actions, inspecting pictures, etc.", or by a reduction in "executive functions in the form of deployment of attention, maintenance of task goals, uploading mechanisms that support task performance [...], executing those mechanisms, response selection, assessment of success on a trial, and other processes." (Caplan et al., 2013a, p. 28–29). While the RRH is undetermined with respect to what type of resource is affected, the RRH assumes that the capacity of this resource is reduced in IWA in comparison to control participants. Furthermore, the RRH assumes that the resource capacity is subject to random fluctuations caused by noise inherent to a participant. This means that the resources in the processing system can vary from participant to participant and in the same participant from moment to moment. This fluctuation is assumed to be larger in IWA than in control participants. The resource demands depend on the complexity of a task and are stable. Task complexity can be determined by the average performance of IWA or control participants in a sentence comprehension task, i.e., tasks that are difficult for a group of participants are said to be complex (Caplan, 2012). The RRH assumes that tasks with high complexity impose greater resource demands than tasks with lower complexity.

Figure 15 illustrates the interplay between the resource capacity inherent to participants (solid lines) and task demands (broken lines) according to the RRH. The figure displays the randomly fluctuating resource capacity of IWA (black) and control participants (grey) over an arbitrary period of time. When the resources of a given participant meet the task demands, sentence processing proceeds in a normal-like fashion, resulting in a correct response. However, if the task demands exceed the available resources of a given participant, sentence processing is impaired, resulting in an incorrect response. According to the RRH, processing is more impaired in complex sentences (e.g., object relative clauses) than in simple sentences (e.g., subject relative clauses) because the resource demands of complex sentences are more likely to exceed the participant's resource capacity. However, since noise randomly affects resource capacity, the resources of the participant can sometimes be high enough to process a complex sentence correctly, and sometimes too low to even process a simple sentence correctly. From these assumptions of the RRH, we derived the novel prediction that performance in sentence comprehension tasks should be variable both within sessions and between sessions because of the noise that randomly affects processing in IWA.

The RRH explains the offline comprehension performance of IWA, but it also makes predictions regarding the online processing mechanisms in sentence comprehen-

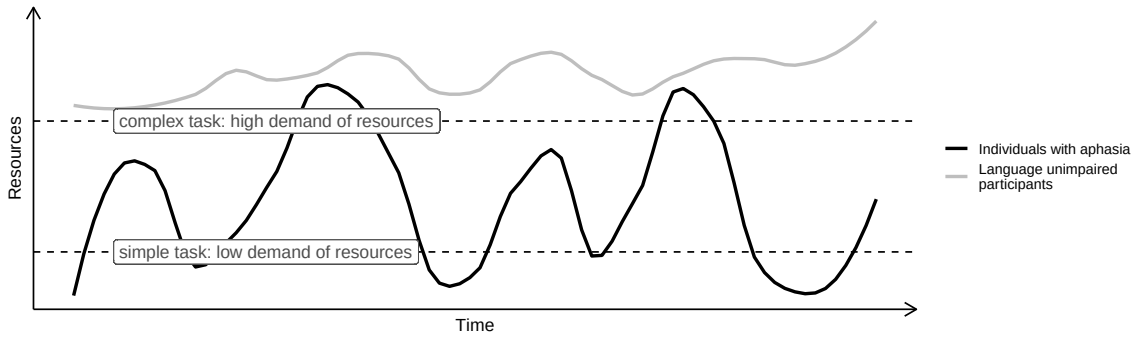

Figure 15: A schematic illustration of the assumed fluctuation of resources according to the resource reduction hypothesis. Solid lines represent the resource capacity of language impaired participants (black) and language unimpaired control participants (grey). Resources randomly fluctuate over arbitrary units of time due to noise in the comprehension system of the participant. Dashed lines represent the resource demand of a simple task (low demand) and of a complex task (high demand). Processing is impaired if the task demand exceeds the resource capacity, otherwise processing is normal.

sion in aphasia. So far, the RRH's predictions regarding online performance have only been investigated with the self-paced listening paradigm (Caplan et al., 2015; Caplan et al., 2007). In the present study, the RRH's predictions were investigated using the visual world paradigm (Cooper, 1974). In this paradigm, participants are simultaneously presented with pictures on a visual display and auditory speech while their proportion of fixations to each picture is recorded. The visual world paradigm is well-established as a means to study sentence processing as it unfolds (Tanenhaus et al., 1995). Next, the predictions of the RRH regarding online processing in aphasia will be presented and it is shown whether previous results are consistent with the predictions. Since the RRH is undetermined with respect to what type of resource is affected in IWA, three options were taken into account that are all discussed in Caplan et al. (2015), namely random fluctuations in resources leading to intermittent deficiencies, slowed processing speed, and syntactic proficiency as expressed by sentence comprehension accuracy.

1.1.1 Prediction 1: Normal-like processing in correct trials

The RRH predicts that correct responses in sentence comprehension should result predominantly from normal syntactic processing (as opposed to accidentally correct responses because of guessing in every trial) while incorrect responses should result from impaired syntactic processing.[1] In line with this prediction, Caplan et al. (2007) found that the self-paced listening times in IWA differed between correct and incorrect trials, and that the listening times were qualitatively similar to control participants in correct trials. Visual world studies also found that the proportion of fixations to the target pic-

[1] However, Caplan et al. (2015) point out that severely impaired IWA could be indeed guessing when they answer correctly. Caplan et al. (2015) define severely impaired IWA as those with accuracies below chance level in sentence-picture matching.

ture (henceforth target fixations) differed between correct and incorrect trials in IWA (Arantzeta et al., 2017; Choy & Thompson, 2010; Dickey et al., 2007; Dickey & Thompson, 2009; Hanne et al., 2012, 2015; Hanne et al., 2016; Hanne et al., 2011; Meyer et al., 2012). In correct trials, target fixations of IWA and control participants were qualitatively similar (Arantzeta et al., 2017; Dickey et al., 2007; Dickey & Thompson, 2009; Hanne et al., 2015; Hanne et al., 2016; Hanne et al., 2011; Meyer et al., 2012). Thus, as predicted by the RRH, syntactic processing in IWA tends to proceed normal-like in correct trials.

Besides the normal-like pattern, a number of visual world studies reported delayed target fixations in IWA in comparison to control participants in correct trials (Hanne et al., 2016; Hanne et al., 2011; Meyer et al., 2012; Schumacher et al., 2015). These delays were ascribed to a processing slowdown (Hanne et al., 2016; Hanne et al., 2011; Meyer et al., 2012). The RRH does not make a prediction about processing speed. However, it can account for slowed processing under the assumption that the reduced capacity is processing speed, which Caplan et al. (2015) consider likely. Thus, the finding that syntactic processing in IWA in correct trials seems to proceed normal-like but slower than in control participants is compatible with the RRH.

1.1.2 Prediction 2: Processing difficulty in complex vs. simple sentences, and a complexity-capacity interaction

The RRH predicts processing differences between syntactically simple and complex sentences. In line with this prediction, Caplan et al. (2007) and Caplan et al. (2013a) found complexity effects in the form of lower accuracy scores and slower response times for syntactically complex versus simple sentences. Additionally, the RRH predicts a super-additive interaction of resource capacity and resource demands, i.e., increased demands should affect participants with a lower capacity level far more than participants with a higher capacity level (e.g., Caplan et al., 2007). Caplan et al. (2007) and Caplan et al. (2015) investigated this prediction in two self-paced listening experiments. As a measure of resource capacity, the authors used the accuracy of each IWA in non-canonical sentences.[2] As a measure of task complexity, the authors used the listening times in simple and complex sentences. In line with the RRH, Caplan et al. (2007) and Caplan et al. (2015) found a super-additive effect, i.e., the difference in listening times between simple and complex sentences was larger for IWA with lower accuracy.

A number of visual world studies have investigated the influence of sentence complexity on fixations to a target picture (Hanne et al., 2015; Mack et al., 2016; Meyer et al., 2012; Sheppard et al., 2015). Both IWA and control participants showed more target fixations in simple canonical versus complex non-canonical sentences (Hanne et al.,

[2]To determine capacity, the authors used the accuracy across sentence types (including passives, object clefts and object relative clauses Caplan et al., 2007) or the accuracy for each separate sentence type (for passives, object relative clauses, reflexives and pronouns Caplan et al., 2015).

2015; Mack et al., 2016; Meyer et al., 2012). However, participant groups differed in the sentence region where complexity influenced the target fixations. Control participants showed increased target fixations in canonical versus non-canonical sentences before the region that disambiguated the sentence's reading (e.g., *The man was*). The differences vanished directly after disambiguation (e.g., *shaving / shaved by the boy*). This fixation behavior was interpreted as an agent-first processing pattern, i.e., a tendency to process the first noun of a sentence as the agent followed by a revision in non-canonical sentences (Hanne et al., 2015; Mack et al., 2016; Meyer et al., 2012). In contrast to control participants, IWA showed increased target fixations in canonical versus non-canonical sentences only after the disambiguating region (Hanne et al., 2015; Mack & Thompson, 2017; Mack et al., 2016; Meyer et al., 2012). The fixation pattern in IWA is compatible with the assumption of the RRH that the processing difficulty is larger in complex versus simple sentences but arises more slowly than in control participants.

1.1.3 Prediction 3: Unsystematic variability in the performance between test and retest

The RRH predicts that sentence comprehension varies unsystematically over time within the same IWA. This is because noise should randomly affect the resources available for sentence processing. Little is known about the nature of this variability in online processing. Only one visual world study (Mack et al., 2016) has investigated this issue so far. Mack et al. (2016) tested the processing of active and passive sentences (*The man visited the woman / was visited by the woman*) in a group of 12 IWA and 21 control participants in two sessions spaced one week apart. The authors investigated the test-retest reliability of the eye-tracking measures and found that the reliability was generally strong in IWA (intraclass correlation coefficients between 0.59 and 0.75), and overall stronger in the IWA than in the control participants. Therefore, Mack et al. (2016) concluded that eye-tracking measures can be reliably used to investigate changes over time in the performance of IWA. Furthermore, the authors investigated the intra-individual variability in the eye-tracking measures and observed that it did not differ between the language-impaired and language-unimpaired groups. Thus, Mack et al. (2016) tentatively suggest that day-to-day variability in online sentence processing is not larger in IWA than in language-unimpaired individuals. Finally, both participant groups showed increased target fixations in the second compared to the first session independent of the sentence type. Mack et al. (2016) interpreted the increase in target fixations in the retest as a practice effect. The practice effect might indicate that variable processing between sessions is not just random fluctuation but reflects systematic changes. Such changes are currently not accounted for by the RRH.

1.2 Aim of the study

The present study aimed to investigate the RRH's predictions regarding sentence processing in IWA. To this end, the visual world paradigm was used. This paradigm allows us to investigate automatic processing during auditory sentence presentation. The paradigm has the advantage over other online paradigms (e.g., self-paced listening, or cross-modal priming) that it is easy to carry out for IWA (Dickey et al., 2007). Furthermore, the paradigm offers more direct information on syntactic processing than offline analyses, because the data are gathered during sentence presentation and thus can reveal how participants arrive at a sentence interpretation (Dickey et al., 2007). Offline responses also require additional conscious processes that might be impaired in IWA making it difficult to draw conclusions about underlying processing abilities (Caplan et al., 2013a). Therefore, the visual world paradigm is suitable to test the predictions of the RRH regarding processing in IWA.

Our experimental design was unique in that sentence processing was investigated across two test phases and four sentence types. This design was chosen to assess the fluctuation in sentence processing in IWA. Furthermore, our study included a relatively large group of 21 IWA. According to a review by Sharma et al. (2021) including 13 visual world studies on sentence comprehension in aphasia, the average number of participants amounts to less than ten IWA (mean = 9 IWA, range = 4 to 16 IWA).[3] Furthermore, our study tested sentence comprehension in German while previous studies investigating the RRH focused on English (Caplan et al., 2015, 2013a; Caplan et al., 2007). Given that the RRH is presumably a language-independent theory, it is vital to test its predictions in other languages. There are several reasons why it is interesting to investigate German. In comparison to English, German has a relatively free word order. Furthermore, German allows disambiguating thematic roles based on case marking. Therefore, word order complexity can be varied based on minimal changes in case marking. To our knowledge, our study is the first comprehensive investigation of the RRH for German, and the first to use the visual world paradigm for this purpose.

The following predictions with respect to the target fixations in aphasia were derived from the RRH:

Fixation patterns derived from prediction 1: Normal-like processing in correct trials. In correct trials, the target fixations of the IWA should be similar to those of control participants. That is, both participant groups should show increases in target fixations over the course of a trial (i.e., increases relative to the beginning of a trial where the proportion of target fixations should be 50%). However, target fixations might increase more slowly in IWA than in control participants, as observed in

[3] Sharma et al. (2021) report a range of 4 to 19 IWA because they did not account for the fact that the study with 19 IWA (Barbieri et al., 2019) had to exclude 3 IWA. This results in the upper bound of 16 IWA reported here.

previous visual world experiments (Hanne et al., 2016; Hanne et al., 2011; Meyer et al., 2012; Schumacher et al., 2015). The RRH would be compatible with slow increases in target fixations in IWA because the reduced capacity likely is processing speed (Caplan et al., 2015). Furthermore, increases in target fixations should be higher in correct versus incorrect trials.

Fixation patterns derived from prediction 2: Processing difficulty in complex vs. simple sentences, and a complexity-capacity interaction. Target fixations should diverge between simple and complex sentences, and the increase in target fixations should be higher in simple sentences. Furthermore, the RRH predicts a super-additive interaction between resource demands and resource capacity. Following Caplan et al. (2007) and Caplan et al. (2015), IWA with a lower overall accuracy are assumed to have a lower resource capacity, thus, they should show a more pronounced complexity effect. Consequently, if the overall accuracy of the IWA decreases, the difference in target fixations between simple and complex sentences should increase.

Fixation patterns derived from prediction 3: Unsystematic variability in the performance between test and retest. Following the RRH, fixation paths should vary randomly between the test and retest phase in IWA. That is, target fixations should not systematically increase faster over the course of a trial in the retest than in the test, as would be expected if practice effects were present (Mack et al., 2016).

2 Methods and Material Study 2

This visual world experiment investigated the processing of declarative sentences (henceforth declaratives), relative clauses and subject and object control structures (henceforth control structures) with an overt pronoun or a covert pronoun (henceforth PRO) in German in language-unimpaired control participants and IWA. In what follows, the specifics of the methods and materials are explained.

2.1 Participants

Overall, 43 participants, all native speakers of German completed the study: 21 IWA (9 females, mean age = 60 years, $SD = 11$, range = 38–78; mean education = 15 years, $SD = 3$, range = 8–22) and 22 age- and education-matched control participants (14 females, mean age = 58 years, $SD = 15$, range = 26–81; mean education = 16 years, $SD = 4$, range = 6–21). All participants had normal or corrected-to-normal hearing and vision. Only control participants without known neurological disorders or language impairments were included. Inclusion criteria for IWA were the presence of chronic aphasia (>12 months post onset), no upper limb apraxia, and intact comprehension of nouns. Aphasia had to be apparent according to the Aachen Aphasia Test (Huber et al.,

1983).[4] Participants gave written consent in accordance with the ethics committee of the University of Potsdam and were paid for participation.

Control participants were recruited from the University of Potsdam and from a church parish. All control participants were right-handed as assessed by the Edinburgh Handedness Inventory (Oldfield et al., 1971). Control participants were screened for dementia using the Montreal Cognitive Assessment (MoCA, Nasreddine et al., 2005) and all participants were in the normal range, i.e., they scored at least 26/30 points (mean = 29 points, $SD = 1$, range = 26–30). Originally, data from 50 control participants were gathered. For age and education matching, 28 control participants were excluded prior to the analyses. Figure 22 in the appendix shows that the fixation paths of the 50 and the 22 control participants are qualitatively similar for all sentence types. Five additional control participants were excluded prior to the analyses because of neurological impairments (3 participants), or because they did not complete all tasks (2 participants).

IWA were recruited from a database of the University of Potsdam and from aphasia self-help groups in Potsdam and Berlin. Demographic and neurological information about the IWA is summarized in Table 6. All but one participant experienced a single stroke at least one year prior to the study. All but three participants were right-handed pre-morbidly as assessed by the Edinburgh Handedness Inventory (Oldfield et al., 1971). The Aachen Aphasia Test (Huber et al., 1983) was administered to determine the type and severity of aphasia (see Table 6). All IWA showed good auditory processing abilities for single nouns, assessed with an auditory word-picture matching task (all scores at least 90% correct) and a lexical decision task (all scores at least 88% correct) of the German psycholinguistic test battery LEMO 2.0 (Stadie et al., 2013). Although accuracy in the lexical decision task was lower in IWA compared to the control group, both participant groups were similarly influenced by psycholinguistic variables: Both groups gave faster responses for words than for non-words (lexicality effect), for high-frequency than for low-frequency words (frequency effect), and for concrete than for abstract words (effect of abstractness). Six additional IWA were excluded prior to data analysis due to no apparent aphasia in the Aachen Aphasia Test (3 participants), less than 90% accuracy in auditory word-picture matching (2 participants), or withdrawal (1 participant).

[4] We did not exclude IWA with certain types of aphasia. It has been hypothesized that sentence complexity specifically affects people with Broca's aphasia (e.g., Drai & Grodzinsky, 1999). However, other authors (e.g., Caplan et al., 2015; Luzzatti et al., 2001) did not confirm a generalization of such a comprehension pattern to all people with Broca's aphasia and found a similar influence of sentence complexity on comprehension performance in individuals with different aphasia types. Therefore, we decided not to restrict our sample to people with a specific type of aphasia.

Table 6: Demographic and neurological data of the individuals with aphasia.

IWA	Gender	Years Age	Years Education	Years P.O.	Etiology	Localization	LEMO[1] (raw scores) T3 (n=80)	LEMO[1] (raw scores) T11 (n=20)	AAT[2] Aphasia type	AAT[2] Severity (standard nine)
2	F	72	8	7	IMI	L	77	19	Anomic	6.8 (mild)
3	M	76	20	17	IMI	L/R	61	20	Not-classifiable	7 (mild)
4	F	47	13	21	IMI	L	78	20	Anomic	7.8 (mild)
6	M	55	14	10	IMI	L	67	20	Anomic	6.8 (mild)
8	F	51	19	7	MA	L	74	20	Anomic	7.4 (mild)
9	M	64	15	2	IMI	L	73	20	Anomic	7.4 (mild)
10	M	58	18	1	IMI	L	52	20	Broca	5 (moderate)
11	F	63	12	1	IMI	L	73	20	Broca	6.8 (mild)
12	F	46	12	13	IMI	L	65	20	Broca	4.2 (moderate)
13	M	74	13	8	IMI	L	57	20	Broca	4.4 (moderate)
14	M	66	13	17	IMI	L	75	20	Anomic	6.4 (mild)
15	F	59	21	4	I	L	77	20	Broca	5.2 (moderate)
16	M	67	17	26	VH	R	72	19	Broca	5.4 (moderate)
17	F	43	14	10	IMI	L	65	20	Broca	6.6 (mild)
18	M	57	13	1	I	L	67	18	Wernicke	not available
19	F	52	19	8	IMI	L	76	20	Broca	5.8 (moderate)
20	M	38	13	3	IMI	L	73	19	Broca	4.2 (moderate)
21	M	57	18	2	IMI	L	66	18	Broca	6 (mild)
22	F	67	16	5	IMI	L	76	20	Anomic	6.2 (mild)
23	M	74	15	7	IMI	L	67	20	Anomic	6.6 (mild)
24	M	78	15	6	IMI	L	not available	19	Wernicke	5.6 (moderate)

Note. IWA = individual with aphasia, F = female, M = male, P.O. = post onset, IMI = ischemic arteria cerebri media infarct, I = infarct, MA = arteria cerebri media aneurysm, VH = vertebrobasilar hemorrhage, L = left, R = right, [1] LEMO 2.0 (Stadie et al., 2013) T3 = auditory lexical decision, T11 = auditory word-picture matching, [2] Aachen Aphasia Test (Huber et al., 1983).

2.2 Procedure

This visual world experiment was part of a larger number of experiments that were carried out in a pseudo-randomized order with the same participants. All experiments were administered twice, i.e., in a test and retest phase spaced approximately two months apart. The specifics of the overall structure of the study are provided in Pregla et al. (2021).

The visual world experiment had two parts. The first part investigated the comprehension of control structures (see part *control structures* in the *materials* section), and the second part investigated the comprehension of declaratives and relative clauses (see part *declaratives and relative clauses* in the *materials* section). The two parts were presented to participants in pseudo-randomized order. Both parts included five practice items for which feedback about response accuracy was provided, followed by the experimental items for which no feedback was provided. The part on control structures included one break after half of the items. The part on declaratives and relative clauses included breaks after each quarter of the items. Control participants and IWA completed the experiment in approximately 30 and 60 minutes respectively.

Prior to the experiment, participants were instructed that they were going to perform a sentence-picture matching task with two pictures and that their eye-movements would be recorded during the task. Items were presented in the following manner: 1) Preview of the pictures for 4000ms and introduction of the displayed characters with a short sentence presented auditorily (e.g., *Hier sind Lisa und Peter.* 'Here are Lisa and Peter.' or *Hier sind Tiger und Esel.* 'Here are tigers and donkeys.'), 2) display of a central fixation cross for 500ms, and 3), reappearance of the pictures and simultaneous auditory presentation of the experimental sentence. Pictures were shown until a picture was selected by the participant or for maximally 30 seconds. For picture selection, the lower left or right button on a Cedrus response pad (key layout RB-840) had to be pressed. In the experiment testing the comprehension of control structures, participants had to select the picture with the person (e.g., *Lisa*) that, according to the sentence, "does something with the animal". In the experiment testing the comprehension of declaratives and relative clauses, participants had to select the picture "that fits with the sentence" (see examples in Figure 16). None of the participants had difficulties understanding the task or responding using the response pad.

A SensoMotoric Instruments (SMI RED250mobile) eye-tracker (binocular eye-tracking, Experiment Center version 3.7, sampling rate 250 Hz) was used. Pictures were presented on a separate monitor (resolution: 1920×1080 pixels) on a grey screen with a distance of 60 pixels between the right border of the left picture and the left border of the right picture, which corresponded to a visual angle of 3°. Each picture subtended a visual angle of 37°. Participants were seated in front of the screen with a distance of

approximately 60 cm. No chin-rest was used but participants were instructed to sit still. A 5-point calibration and validation were carried out before the practice phase, the test phase, and the second half of the test phase. If necessary, calibration could be manually initiated during the experiment. Both eyes were recorded and the fixation locations were determined based on the mean x and y coordinates of the eyes. Blinks, saccades, and fixations were detected with the velocity based algorithm of the SMI software BeGaze (version 3.7). Temporarily adjacent samples that did not exceed a velocity of 40°/s for at least 50ms were treated as a fixation. Areas of interest (AoIs) consisted of the two pictures, and the number of fixations on the target picture (correct, counted as 1) in proportion to the fixations on the foil picture or no picture (counted as 0) was calculated. In the results section, the proportion of target fixations will be reported.

2.3 Materials

Below, the sentence structures, the auditory stimuli and the pictures will be presented.

2.3.1 Control structures

Examples for the sentences are given in Table 7 (for all items, see appendix). These sentences were used to test for the comprehension of subject and object control structures. In control structures, the subject of an embedded clause is identified with an argument of a matrix clause (Stiebels, 2007), i.e., the argument in the matrix clause controls the meaning of the subject in the embedded clause. Participants had to decide which of the arguments of the matrix clause, the subject or the object, controls the subject of the embedded clause. This decision depended on the matrix clause verb that either led to a subject control interpretation (e.g., *versprechen*, 'promise') or an object control interpretation (e.g., *erlauben*, 'allow'). The critical region of the sentence was the first phrase of the embedded clause. This phrase included the overt pronoun or PRO and thus was the point where the decision about the controlling argument should take place (highlighted in bold in Table 7).

A set of 50 control structures was used. In 20 sentences, the subject of the embedded clause was a pronoun controlled by the subject of the matrix clause (see Table 7, match and mismatch). In a further 20 sentences, the subject of the embedded clause was PRO, i.e., the pronoun was not pronounced overtly. PRO was controlled by the subject or the object in ten sentences respectively (see Table 7, s-ctrl and o-ctrl). Finally, ten filler sentences were included. Sentences were pseudo-randomized with at most three consecutive repetitions of the same sentence type.

To construct the sentences, 10 control verbs (5 subject control, 5 object control) with a mean lemma frequency of 4,713 ($SD = 2,146$) per million tokens in dlexDB (Heister et al., 2011) were used. In the sentences with PRO, control type was manipulated to

Table 7: Example of the control structures with PRO or an overt pronoun used in the experiment.

Condition	Sentence
s-ctrl	Peter$_i$ verspricht nun Lisa$_j$ PRO$_i$ **das kleine Lamm** zu streicheln und zu kraulen.
(n=10)	Peter$_i$ promises now Lisa$_j$ PRO$_i$ to pet and to ruffle the little lamb.
o-ctrl	Peter$_i$ erlaubt nun Lisa$_j$ PRO$_j$ **das kleine Lamm** zu streicheln und zu kraulen.
(n=10)	Peter$_i$ allows now Lisa$_j$ PRO$_j$ to pet and to ruffle the little lamb.
match	Peter$_i$ verspricht nun Thomas$_{MASC}$, **dass er**$_i$ das kl. Lamm streichelt u. krault.
(n=10)	Peter$_i$ promises now Thomas$_{MASC}$ that he$_i$ will pet and ruffle the little lamb.
mismatch	Peter$_i$ verspricht nun Lisa$_{FEM}$, **dass er**$_i$ das kleine Lamm streichelt und krault.
(n=10)	Peter$_i$ promises now Lisa$_{FEM}$ that he$_i$ will pet and ruffle the little lamb.

Note. s-ctrl/o-ctrl = subject/object control, match/mismatch = gender match or mismatch of the main clause nouns. Critical region in bold.

vary the distance between the controlling argument and PRO. Based on earlier findings, subject control structures were regarded as complex because the distance between the controlling argument an PRO is longer than in object control structures (e.g., Caplan & Hildebrandt, 1988; Kwon & Sturt, 2016). In the sentences with a pronoun, only subject control verbs were used. The main clause nouns were common two-syllable German unambiguously male or female first names. In the sentences with PRO, nouns were always of different gender. In the sentences with a pronoun, the gender of the second noun of the matrix clause was manipulated such that it either matched or mismatched in gender with the first noun. This was done to manipulate the similarity of the nouns. Based on previous findings, sentences with gender-matching nouns were regarded as complex because the nouns are more similar than in sentences with gender-mismatching nouns (e.g., Schroeder, 2007; Stewart et al., 2000). Fillers included an object control verb and an overt pronoun (e.g. *Peter erlaubt nun Lisa, dass sie das kleine Lamm streichelt und krault.*, 'Peter allows now Lisa that she pets and ruffles the little lamb.').

2.3.2 Declaratives and relative clauses

Examples of the sentences are given in Table 8 (for all items, see appendix). In these sentences, the order of the nominative subject and the accusative object was varied. They were used to study the processing of canonical and non-canonical word order. In German, word order is canonical when the subject precedes the object, and it is non-canonical when the subject follows the object. Participants had to decide which of the arguments was the subject and the object. This decision depended on the case marking of the determiners and relative pronouns that were unambiguously marked for nominative case or accusative case. The critical region of the sentence was the phrase where the order of the arguments was disambiguated (highlighted in bold in Table 7). In the declaratives, this region consisted of the first noun phrase. In relative clauses, this region

consisted of the relative pronoun.

A set of 80 sentences was used: 20 declaratives and 60 relative clauses.[5] 10 declaratives had a canonical word order, i.e., the subject preceded the object (SO). The other 10 declaratives had a non-canonical word order, i.e., the subject followed the object (OS). Based on previous studies for German, SO declaratives will be regarded as simple and OS declaratives as complex (e.g., Hanne et al., 2011; Vogelzang et al., 2019). Relative clauses consisted of 30 subject and 30 object relative clauses. They were further divided into subject and object modifying relative clauses, and relative clauses with a plural noun (10 items respectively). In the present study, only the 20 subject modifying subject and object relative clauses with singular nouns were analyzed. The analysis was restricted to these sentences because the study investigated the predictions of the RRH with respect to the processing of simple and complex sentences. The other conditions were included to test predictions with respect to changes of number and case between main clause and subclause which were not the focus of this paper.[6] Based on previous findings for German, subject relative clauses will be regarded as simple and object relative clauses as complex (e.g., Adelt et al., 2017; Bader & Meng, 1999). Sentences were pseudo-randomized with a maximum of three consecutive repetitions of the same sentence type.

Table 8: Example of the declaratives and relative clauses used in the experiment.

Sentence type	Condition	Sentence
Declaratives	SO	Hier tröstet **der**$_{NOM}$ **Tiger** gerade den$_{ACC}$ Esel
	(n=10)	Here the$_{NOM}$ tiger just comforts the$_{ACC}$ donkey
	OS	Hier tröstet **den**$_{ACC}$ **Tiger** gerade der$_{NOM}$ Esel
	(n=10)	Here the$_{ACC}$ tiger just comforts the$_{NOM}$ donkey
Relative clause	SRC	Hier ist der Tiger **der**$_{NOM}$ den$_{ACC}$ Esel gerade tröstet
	(n=10)	Here is the tiger who$_{NOM}$ comforts the$_{ACC}$ donkey
	ORC	Hier ist der Tiger **den**$_{ACC}$ der$_{NOM}$ Esel gerade tröstet
	(n=10)	Here is the tiger who$_{ACC}$ the$_{NOM}$ donkey comforts

Note. S = subject O = object, SRC/ORC = subject/object relative clause. Critical region in bold.

To construct the sentences, 10 transitive action verbs with two syllables and a mean lemma frequency of 85 ($SD = 211$) per million tokens in dlexDB (Heister et al., 2011) were used. The nouns referred to animals with masculine gender, and had

[5]This part of the experiment did not include filler sentences

[6]Examples of these sentences are: Object modifying subject / object relative clause: *Ich seh den Tiger, der den Esel gerade tröstet / den der Esel gerade tröstet.*, 'I see the tiger who just comforts the donkey / who the donkey just comforts.'. Subject modifying subject / object relative clause with plural noun in the relative clause: *Hier ist der Tiger, der die Esel gerade tröstet / den die Esel gerade trösten.*, 'I see the tiger who just comforts the donkeys / who the donkeys just comfort.'.

a length of two-syllable and a mean lemma frequency of 356 ($SD = 400$) per million tokens in dlexDB (Heister et al., 2011). Twenty-three students rated the plausibility of the animals as agent or patient of the actions to ensure that all sentences were pragmatically reversible.

2.3.3 Auditory stimuli

Sentences were spoken with a neutral prosodic contour at a rate of 4.79 syllables per second in the experiment on declaratives and relative clauses and at a rate of 3.95 syllables per second in the experiment on control structures. These rates fall in the range of 3–6 syllables per second, which is considered a normal speech rate (Levelt, 2001). Sentences were recorded in a sound-proof booth with a trained female native speaker of German. Recordings were post-processed with Praat (Boersma & Weenink, 2018). The same recordings were used for pairs of simple and complex sentences (e.g., subject and object relative clauses) by exchanging the manipulated region (e.g., *der* 'the.NOM' and *den* 'the.ACC') in the sound files.[7]

2.3.4 Pictures

Pictures consisted of pairs of black-and-white drawings. In the part of the experiment on declaratives and relative clauses, the target picture displayed the agent acting on the patient, and the foil picture displayed the referents with reversed thematic roles (e.g., Figure 16, A). In the part of the experiment on control structures, target and foil picture displayed the target or distractor referent respectively interacting with the animal mentioned in the sentence (e.g., Figure 16, B). Referents had the same size and adopted the same postures. Human referents were identifiable by their initials (e.g., *L* for *Lisa*). The action direction (from left to right or reversed) was balanced. Target and foil pictures were presented in the center of the screen adjacent to each other. The order of the pictures was counterbalanced so as to avoid any bias due to presentation order.

2.4 Data analysis

Data were analyzed separately for IWA and control participants, and for the four sentence structures (i.e., declaratives, relative clauses, control structures with an overt pronoun and control structures with PRO). The data of the test and retest were pooled, i.e., the statistical models included 20 observations per sentence type. The data of the two participant groups were not combined into one model to reduce computation time, which was up to a week for the models presented here. Blinks and saccades were excluded from the analyses. The data were analyzed in two different ways: 1) Time bin

[7] It was checked in a pilot with four students and four elderly control participants that the spliced stimuli sounded natural.

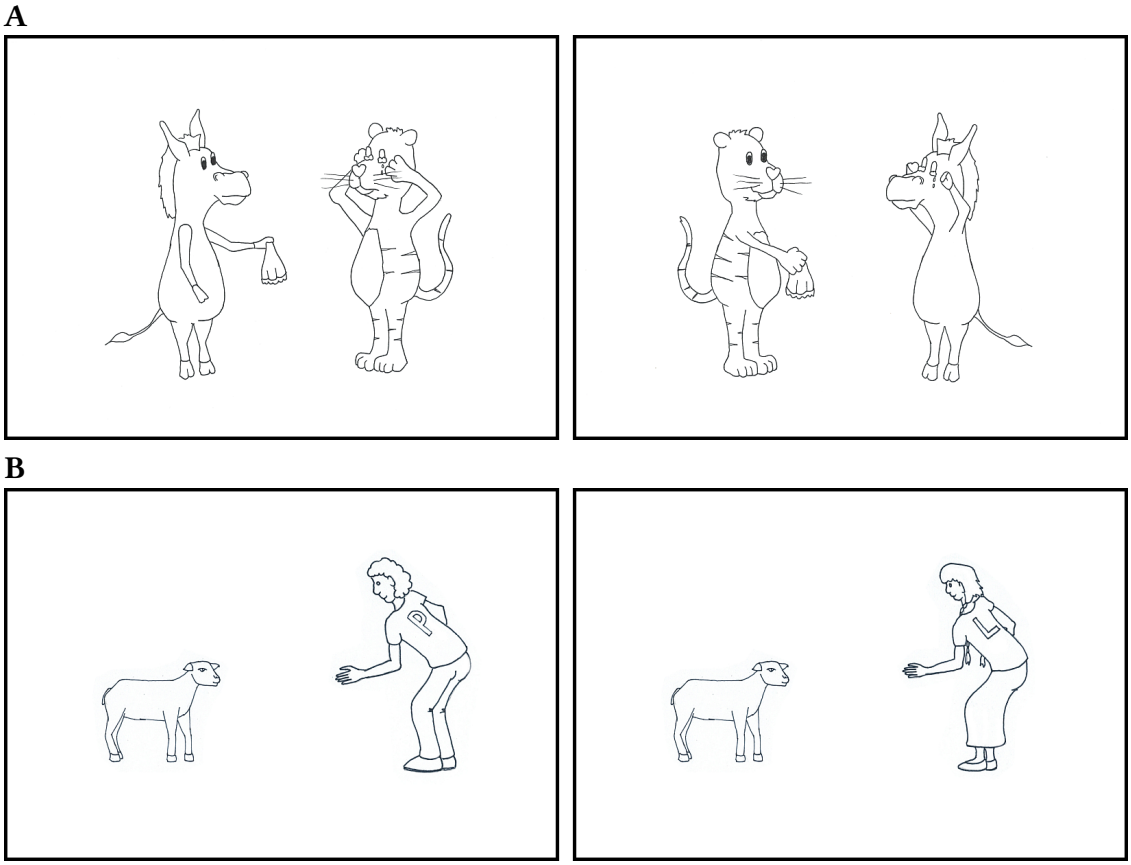

Figure 16: Sample pictures of the part of the experiment with declaratives and relative clauses (A) and the part of the experiment with control structures (B). For the subject relative clause *Here is the tiger that$_{nom}$ comforts the$_{acc}$ donkey.*, the right picture A is the target and the left picture A the foil. For the object relative clause *Here is the tiger that$_{acc}$ the$_{nom}$ donkey comforts.*, the left picture A is the target and the right picture A the foil. For the object control sentence *Peter allows Lisa to pet the lamb.*, the right picture B is the target and the left picture B the foil. For the subject control sentence *Peter promises Lisa to pet the lamb.*, the left picture B is the target and the right picture B the foil.

analysis, in which the data were sliced in 50 ms time bins as done in a growth curve analysis (Mirman, 2014). This fine grained measure was used to determine where in the sentence a change in target fixations occurred at the group level. 2) Time window analysis, in which target fixations were averaged across three broad time windows and the two test phases. This broad measure was chosen to estimate the target fixations for each individual participant as recommended by McMurray (2020). It was not possible to determine the fine grained fixation path of each individual participant from the time bin analysis because the number of observations per participant was too low to get reliable participant-level estimates of the target fixations in each time bin. All data and code are available from https://osf.io/mc2rn/.

2.4.1 Time bin analysis

Analyses included all fixations from the onset of the critical region to a designated cutoff point after the end of the sentence. They were limited to this period to reduce computation time. In declaratives and relative clauses, the critical region was the first determiner or the relative pronoun (see region in bold in Table 8). In control structures, the critical region was the onset of the subclause (see region in bold in Table 7). The designated cut-off point after the end of the sentence was the mean reaction time of the respective participant group for the respective sentence type. Consequently, the analyses of the IWA include longer periods of silence than the analyses of the control participants because IWA had longer mean response times.

Fixations were averaged across 50ms bins. About 99% of the obtained mean fixations were binary (i.e., 1 target fixated, 0 target not fixated), the remaining mean fixations were binarized (cf. Huang & Snedeker, 2020): If the mean proportion of target fixations in a particular bin was smaller or equal to 0.5, a 0 was inserted, otherwise, a 1 was inserted. The mean fixations were analyzed using R (Version 3.6.3; R Core Team, 2020) and the R-package *brms* (Version 2.17.0; Bürkner, 2017; Bürkner, 2018) with Bayesian hierarchical generalized linear mixed models with a logit link and full variance covariance matrices for the random effects of participants and items. Model estimates were back-transformed into proportions for ease of interpretation.

All models included the predictors COMPLEXITY, TEST PHASE and TIME and their interaction. The models of the IWA additionally included the predictor ACCURACY and interactions of all predictors. For COMPLEXITY, sum contrasts were used, where complex sentences were coded as -1 (i.e., OS declaratives, object relative clauses, subject control structures, and control structures with gender matching nouns) and simple sentences as $+1$ (i.e., SO declaratives, subject relative clauses, object control structures, and control structures with gender mismatching nouns). Similarly, a sum coding was used for TEST PHASE (-1 test, $+1$ retest) and ACCURACY ($+1$ correct, -1 incorrect). Following Mirman (2014), higher-order orthogonal polynomials were used for the predictor TIME to account for the fact that the change in proportion of target fixations over time is not linear. In all models, fourth order polynomials were used.

The prior distributions for the parameters in our models were specified as follows: The prior of the intercept was set to $\mathit{Normal}(0, 1.5)$, the priors of the slopes were set to $\mathit{Normal}(0, 1)$, and the prior standard deviations of the random effects to $\mathit{Normal}_{+}(0, 1)$ truncated at zero because standard deviations cannot be negative. The prior of the correlation between the random intercepts and slopes was set to $LKJ = 2$ (Lewandowski et al., 2009) to disfavor extreme correlations. The model output consisted of the posterior distributions of the parameters. The estimated 95% credible interval (CrI) of the posterior was extracted. The CrI is the range of plausible values of the parameters given the data

and model.

The 95% CrIs were used to estimate the point in time of a divergence in proportion of target fixations between two conditions. These divergences were calculated to investigate the predictions of the RRH. More specifically, the following divergences were scrutinized based on the predictions: For prediction 1 (Normal-like processing in correct trials), it was checked whether there was a divergence in target fixations between control participants and IWA, a divergence in target fixations between the correct and incorrect trials of the IWA, and a divergence from 50% target fixations (i.e., the point in time where participants started to fixate the target picture more than the foil picture, cf. Wendt et al., 2014). For prediction 2 (Processing difficulty in complex vs. simple sentences, complexity-capacity interaction), it was checked whether there was a divergence in target fixations between simple and complex sentences, and an interaction between target fixations and response accuracy (the latter analysis is described in detail in the section *Time window analysis*). For prediction 3 (Unsystematic variability in the performance between test and retest), it was checked whether there was a divergence in target fixations between test and retest.

To be counted as a divergence, the 95% CrIs of the respective two conditions were not allowed to overlap for at least 4 consecutive time bins (i.e., 200 ms). To determine a confidence interval (CI) for a divergence point, bootstrapping analyses were carried out (Stone et al., 2020). Different from Stone et al. (2020), we did not fit t-tests for each time bin to determine divergence between two conditions but we used the 95% CrIs of the models previously fit with *brms*. Thus, to determine the CIs for the divergence points we only had to fit one Bayesian model for each sentence type instead of multiple t-tests. The 95% CrIs were resampled for each participant in each time bin, and the divergence between CrIs was calculated for the resampled data. Resampling was done 2000 times to generate a distribution of divergence points.

2.4.2 Time window analysis

This analysis was carried out to test the prediction of the RRH that there is an interaction between resource capacity of an IWA and the complexity of the sentence structure. For this analysis, the data of the test and retest were pooled and trials were divided into three regions of interest: 1) the first half of the target sentence up to and including the critical region, 2) the second half of the target sentence, and 3) the silence region after the sentence until the response key was pressed. For each region, the sum of target fixations and the total number of fixations in each trial was calculated. The sum of target fixations and the total number of fixations were entered as the dependent variables of binomial models with a logit link which were fit in *brms*. Model estimates were back-transformed into proportions for ease of interpretation.

The models included the following predictors: COMPLEXITY, ACCURACY, OVER-

all accuracy and their interactions. The predictors complexity and accuracy were sum coded (+1 simple, −1 complex; +1 correct, −1 incorrect). For the predictor overall accuracy, the overall response accuracy of each IWA for each of the four sentence types were calculated. The response accuracy was then centered per sentence type, i.e., per sentence type, the average response accuracy was subtracted from the response accuracy of each IWA (Schad et al., 2020). In an additional model, overall accuracy was replaced by severity which was the centered severity of each IWA in the Aachen Aphasia Test (see Table 6). The same priors as in the time bin analyses were used.

3 Results Study 2

First, the results of the time bin analyses for the two participant groups will be reported. Afterwards, the results of the time window analyses for each single IWA will be presented. Accuracy and response times of the sentence-picture matching task have been analyzed and reported in Pregla et al. (2021). We will give a summary of the offline results before turning to the target fixations.

3.1 Summary of the offline results

Accuracy and response times are summarized in Table 9. Control participants responded faster and displayed more correct responses than IWA. Both participant groups responded faster and displayed more correct responses in simple versus complex sentences, and in the retest versus the test phase. As visible in Table 9, the response accuracy of the IWA was at or below 50% in OS declaratives, object relative clauses, and in the complex control structures (i.e., match and subject control) in the test phase. This result is addressed in the discussion.

3.2 Results of the Time Bin Analyses

The fixation paths in correct trials of the two participant groups are shown in Figure 17. The fixation paths of correct versus incorrect trials of the IWA are shown in Figure 18. In the following, the results are presented according to the ordering of the three predictions of the RHH as outlined in the theoretical background.

3.2.1 Normal-like processing in correct trials

This prediction of the RRH was tested with the following comparisions: 1) comparisons of the fixation paths of IWA and the control participants for each sentence type and test phase in correct trials, 2) comparison of the fixation paths against a threshold of 50% target fixations (i.e., the threshold above which participants fixated the target picture

Table 9: Mean and standard error of the accuracy (in %) and response times (in ms) in the sentence-picture matching task of the visual world experiment in individuals with aphasia and control participants.

		SO	OS	SRC	ORC	mis-match	match	o-ctrl	s-ctrl
Accuracy									
IWA	test	63.8 (3.3)	42.4 (3.5)	67.1 (3.2)	40.5 (3.4)	60.5 (3.5)	50.5 (3.6)	67.5 (3.3)	50.5 (3.6)
	retest	77.6 (2.8)	43.8 (3.5)	67.1 (3.2)	45.2 (3.5)	76.7 (2.9)	70.5 (3.2)	76.2 (2.9)	62.9 (3.3)
CP	test	99.1 (0.6)	89.1 (2.1)	96.4 (1.3)	92.3 (1.8)	100 (0)	96.8 (1.2)	99.5 (0.5)	96.8 (1.2)
	retest	98.2 (0.9)	93.2 (1.7)	96.8 (1.2)	95.9 (1.3)	100 (0)	98.6 (0.8)	100 (0)	99.1 (0.6)
Response time									
IWA	test	4116 (227)	4975 (297)	3935 (247)	4031 (233)	2553 (236)	3109 (300)	1922 (180)	2511 (223)
	retest	4502 (261)	5299 (294)	4252 (275)	4237 (231)	2061 (203)	2936 (279)	2135 (191)	2344 (233)
CP	test	976 (75)	1875 (130)	1421 (124)	1695 (116)	400 (42)	625 (74)	500 (88)	463 (56)
	retest	885 (60)	1495 (100)	1372 (123)	1445 (110)	398 (36)	499 (43)	371 (34)	439 (39)

Note. IWA = individuals with aphasia, CP = control participants, SO / OS = declarative sentence with canonical / non-canonical word order, SRC / ORC = subject / object relative clause, match / mismatch = gender of the main clause nouns is the same / different, s-ctrl / o-ctrl = subject / object control.

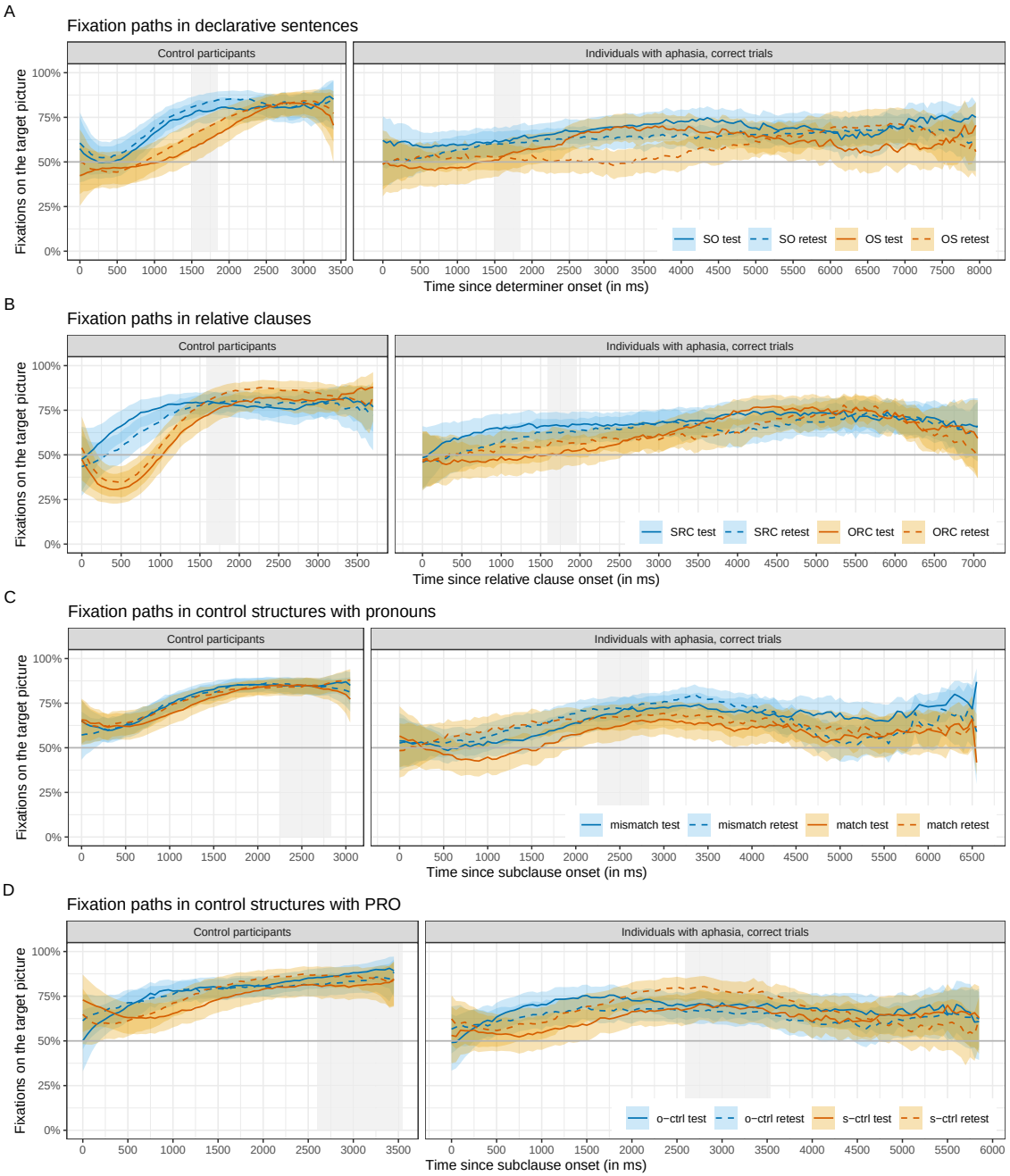

Figure 17: Estimated fixation curves of the correct trials of the control participants and the individuals with aphasia within the time frame from the onset of the critical region until the response key was pressed. A: canonical (SO) and non-canonical (OS) declaratives; B: subject (SRC) and object (ORC) relative clauses; C: control structures with a pronoun with gender matching (match) and mismatching (mismatch) nouns; D: subject (s-ctrl) and object (o-ctrl) control structures with PRO. Solid and dashed lines represent the mean fixations in simple and complex sentences respectively and shaded areas represent the 95% credible intervals around the mean. Vertical bands shaded in grey mark the sentence end.

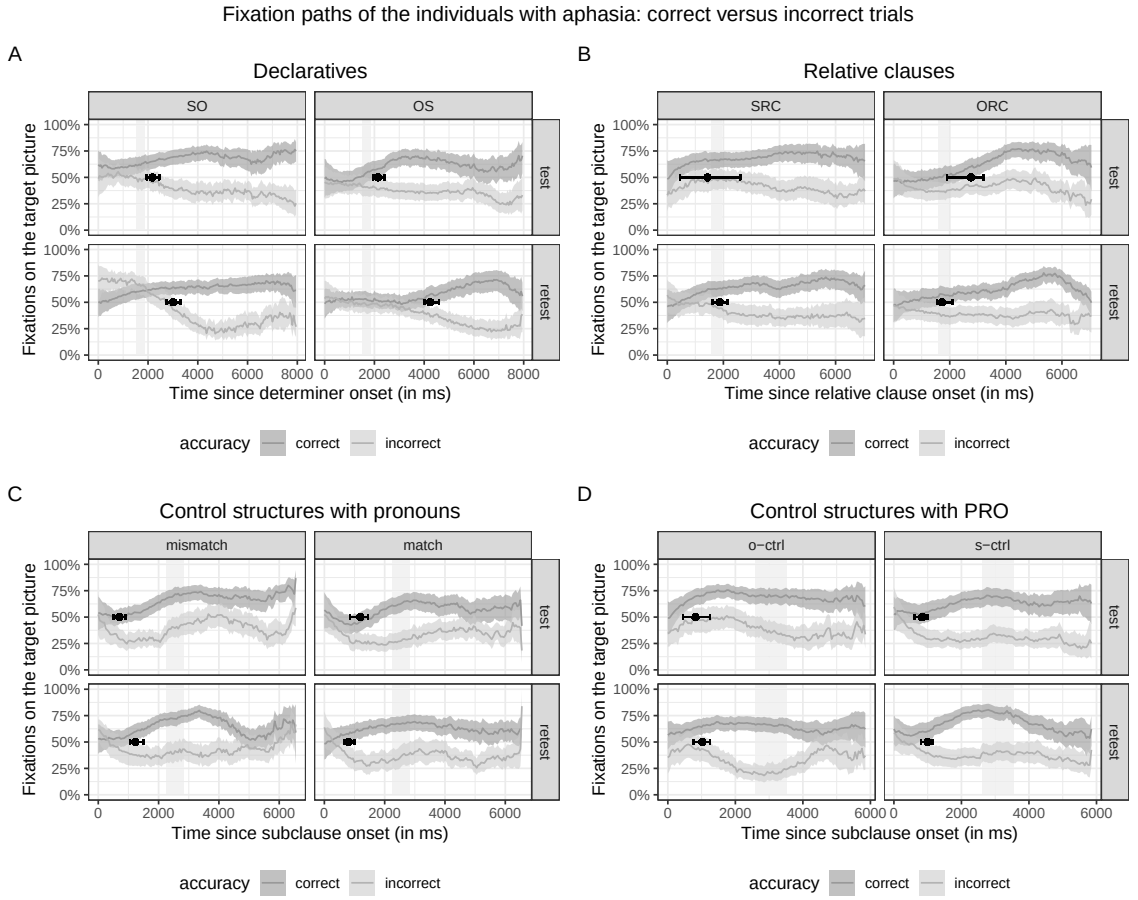

Figure 18: Estimated fixation curves of the individuals with aphasia for the time frame from the onset of the critical region until the response key was pressed. A: canonical (SO) and non-canonical (OS) declaratives; B: subject (SRC) and object (ORC) relative clauses; C: control structures with a pronoun with gender matching (match) and mismatching (mismatch) nouns; D: subject (s-ctrl) and object (o-ctrl) control structures with PRO. Solid dark grey and light grey lines represent the mean fixations in correct and incorrect trials and shaded areas represent the 95% credible intervals around the mean. Dots represent the divergence onsets between correct and incorrect trials. Error bars represent bootstrapped confidence intervals. Vertical bands shaded in grey mark the sentence end. The width of these bands varies because audio files were not of equal length, and therefore, the sentence end can lie somewhere in between these bands. The minimum and maximum audio file length varies per sentence type, i.e., the minimum and maximum audio file length is different in declaratives, relative clauses, control structures with a pronoun and control structures with PRO. As such, the width of the band is different for each sentence type.

more than the foil picture) for each sentence type, test phase and participant group in correct trials, and 3) comparisons of the fixation paths in correct and incorrect trials for each sentence type and test phase in the IWA.

1) Divergence between the participant groups: The increases in target fixations in correct trials were greater in control participants than in IWA. Control participants' target fixations exceeded the IWA's target fixations in all sentence structures except subject control structures, and subject relative clauses in the test phase. The divergence between the groups started less than two seconds after the critical region, which was before or at the sentence end (estimates of the divergence onsets see Table 12 in the appendix).

2) Divergence from 50% target fixations: In both participant groups, the fixation curves of the correct trials exceeded the 50% threshold in all sentence structures (estimates of the divergence onsets see Figure 19 and Table 12 in the appendix). With the exception of SO declaratives in the test phase, subject relative clauses in the test phase, and the subject control structures, the fixation paths of the control participants exceeded the 50% threshold earlier than the fixation paths of the IWA. In both participant groups, target fixations diverged from 50% earlier in the simple sentences than in the complex sentences. This was the case in the declaratives and relative clauses in the control participants and in all sentence types except for control structures with a pronoun in the IWA.

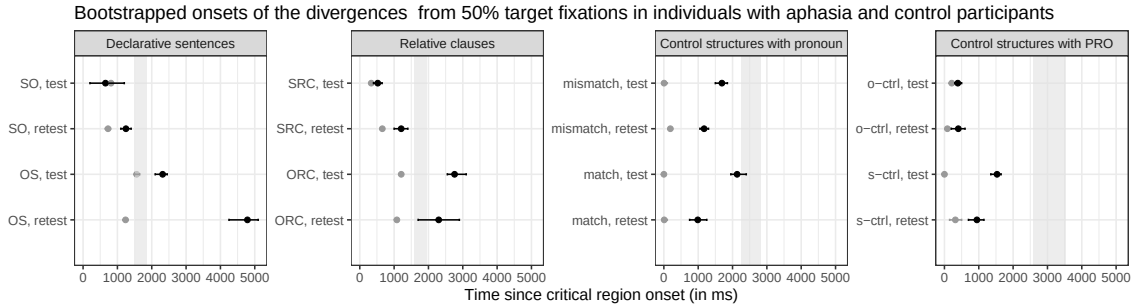

Figure 19: Bootstrapped means (dots) and 95% confidence intervals (error bars) of the divergence onsets from 50% target fixations. Divergence onsets are shown separately for each test phase (test, retest) for canonical and non-canonical declaratives (SO and OS), subject and object relative clauses (SRC and ORC), control structures with gender matching and mismatching nouns with a pronoun (match and mismatch), and subject and object control structures with PRO (s-ctrl and o-ctrl). Vertical bands shaded in grey mark the sentence end.

3) Divergence between correct and incorrect trials: In all sentence types and both test phases, IWA showed more target fixations in correct versus incorrect trials. Divergences occurred earlier in control structures with a pronoun or PRO than in declaratives and relative clauses (see Figure 18, estimates of the divergence onsets see Table 12 in the appendix). In all sentence types, the differences in target fixations between correct and incorrect trials were long lasting, extending over a period of at least two seconds.

3.2.2 Processing difficulty in complex vs. simple sentences, complexity-capacity interaction

This prediction of the RRH was tested by the juxtaposition of fixation paths in simple as opposed to complex sentences for each sentence type, test phase and participant group in correct trials. Furthermore, the RRH predicts an interaction of sentence complexity and resource capacity of the IWA, which will be investigated in the section *Results of the Time Window Analyses*.

In the control participants, the fixation paths of the simple sentences exceeded the fixation paths of the complex sentences in declaratives and relative clauses in both test phases (divergence onsets: declaratives, test: 1008 ms CI: [950, 1100], retest: 1053 ms CI: [950, 1250], relative clauses, test: 223 ms CI: [200, 250], retest: 477 ms CI: [400, 600]). In the correct trials of the IWA, a divergence between simple and complex sentences occurred only in declaratives in the retest (divergence onset: 2847 ms CI: [2200, 3250]).

3.2.3 Unsystematic variability in the performance between test and retest

This prediction of the RRH was tested by comparing the fixation paths in the two test phases for each sentence type and participant group in correct trials.

In both participant groups, the fixation paths of the correct trials overlapped in test and retest. There was one exception: IWA showed earlier increases in target fixations in the test phase compared to the retest phase in OS declaratives (divergence onset: 2860 ms CI: [2550, 3150]).

3.3 Results of the Time Window Analyses

The time window analyses were carried out to investigate the relationship between overall comprehension accuracy of each IWA in a sentence structure and their target fixations. The analysis was based on the prediction of the RRH that there is a complexity-capacity interaction. The results are visualized in Figure 20.

Figure 20 A illustrates the relationship between overall response accuracy of each IWA and their differences in target fixations between simple and complex sentences in the second half of the sentence after the critical word and in the silence region. The interactions between overall response accuracy and sentence complexity were uninformative in all sentence types (for the estimates see Table 10). Figure 20 B shows the relationship between overall response accuracy of each IWA and their differences in target fixations between correctly and incorrectly answered trials in the second half of the sentence after the critical word and in the silence region. As it can be seen, there was no indication that overall response accuracy systematically influenced the differences in target fixations in correct versus incorrect trials (for the estimates see Table 10). Rather, in the silence region, distributions were shifted to the right in all IWA and sentence types

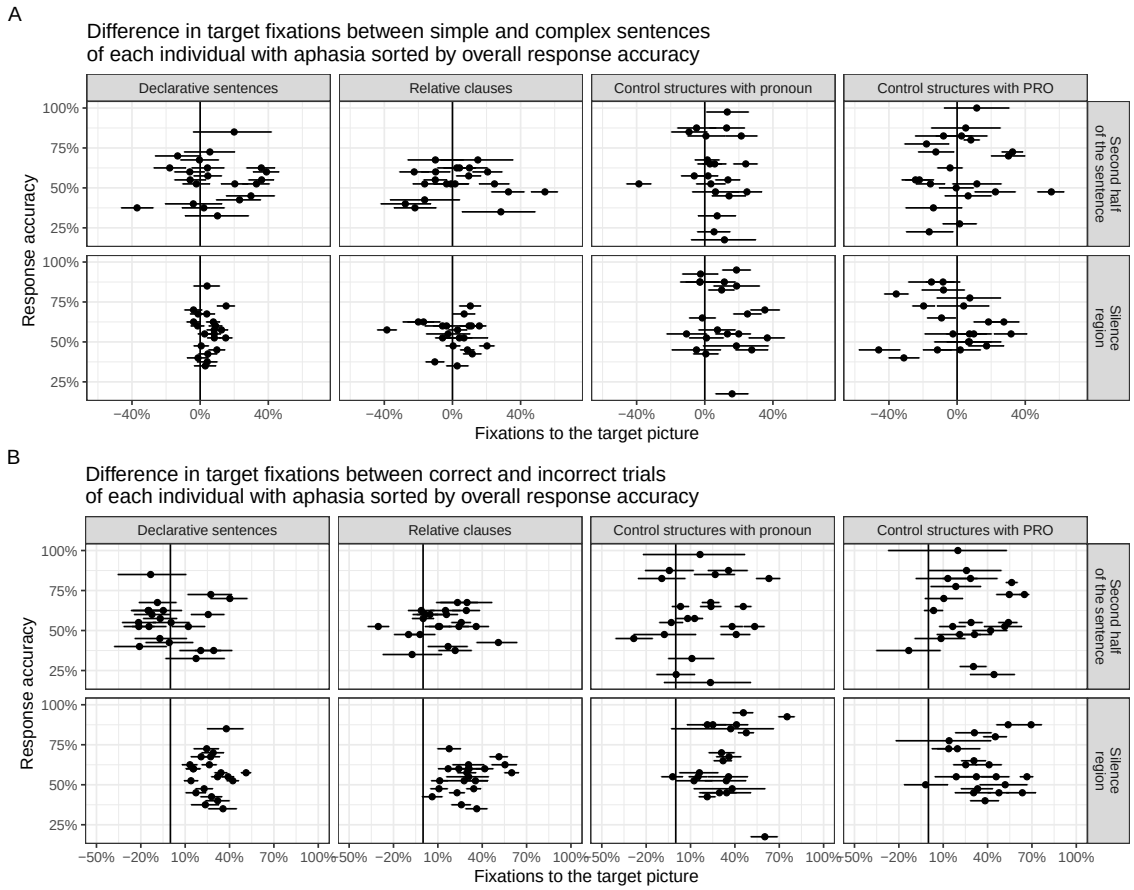

Figure 20: Mean estimates (dots) and 95% credible intervals (horizontal lines) of (A) the difference in target fixations between simple and complex sentences and (B) the difference in target fixations between correct and incorrect trial in each individual with aphasia in the four investigated sentence types in the second half of the sentence after the critical word and in the silence region. Participants are displayed in descending order by their overall response accuracy in the respective sentence type. Distributions that are right-shifted denote higher proportions of target fixations in simpler sentences (A) or correct trials (B).

as visible in the lower part of Figure 20 B. This means that all IWA fixated the target picture more in trials in which they answered correctly across sentence types.

3.3.1 Additional Time Window Analysis

In addition to the analyses above, an analysis that was not based on the predictions of the RRH was carried out in order to test whether there was a relationship between the severity grade measured with the Aachen Aphasia Test (see Table 6) and the individual target fixations in the second half of the sentence after the critical word or in the silence region. In our group of IWA with an aphasia severity grade ranging from mild to moderate there was no indication that the severity grade influenced the overall amount of target fixations or the differences in target fixations between simple and complex as well as between correct and incorrect trials (for the estimates see Table 13 in the appendix). However, it cannot be ruled out that there is an influence of severity on the

Table 10: Means and 95% Credible Intervals (CrI) for the influence of the overall response accuracy in the sentence type on the target fixations in the second half of the sentence and in the silence region.

	Overall Accuracy	Overall Acc × Complexity	Overall Acc × Trial Acc
Region 2			
declaratives	0.3% CrI: [-1.5, 2.1]	0.1% CrI: [-1.9, 2.1]	-0.1% CrI: [-2, 1.9]
RC	-0.5% CrI: [-2.4, 1.3]	1.4% CrI: [-1.6, 4.3]	-1.8% CrI: [-4.4, 0.7]
pronoun	0.5% CrI: [-0.8, 1.8]	0.5% CrI: [-0.5, 1.5]	0.3% CrI: [-1.3, 2]
PRO	0.2% CrI: [-2.3, 2.6]	-0.3% CrI: [-2.2, 1.6]	1.3% CrI: [-1.4, 3.8]
Region 3			
declaratives	0.9% CrI: [-0.1, 1.8]	-0.4% CrI: [-0.9, 0.3]	0.5% CrI: [-0.7, 1.7]
RC	0.9% CrI: [-0.4, 2.3]	-0.6% CrI: [-2.4, 1.2]	0.1% CrI: [-2, 2.3]
pronoun	0.6% CrI: [-1.3, 2.4]	0.1% CrI: [-0.9, 1.1]	0.7% CrI: [-0.8, 2.2]
PRO	0.3% CrI: [-1.4, 1.8]	0.1% CrI: [-1.7, 1.9]	1.5% CrI: [-0.6, 3.5]

Note. Acc = Accuracy, Region 2 = second half of the sentence after the critical word, Region 3 = silence region, RC = relative clauses, pronoun = control structures with a pronoun, PRO = control structures with PRO.

target fixations for a group of IWA with a wider range of severity levels.

4 Discussion Study 2

This study investigated predictions of the RRH (Caplan, 2012) regarding sentence processing in IWA. Sentence processing abilities were assessed with an auditory sentence-picture matching task by measuring the proportion of target fixations in the visual world paradigm. Fixation patterns were investigated across two test phases and four sentence types.

Before we discuss the fixation patterns with respect to the predictions of the RRH, it is important to check whether the response accuracies of the IWA in our study are representative for IWA. This validation check can be carried out by comparing our response data to previous visual world studies. We show below that our accuracy data exhibit very similar patterns to the patterns observed in 13 previously published visual world studies.

After presenting the validation check, we will discuss the fixation patterns with respect to the three investigated predictions of the RRH, namely: 1) Normal-like processing in correct trials, 2) Processing difficulty in complex vs. simple sentences, and complexity-capacity interaction, and 3) Variability in the performance between test and retest due to random noise.

4.1 Validation check: Comparison of the accuracy in this study to that of previous studies

As shown in Table 9, the accuracy of the IWA in the current study was at chance in the comprehension of complex sentences. To exclude the conclusion that this performance of the IWA was exceptionally low, we did a comparison with studies using similar tasks in the visual world paradigm with similar participants. The comparison included accuracy data in the comprehension of several sentences types from the following visual world studies: Adelt et al. (2017), Bos et al. (2014), Choy and Thompson (2005, 2010), Dickey et al. (2007), Dickey and Thompson (2009), Engel et al. (2018), Hanne et al. (2015), Hanne et al. (2016), Hanne et al. (2011), Mack and Thompson (2017), Mack et al. (2016), Meyer et al. (2012), and Schumacher et al. (2015), Sheppard et al. (2015), Thompson et al. (2004). The extracted accuracies are provided in Table 14 in the appendix. As shown in Figure 21, the accuracies of the IWA in this study are within the range of accuracies of the IWA in previous studies. A linear model was fit with *lme4* (Bates et al., 2015) to the arcsine-transformed mean accuracy with study as random effect. According to the model, the mean accuracy of our study (65%, coded as 1) was not significantly different from the mean accuracy of earlier studies (60%, coded as -1, $\hat{\beta}$ = -0.04%, SE = 0.05, t = -0.67). This shows that there is no evidence that the accuracies in our study are atypical in any respect.

4.2 Processing in correct trials

According to the RRH, processing in correct trials should be normal-like. Although the accuracy of the IWA lies within chance range, the observation of an increase in target fixations above 50% speaks against guessing and in favor of normal-like processing. This assumption is further supported by the fact that the increase occured early (on average 2 seconds after onset of the critical region, estimates see Table 12) during the trial and not shortly before response selection in all sentence types of both test phases (Burchert et al., 2013; Hanne et al., 2011). Furthermore the early and stable difference in target fixations between correct and incorrect trials corroborates the notion of normal-like target decision (Burchert et al., 2013; Hanne et al., 2011). As in Hanne et al. (2012), the oberservation that each individual IWA displayed these differences in correct and incorrect trials (irrespective of the overall response accuracy) further advocates the assumption of normal-like processing (see Figure 20 B).

In addition to normal-like processing, it was predicted that processing speed is slowed down in IWA. To evaluate this prediction, the fixation paths of the IWA were compared to those of the control participants. Similar to previous studies (e.g., Mack et al., 2016; Meyer et al., 2012), IWA showed later increases in target fixations than control participants (i.e., the lower bound of the 95% CrI estimated for the fixation paths of the

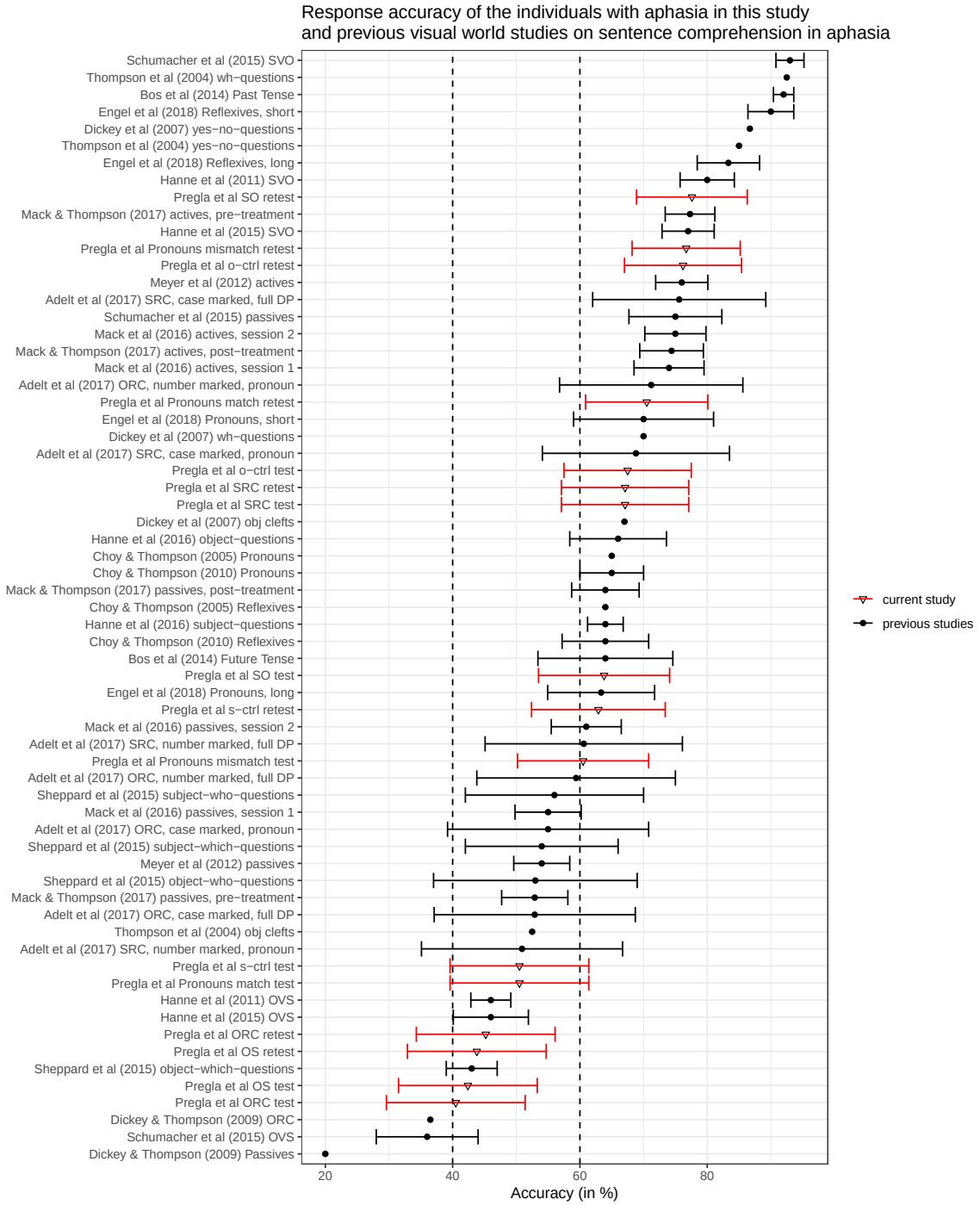

Figure 21: Response accuracy of the individuals with aphasia in the current and previous visual world studies on sentence comprehension in aphasia sorted increasingly by mean accuracy. Dots and triangles represent the mean accuracy and error bars represent standard errors (if error bars are missing, standard errors could not be derived from the information provided in the study). Dashed lines mark an area of 40–60% accuracy which would be the chance area for a task in which the probability of getting a correct response by guessing is 50% and the number of items is 100.

control participants exceeded the upper bound of the 95% CrI estimated for the fixation paths of the IWA). This suggests that IWA do not process morpho-syntactic information as a cue to sentence processing as rapidly as control participants. Furthermore, the delayed increase in target fixations was visible across sentence structures with different types of morpho-syntactic information, namely case information (declaratives, relative clauses), gender information (control structures with a pronoun), and information about the verb's control type (control structures with a pronoun or PRO). Thus, it does not seem to be one specific type of morpho-syntactic information that leads to sentence processing difficulty in IWA. Rather, morpho-syntactic processing in IWA seems to be slowed down in general. This finding is in line with the RRH under the assumption that reduced resources in IWA are reflected by a reduction of processing speed, an assumption that was also put forward by Caplan et al. (2015).

Finally, the increase of target fixations in IWA was not as pronounced as in language unimpaired control participants. This is not in line with Nozari et al. (2016) who found similar increases in both in language impaired and unimpaired participants, however the increase was delayed in IWA. Following the reasoning of Nozari et al. (2016), if IWA would trade speed for accuracy target fixations should increase more slowly but to the same maximum as in control participants. Our finding of a less noticable increase in target fixations in addition to a delay might suggest that in IWA the decision for the target picture was taken with less certainty. We will elaborate on what might lead to the reduced certainty in picture selection in the summary section below.

Overall, the data are consistent with the general conclusion of visual world studies in aphasia, namely that IWA do not guess and deliberately decide on the target picture in correct trials. This conclusion was confirmed at the group level and the individual participant level. This result is in line with the prediction of the RRH that processing in correct trials is normal-like.

4.3 Processing of complex sentences

According to the RRH, participants should have more processing difficulty in complex vs. simple sentences, and there should be an interaction between sentence complexity and resource capacity. To test this prediction, sentence complexity (i.e., canonicity, similarity of noun phrases, dependency length) in different sentence types (i.e., declaratives, relative clauses, control structures with pronoun or PRO) was varied, and the fixation paths of the simple and complex sentences in each sentence structure were compared.

In the control structures, neither the control participants nor the IWA showed differences in target fixations between the simple and complex sentences. Control structures with pronouns were regarded as simple when the gender of the pronoun's antecedent and a distractor noun mismatched and as complex when the gender of the

two nouns matched. Control structures with PRO were regarded as simple when the antecedent directly preceded PRO (object control) and as complex when a noun intervened between the antecedent and PRO (subject control). Irrespective of the pronoun type, target fixations overlapped between the simple and complex sentences. Similar results have been obtained for language-unimpaired participants for reflexive pronouns, in which the distractor noun also did not influence pronoun resolution (Dillon et al., 2013; Schroeder, 2007; Sturt, 2003). A possible explanation for the lack of influence of the distractor might be that only antecedents accessible for binding are considered during pronoun resolution (Sturt, 2003).

In the declaratives and relative clauses, control participants showed differences between sentences with a canonical and non-canonical word order in both test phases. That is, irrespective of sentence type, there were more target fixations in canoncial sentences than in non-canonical sentences. As in previous studies (e.g., Hanne et al., 2015; Mack et al., 2016; Meyer et al., 2012), these differences in fixations between sentences with a canonical and non-canonical word order can be regarded as agent-first processing pattern. That is, control participants expected the canonical word order, which is more frequent in German than the non-canonical order in non-experimental settings (Bader & Häussler, 2010). Control participants rapidly revised this expectation in non-canonical sentences after they encountered the disambiguating information in the input. These results are consistent with the established findings regarding processing in language-unimpaired control participants (e.g., Hanne et al., 2015; Mack et al., 2016; Meyer et al., 2012).

In contrast to the control parcitipants, IWA displayed no differences in target fixations between sentences with a canonical and non-canonical word order, with the exception of declaratives in the retest. At first glance, the absence of differences in online processing (i.e., the overlap in target fixations between canonical and non-canonical sentences) is surprising given the fact that we observed differences in offline processing (i.e., lower response accuracy in non-canonical versus canonical sentences, Pregla et al., 2021). This contradiction between the offline and online data can be explained by the fact that only the fixations of the correct trials entered the analyses. The result can therefore be interpreted as follows: Non-canonicl sentences induced a higher number of incorrect responses as compared to canonical sentences. However, if a correct response was given, processing (as indicated by fixation patterns) was similar for both non-canonical and canonical sentences. Thus, the overlapping fixation paths in correct trials suggest that IWA were able to process the sentences correctly regardless of complexity. In principle, this conclusion is consistent with the RRH, according to which both canonical and non-canonical sentences are processed normal-like provided the randomly fluctuating resources of the IWA are high enough. However, the conclusion that processing in IWA is normal-like may be premature as the comparison of the participant groups in the next

section shows.

The results of the current study and previous studies suggest an agent-first fixation pattern for control participants (Hanne et al., 2015; Mack & Thompson, 2017). This pattern consists of increasing fixations to the distractor picture in non-canonical trials followed by increasing fixations to the target picture reflecting a revision of the prediction. In contrast, the results of the current study and previous studies suggest no agent-first fixation pattern for IWA (Hanne et al., 2015; Mack et al., 2016; Meyer et al., 2012). How can the absence of the agent-first fixation pattern in IWA be explained? The RRH assumes that IWA have a reduced and fluctuating resource, and this reduced resource likely manifests itself in a reduction of processing speed. Importantly, a slowdown in processing speed should entail a slow emergence of agent-first predictions. Moreover, resource fluctuation should lead to variation with respect to the emergence of agent first predictions during sentence processing. Due to this fluctuation, we assume that IWA may or may not create an agent-first prediction before the unambiguous case cue occurs in the input. If they do not create an agent-first prediction before the unambiguaous case cue occurs, there is no mismatch between the agent-first prediction and the information that is provided by the cue. As a result a correct response is given. This processing pattern matches with the fixation paths in correct trials: Due to the absence of an agent-first prediction, no revision of the prediction is needed in non-canonical sentences. Therefore, fixation paths overlap in canonical and non-canonical sentences. In contrast, if IWA do engage in an agent-first interpretation befor the cue information is given, a mismatch arises between this prediction and the cue. We assume that this conflict cannot be solved because IWA are unable to revise a previously made prediction, thus resulting in an incorrect response. This could explain IWA's high number of incorrect responses in non-canonical trials. This interpretation is supported by the fixation patterns in incorrect non-canonical trials: Due to the agent-first prediction, IWA show increasing distractor fixations, and as they are not able to revise their prediction, these fixation patterns do not change, i.e. IWA continue to fixate the distrator picture. The conclusion that IWA might be impaired in revising initial sentence interpretations is consistent with the results of Lissón et al. (2021). Using computational modeling, these authors found that IWA have a much lower probability of *backtracking* (i.e., revision of an incorrect sentence interpretation to the correct one) than control participants. That is, incorrect initial sentence interpretations, e.g., agent-first predictions in non-canonical sentences, might result in incorrect responses, as incorrect interpretations cannot be revised. Put differently, the results do not suggest that IWA are in general unable to make agent-first predictions, but that IWA have difficulties *revising* their agent-first predictions based on the morpho-syntactic information of the input. Overall, the fixation patterns of IWA for non-canonical sentences hint at a processing pattern that is not only slower but also different from normal-like processing in that the revision of agent-first predictions is

impaired.

With respect to individual participants, the RRH predicted an interaction between sentence complexity and resource capacity. Applied to fixation data, it was assumed that the differences in target fixations between simple and complex sentences should be larger in IWA with lower resource capacity. Resource capacity was operationalized as the overall response accuracy (i.e., low accuracy = low capacity and high accuracy = high capacity). As shown in Figure 20, the patterns were not consistent with the predicted interaction. Possibly, the interaction could not be detected since the IWA as a group also did not show clear differences between complex and simple sentences, as discussed in the previous paragraphs. Furthermore, the number of 20 observations per sentence type might have been too low to find differences between individuals.

In sum, the data discussed in this section are consistent with the RRH's prediction that processing difficulty is higher in complex versus simple sentences. More specifically, although the number of target fixations in IWA was not clearly different in simple and complex sentences, the number of incorrect trials was higher in complex versus simple sentences. One unexpected finding is that IWA did not show an agent-first fixation pattern. The absence of this pattern might indicate that agent-first predictions emerge slow in IWA and that IWA have difficulties successfully revising agent-first predictions once they emerged.

4.4 Processing variability between test phases

The RRH predicts variability in the performance of IWA caused by random fluctuations in resources. To test this prediction, the target fixations of the test phase and the retest phase were compared.

The control participants did not exhibit notable changes (i.e., increases or decreases) in target fixations in the retest phase. This result is inconsistent with Mack et al. (2016), where control participants showed systematic increases in target fixations in the retest. A reason for the diverging results could be that the interval between the test phase and the retest phase was different between this study and Mack et al. (2016). In this study, the gap between test phases was two months, whereas, in Mack et al. (2016), it was only one week. The short gap in Mack et al. (2016) could have enabled participants to remember the task better than in this study, which could explain the differences in practice effects between the two studies.

With respect to IWA, Mack et al. (2016) observed a systematic increase in target fixations between test and retest which they interpreted as a practice effect. In our study, we did not observe such a systematic increase in target fixations in the retest. Furthermore, clear changes in fixation paths between the test phases occurred only in one sentence structure, namely the OS declaratives. In OS declaratives, target fixations in-

creased more slowly in the retest than in the test phase. This result is unexpected when assuming a practice effect, because a practice effect should have led to a faster increase in target fixations. A similar slowdown in sentence processing has been observed by Warren et al. (2016). In their study, the IWA became slower in reading low predictable sentences over the course of the experiment, while the control group became faster. The authors speculated that IWA do not adjust to experimental sentences in the same way as control participants (Warren et al., 2016). The slower increase in the retest in our study might therefore suggest that IWA had difficulties adapting to sentences with a non-canonical word order. However, the results of both studies are not fully comparable with each other since Warren et al. (2016) studied changes in the behaviour within a single session and not between different sessions. Furthermore, the slowdown in target fixations is only present in the OS declaratives and not in the other sentence structures. Therefore, the difference in target fixations in OS declaratives could be an accidental finding.

In sum, the RRH predicted variability in the performance because of random fluctuations in processing resources and the pattern of target fixations in test and retest is overall in line with this prediction.

5 Summary and Conclusion Study 2

Table 11 provides an overview of the predictions of the RRH, the expected fixation patterns, and our results.

Four findings were consistent with the RRH. First, there were stable increases over 50% target fixations in correct trials, and early and stable divergences between correct and incorrect trials. These fixation patterns occurred in simple and complex sentences, across all sentence structures and test phases, both at the group level and at the individual participant level. The latter results indicate that IWA do not choose a picture at random but settle on a picture in correct trials in the sentence-picture matching task. This finding is consistent with the prediction of the RRH that the processing of IWA in correct trials is normal-like. Second, IWA showed a slower-than-normal increase in target fixations. This slowdown is compatible with the RRH because processing speed might be the resource that is reduced in IWA. Third, while the expected divergence in target fixation between simple and complex sentences was not confirmed, the number of incorrect trials was higher for complex sentences. Taking response accuracy into account, this finding is in line with the RRH according to which processing should be successful irrespective of sentence complexity once the resource demands are met. Fourth, IWA did not show systematic increases in target fixations in the retest. This finding is consistent with the prediction of the RRH that sentence processing should be variable in IWA.

Table 11: Predictions of the resource reduction hypothesis, their expected expression in the visual world experiment and actual findings.

Predictions of the resource reduction hypothesis, expected fixation pattern for individuals with aphasia	Findings consistent with predictions?
1) Normal-like processing in correct trials	
– increases over 50% in TF in correct trials	yes, but reduced magnitude in TF
– higher increases in TF in correct vs. incorrect trials	yes
– slower increase in TF compared to control participants	yes
2) Processing difficulty in complex vs. simple sentences, complexity-capacity interaction	
– higher increases in TF in simple vs. complex sentences	no*, similar TF in correct trials
– interaction complexity effect and overall response accuracy	inconclusive
3) Unsystematic variability in the performance between test and retest	
– unsystematic changes in fixation paths between test and retest	yes

Note. *The predicted pattern was only observed in declaratives in the retest phase. TF = target fixations.

Three findings diverged from the predictions of the RRH. First, the magnitude of target fixations was lower in IWA than in the control participants in correct trials, which could reflect a reduced certainty in picture selection. Second, IWA did not exhibit an agent-first fixation pattern, which points towards an impairment in the revision of structural predictions. Third, IWA showed increased canonicity effects in the retest phase, which might indicate that IWA have difficulties adjusting to the input. Under the RRH, these findings for correct trials are unexpected given that processing in correct trials should be normal-like. Caplan et al. (2015) also found differences in processing between IWA and control participants for correct trials. They concluded that the impairment has graded effects, "… at times slowing incremental processing without leading to an error" (Caplan et al., 2015, p. 305). While our results also indicated a slowdown, it is questionable whether slowed processing alone can explain the difficulties in correct trials. Possibly, an additional source might cause these difficulties. For example, IWA might struggle matching their expectations about the sentence structure with the actual linguistic input, which requires correct perception of the input, detection of the mismatch between input and expectation, and updating the expectations (Cope et al., 2017). Difficulty in matching expectation and input might cause IWA to be less certain in picture selection than control participants and impaired in revising incorrect expectations. The impairment in revising expectations might eventually lead to difficulties adjusting to complex non-canonical sentences. Overall, it seems that processing difficulties may not only underlie incorrect trials but also correct trials. Thus, our results confirm the

RRH in part, but not completely, since in some respects processing of the IWA in correct trials is not normal-like.

To conclude, our findings were consistent with the predictions of the RRH that processing difficulty is more frequent in complex versus simple trials, and that processing varies unsystematically between test phases. Also the observed processing slowdown in IWA is compatible with the RRH. However, our results were mixed with respect to the prediction that processing in correct trials is normal-like. On the one hand, IWA showed a deliberate decision for the target picture in correct trials in all sentence structures and both test phases, which speaks for normal-like processing. On the other hand, IWA showed a reduced certainty in picture selection, difficulty in revising sentence interpretations, and difficulty in adjusting to complex sentences, which speaks for processing difficulties in correct trials. Further research is needed to investigate whether these performance patterns can be attributed to the effects of slowed processing.

Acknowledgements We would like to thank the individuals with aphasia and the language unimpaired volunteers who participated in this study. We are grateful to Andreas Schmidt for helpful discussions of the work. We would also like to thank Silke Böttger, Sarah Düring and Therese Mayr for assisting with data collection. This research was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317633480 – SFB 1287, project B02 (PIs: Shravan Vasishth, Frank Burchert, and Nicole Stadie).

A2 Appendix Study 2

A2.1 Sentence stimuli

Declarative sentences

The manipulated determiners (*nominative / accusative*) are presented in italics.

1. Hier badet *der / den* Esel gerade *den / der* Tiger.
   Here *the (nom /acc)* donkey just bathes *the (acc /nom)* tiger.

2. Hier zeichnet *der / den* Büffel gerade *den / der* Panther.
   Here *the (nom /acc)* buffalo just draws *the (acc /nom)* panther.

3. Hier kitzelt *der / den* Hamster gerade *den / der* Igel.
   Here *the (nom /acc)* hamster just tickles *the (acc /nom)* hedgehog.

4. Hier rettet *der / den* Pudel gerade *den / der* Kater.
   Here *the (nom /acc)* poodle just rescues *the (acc /nom)* tomcat.

5. Hier bürstet *der / den* Kater gerade *den / der* Pudel.
   Here *the (nom /acc)* tomcat just brushes *the (acc /nom)* poodle.

6. Hier tröstet *der / den* Tiger gerade *den / der* Esel.
   Here *the (nom /acc)* donkey just comforts *the (acc /nom)* tiger.

7. Hier leitet *der / den* Panther gerade *den / der* Büffel.
   Here *the (nom /acc)* panther just guides *the (acc /nom)* buffalo.

8. Hier füttert *der / den* Igel gerade *den / der* Hamster.
   Here *the (nom /acc)* hedgehog just feeds *the (acc /nom)* hamster.

9. Hier findet *der / den* Eber gerade *den / der* Otter.
   Here *the (nom /acc)* boar just finds *the (acc /nom)* otter.

10. Hier streichelt *der / den* Otter gerade *den / der* Eber.
    Here *the (nom /acc)* otter just pets *the (acc /nom)* boar.

Relative clauses

The manipulated determiners to get subject and object relative clauses (*nominative / accusative*) are presented in italics.

1. Hier ist der Esel, *der / den den / der* Tiger gerade badet.
   Here is the donkey *who (nom / acc) the (nom / acc)* tiger just bathes.

2. Hier ist der Büffel, *der / den den / der* Panther gerade zeichnet.
   Here is the buffalo *who (nom / acc) the (nom / acc)* panther just draws.

3. Hier ist der Hamster, *der / den den / der* Igel gerade kitzelt.
   Here is the hamster *who (nom / acc) the (nom / acc)* hedgehog just tickles.

4. Hier ist der Pudel, *der / den den / der* Kater gerade rettet.
   Here is the poodle *who (nom / acc) the (nom / acc)* tomcat just rescues.

5. Hier ist der Kater, *der / den den / der* Pudel gerade bürstet.
   Here is the tomcat *who (nom / acc) the (nom / acc)* poodle just brushes.

6. Hier ist der Tiger, *der / den den / der* Esel gerade tröstet.
   Here is the tiger *who (nom / acc) the (nom / acc)* donkey just comforts.

7. Hier ist der Panther, *der / den den / der* Büffel leitet.
   Here is the panther *who (nom / acc) the (nom / acc)* buffalo just guides.

8. Hier ist der Igel, *der / den den / der* Hamster gerade füttert.
   Here is the hedgehog *who (nom / acc) the (nom / acc)* hamster just feeds.

9. Hier ist der Eber, *der / den den / der* Otter gerade findet.
   Here is the boar *who (nom / acc) the (nom / acc)* otter just finds.

10. Hier ist der Otter, *der / den den / der* Eber gerade streichelt.
    Here is the otter *who (nom / acc) the (nom / acc)* boar just pets.

Sentences with PRO

The manipulated verb (*subject control / object control*) is presented in italics.

1. Peter *verspricht / erlaubt* nun Lisa, das kleine Lamm zu streicheln und zu kraulen.
   Peter now *promises / allows* Lisa to pet and to ruffle the little lamb.

2. Thomas *versichert / gestattet* nun Anna, das dicke Rind zu melken und zu hüten.
   Thomas now *assures / allows* Anna to milk and to tend the thick cattle.

3. Thomas *droht / befielt* nun Lisa, das schnelle Huhn zu jagen und zu fangen.
   Thomas now *threatens / commands* Lisa to chase and to catch the fast chicken.

4. Peter *garantiert / empfiehlt* nun Anna, das stolze Ross zu bürsten und zu striegeln.
   Peter *guarantees / recommends* now Anna to brush and to comb the proud steed.

5. Thomas *schwört / rät* nun Anna, das süße Ferkel zu waschen und zu säubern.
   Thomas now *swears / advises* Anna to wash and to clean the sweet piglet.

6. Lisa *verspricht / erlaubt* nun Peter, das alte Schaf zu impfen und zu pflegen.
   Lisa now *promises / allows* Peter to vaccinate and to nurse the old sheep.

7. Anna *versichert / gestattet* nun Thomas, das junge Kalb zu malen und zu zeichnen.
   Anna now *assures / allows* Thomas to paint and to draw the young calf.

8. Anna *droht / befielt* nun Peter, das kluge Schwein zu füttern und zu mästen.
   Anna now *threatens / commands* Peter to feed and to fatten the clever pig.

9. Lisa *garantiert / empfiehlt* nun Thomas, das scheue Reh zu locken und zu suchen.
   Lisa now *guarantees / recommends* Thomas to lure and to search the shy deer.

10. Lisa *schwört / rät* nun Peter, das schöne Pferd zu satteln und zu zäumen.
    Lisa now *swears / advises* Peter to saddle and to bridle the nice horse.

Sentences with a pronoun

The manipulated noun (*same gender / different gender*) is presented in italics.

1. Peter verspricht nun *Thomas / Lisa*, dass er das kleine Lamm streichelt und krault.
   Peter now promises *Thomas / Lisa* that he will pet and ruffle the little lamb.

2. Thomas versichert nun *Peter / Anna*, dass er das dicke Rind melkt und hütet.
   Thomas now assures *Peter / Anna* that he will milk and tend the thick cattle.

3. Thomas droht nun *Peter / Lisa*, dass er das schnelle Huhn jagt und fängt
   Thomas now threatens *Peter / Lisa* that he will chase and catch the fast chicken.

4. Peter garantiert nun *Thomas / Anna*, dass er das stolze Ross bürstet und striegelt.
   Peter guarantees now *Thomas / Anna* that he will brush and comb the proud steed.

5. Thomas schwört nun *Peter / Anna*, dass er das süße Ferkel wäscht und säubert.
   Thomas now swears *Peter / Anna* that he will wash and clean the sweet piglet.

6. Lisa verspricht nun *Anna / Peter*, dass sie das alte Schaf impft und pflegt.
   Lisa now promises *Anna / Peter* that she will vaccinate and nurse the old sheep.

7. Anna versichert nun *Lisa / Thomas*, dass sie das junge Kalb malt und zeichnet.
   Anna now assures *Lisa / Thomas* that she will paint and draw the young calf.

8. Anna droht nun *Lisa / Peter*, dass sie das kluge Schwein füttert und mästet.
   Anna now threatens *Lisa / Peter* that she will feed and fatten the clever pig.

9. Lisa garantiert nun *Anna / Thomas*, dass sie das scheue Reh lockt und sucht.
   Lisa now guarantees *Anna / Thomas* that she will lure and search the shy deer.

10. Lisa schwört nun *Anna / Peter*, dass sie das schöne Pferd sattelt und zäumt.
    Lisa now swears *Anna / Peter* that she will saddle and bridle the nice horse.

A2.2 Estimates of the bootstrapped divergence onsets

Table 12: Bootstrapped onsets of the divergence of the fixation path (1) between participant groups, (2) from 50% target fixations, and (3) between correctly and incorrectly answered trials. Divergence onsets are represented as mean in milliseconds with a bootstrapped 95% confidence interval [CI].

	Declarative sentences	Relative clauses	Control structures with pronoun	Control structures with PRO
(1) Group				
test	s: 1535 [1400, 1800]	s: NA	s: 677 [500, 800]	s: 2456 [2300, 2750]
	c: 2243 [2100, 2500]	c: 1414 [1300, 1550]	c: 626 [500, 750]	c: NA
retest	s: 1138 [1050, 1250]	s: 1367 [1100, 1650]	s: 760 [650, 850]	s: 2046 [1100, 2750]
	c: 1705 [1600, 1800]	c: 1333 [1250, 1450]	c: 1101 [950, 1250]	c: NA
(2) 50%				
IWA: test	s: 653 [200, 1200]	s: 524 [400, 650]	s: 1690 [1500, 1850]	s: 389 [300, 500]
	c: 2317 [2100, 2450]	c: 2760 [2550, 3100]	c: 2133 [1950, 2400]	c: 1532 [1350, 1650]
IWA: retest	s: 1248 [1100, 1400]	s: 1206 [1000, 1400]	s: 1172 [1050, 1300]	s: 404 [200, 600]
	c: 4787 [4250, 5100]	c: 2299 [1700, 2900]	c: 987 [750, 1250]	c: 945 [700, 1150]
CP: test	s: 814 [750, 850]	s: 334 [300, 350]	s: 7 [0, 100]	s: 211 [150, 250]
	c: 1560 [1500, 1650]	c: 1207 [1150, 1250]	c: 1 [0, 0]	c: 0 [0, 0]
CP: retest	s: 726 [650, 800]	s: 656 [600, 700]	s: 189 [150, 250]	s: 86 [50, 150]
	c: 1237 [1200, 1300]	c: 1080 [1050, 1100]	c: 10 [0, 100]	c: 318 [150, 500]
(3) Accuracy				
IWA: test	s: 2176 [1950, 2450]	s: 1435 [450, 2600]	s: 699 [500, 900]	s: 819 [450, 1250]
	c: 2143 [1950, 2400]	c: 2762 [1900, 3200]	c: 1189 [850, 1450]	c: 834 [600, 1000]
IWA: retest	s: 3013 [2750, 3300]	s: 1872 [1600, 2150]	s: 1225 [1050, 1500]	s: 1022 [750, 1250]
	c: 4245 [4000, 4600]	c: 1725 [1550, 2100]	c: 808 [650, 1000]	c: 1001 [800, 1150]

Note. IWA = individuals with aphasia, CP = control participants, s = simple canonical or low interference conditions, c = complex non-canonical or high interference conditions, NA = no divergence.

A2.3 Comparison of the fixation paths of 50 and 22 control participants

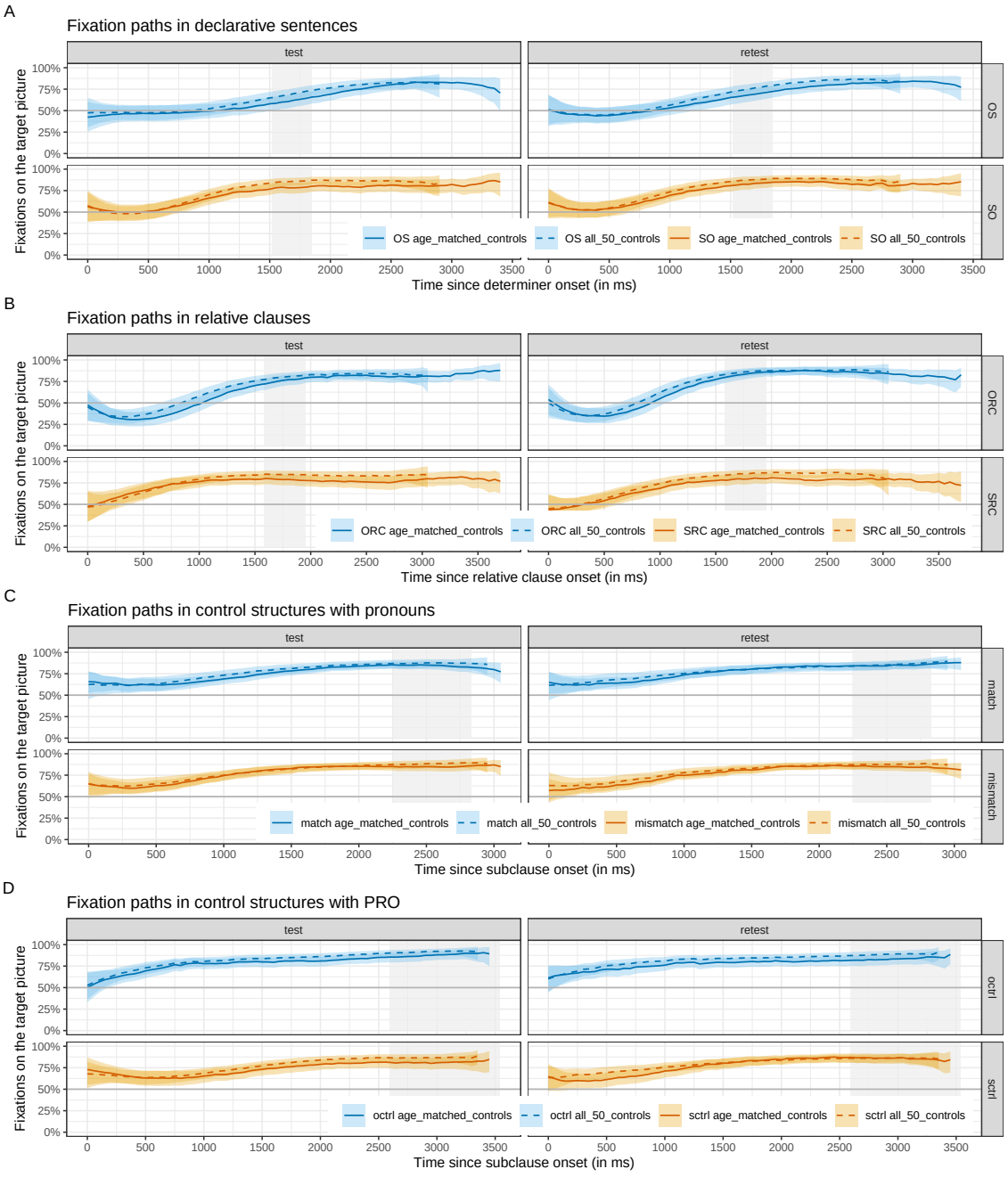

Figure 22: Estimated fixation paths of the whole control group (n = 50 participants, mean age = 48 years, SD = 20, range = 19–83; mean education = 18 years, SD = 4, range = 6–26, dashed lines) and the age and education matched control group (n = 22 participants, mean age = 58 years, SD = 15, range = 26–81; mean education = 16 years, SD = 4, range = 6–21, solid lines). A: canonical (SO) and non-canonical (OS) declaratives; B: subject (SRC) and object (ORC) relative clauses; C: control structures with a pronoun with gender matching (match) and mismatching (mismatch) nouns; D: subject (s-ctrl) and object (o-ctrl) control structures with PRO. Solid and dashed lines represent the mean fixations in simple and complex sentences respectively and shaded areas represent the 95% credible intervals around the mean. Vertical bands shaded in grey mark the sentence end.

A2.4 Estimates of the time window analysis comparing target fixations and aphasia severity

Table 13: Means and 95% Credible Intervals (CrI) for the influence of the aphasia severity grade in the Aachen Aphasia Test (AAT) on the target fixations in the second half of the sentence and in the silence region.

	Severity	Severity × Complexity	Severity × Trial Accuracy
Region 2			
declaratives	0.2% CrI: [-9, 9.6]	4.3% CrI: [-6, 15.1]	-5.1% CrI: [-15.3, 5.2]
RC	-3.6% CrI: [-10.4, 3]	6.8% CrI: [-4.1, 17.4]	-3% CrI: [-13.3, 7.2]
pronoun	-1.3% CrI: [-12.5, 9.9]	2.2% CrI: [-6.5, 10.7]	-8% CrI: [-21.3, 5.7]
PRO	-3.7% CrI: [-17.9, 10.8]	4.3% CrI: [-7.2, 15.7]	6% CrI: [-9.8, 21.6]
Region 3			
declaratives	1.4% CrI: [-4, 6.9]	-1.4% CrI: [-4.7, 2.1]	6% CrI: [0.4, 11.6]
RC	-1.8% CrI: [-7.5, 4]	-4% CrI: [-11.4, 3.2]	3.2% CrI: [-5, 11.5]
pronoun	-4.7% CrI: [-19.7, 9.9]	3.7% CrI: [-4.9, 12.4]	3% CrI: [-9.4, 15.7]
PRO	0.6% CrI: [-8.8, 10.2]	6.9% CrI: [-3.8, 17.4]	-2.7% CrI: [-16.2, 10.6]

Note. Region 2 = second half of the sentence after the critical word, Region 3 = silence region, RC = relative clauses, pronoun = control structures with a pronoun, PRO = control structures with PRO.

A2.5 Comparison of the response accuracy in this and previous visual world studies on sentence comprehension in aphasia

Table 14: Response accuracy in visual world studies on sentence comprehension in aphasia

Study and sentence type	Mean accuracy	Reported uncertainty measure	Reported uncertainty values	N IWA	N items	Calculated standard error
Adelt et al (2017) ORC, case marked, full DP	0.53	sd	0.501	10	16	0.16
Adelt et al (2017) ORC, case marked, pronoun	0.55	sd	0.499	10	16	0.16
Adelt et al (2017) ORC, number marked, full DP	0.59	sd	0.493	10	16	0.16
Adelt et al (2017) ORC, number marked, pronoun	0.71	sd	0.454	10	16	0.14
Adelt et al (2017) SRC, case marked, full DP	0.76	sd	0.431	10	16	0.14
Adelt et al (2017) SRC, case marked, pronoun	0.69	sd	0.465	10	16	0.15
Adelt et al (2017) SRC, number marked, full DP	0.61	sd	0.49	10	16	0.16
Adelt et al (2017) SRC, number marked, pronoun	0.51	sd	0.501	10	16	0.16
Bos et al (2014) Future Tense	0.64	sd	0.26	6	20	0.11
Bos et al (2014) Past Tense	0.92	sd	0.04	6	20	0.02
Choy & Thompson (2005) Pronouns	0.65	NA	NA	8	20	NA
Choy & Thompson (2005) Reflexives	0.64	NA	NA	8	20	NA
Choy & Thompson (2010) Pronouns	0.65	sd	0.141	8	20	0.05
Choy & Thompson (2010) Reflexives	0.64	sd	0.192	8	20	0.07
Dickey & Thompson (2009) ORC	0.36	range	0.16 to 0.75	8	12	NA
Dickey & Thompson (2009) Passives	0.20	range	0 to 0.67	8	12	NA
Dickey et al (2007) obj clefts	0.67	range	0.2 to 1	12	10	NA
Dickey et al (2007) wh-questions	0.70	range	0 to 1	12	10	NA
Dickey et al (2007) yes-no-questions	0.87	range	0.6 to 1	12	10	NA
Engel et al (2018) Pronouns, long	0.63	sd	0.2066	6	10	0.08
Engel et al (2018) Pronouns, short	0.70	sd	0.2683	6	10	0.11
Engel et al (2018) Reflexives, long	0.83	sd	0.1211	6	10	0.05
Engel et al (2018) Reflexives, short	0.90	sd	0.0894	6	10	0.04
Hanne et al (2011) OVS	0.46	se	0.0314	7	20	0.03
Hanne et al (2011) SVO	0.80	se	0.0426	7	20	0.04
Hanne et al (2015) OVS	0.46	sd	0.166	8	20	0.06
Hanne et al (2015) SVO	0.77	sd	0.116	8	20	0.04
Hanne et al (2016) object-questions	0.66	sd	0.186	6	20	0.08
Hanne et al (2016) subject-questions	0.64	sd	0.068	6	20	0.03
Mack & Thompson (2017) actives, post-treatment	0.74	sd	0.159	10	24	0.05
Mack & Thompson (2017) actives, pre-treatment	0.77	sd	0.123	10	24	0.04
Mack & Thompson (2017) passives, post-treatment	0.64	sd	0.169	10	24	0.05
Mack & Thompson (2017) passives, pre-treatment	0.53	sd	0.166	10	24	0.05
Mack et al (2016) actives, session 1	0.74	sd	0.189	12	24	0.06
Mack et al (2016) actives, session 2	0.75	sd	0.167	12	24	0.05
Mack et al (2016) passives, session 1	0.55	sd	0.18	12	24	0.05
Mack et al (2016) passives, session 2	0.61	sd	0.192	12	24	0.06
Meyer et al (2012) actives	0.76	sd	0.13	10	20	0.04
Meyer et al (2012) passives	0.54	sd	0.14	10	20	0.04
Pregla et al o-ctrl retest	0.76	sd	0.42	21	10	0.09
Pregla et al o-ctrl test	0.68	sd	0.46	21	10	0.10
Pregla et al ORC retest	0.45	sd	0.5	21	10	0.11
Pregla et al ORC test	0.40	sd	0.5	21	10	0.11
Pregla et al OS retest	0.44	sd	0.5	21	10	0.11
Pregla et al OS test	0.42	sd	0.5	21	10	0.11
Pregla et al Pronouns match retest	0.70	sd	0.44	21	10	0.10
Pregla et al Pronouns match test	0.50	sd	0.5	21	10	0.11
Pregla et al Pronouns mismatch retest	0.77	sd	0.39	21	10	0.08
Pregla et al Pronouns mismatch test	0.60	sd	0.47	21	10	0.10
Pregla et al s-ctrl retest	0.63	sd	0.48	21	10	0.10
Pregla et al s-ctrl test	0.50	sd	0.5	21	10	0.11
Pregla et al SO retest	0.78	sd	0.4	21	10	0.09
Pregla et al SO test	0.64	sd	0.47	21	10	0.10
Pregla et al SRC retest	0.67	sd	0.46	21	10	0.10
Pregla et al SRC test	0.67	sd	0.46	21	10	0.10
Schumacher et al (2015) OVS	0.36	se	0.08	12	16	0.08
Schumacher et al (2015) passives	0.75	se	0.073	12	16	0.07
Schumacher et al (2015) SVO	0.93	se	0.022	12	16	0.02
Sheppard et al (2015) object-which-questions	0.43	se	0.04	8	46	0.04
Sheppard et al (2015) object-who-questions	0.53	se	0.16	8	46	0.16
Sheppard et al (2015) subject-which-questions	0.54	se	0.12	8	46	0.12
Sheppard et al (2015) subject-who-questions	0.56	se	0.14	8	46	0.14
Thompson et al (2004) obj clefts	0.52	NA	NA	4	10	NA
Thompson et al (2004) wh-questions	0.92	NA	NA	4	10	NA
Thompson et al (2004) yes-no-questions	0.85	NA	NA	4	10	NA

Note. IWA = individuals with aphasia, SRC/ORC = subject / object relative clause, DP = determiner phrase, obj = object, SVO/SO = declarative sentence with subject (verb) object order, OVS/OS = declarative sentence with object (verb) subject order, o-ctrl / s-ctrl= object / subject control, se = standard error, sd = standard deviation, NA = not reported.

Bibliography

Adelt, A., Stadie, N., Lassotta, R., Adani, F., & Burchert, F. (2017). Feature dissimilarities in the processing of German relative clauses in aphasia. *Journal of Neurolinguistics*, *44*, 17–37.

Andrews, C. (2021). *There and gone again: Syntactic structure in memory* (Doctoral dissertation). University of Massachusetts Amherst.

Arantzeta, M., Bastiaanse, R., Burchert, F., Wieling, M., Martinez-Zabaleta, M., & Laka, I. (2017). Eye-tracking the effect of word order in sentence comprehension in aphasia: Evidence from Basque, a free word order ergative language. *Language, Cognition and Neuroscience*, *32*, 1320–1343.

Badecker, W., & Straub, K. (2002). The processing role of structural constraints on interpretation of pronouns and anaphors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 748–769.

Bader, M., & Häussler, J. (2010). Word order in German: A corpus study. *Lingua*, *120*, 717–762.

Bader, M., & Meng, M. (1999). Subject-object ambiguities in German embedded clauses: An across-the-board comparison. *Journal of Psycholinguistic Research*, *28*, 121–143.

Bader, M., & Meng, M. (2018). The misinterpretation of noncanonical sentences revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 1286.

Barbieri, E., Mack, J. E., Chiappetta, B., Europa, E., & Thompson, C. K. (2019). Recovery of offline and online sentence processing in aphasia: Language and domain-general network neuroplasticity. *Cortex*, *120*, 394–418.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Berndt, R. S., Mitchum, C. C., & Haendiges, A. N. (1996). Comprehension of reversible sentences in "agrammatism": A meta-analysis. *Cognition*, *58*, 289–308.

Boersma, P., & Weenink, D. (2018). *Praat: Doing phonetics by computer [computer program]. version 6.0.37*. Retrieved February 13, 2018, from http://www.praat.org/

Bos, L. S., Hanne, S., Wartenburger, I., & Bastiaanse, R. (2014). Losing track of time? Processing of time reference inflection in agrammatic and healthy speakers of German. *Neuropsychologia*, *65*, 180–190.

Boyle, M. (2014). Test–retest stability of word retrieval in aphasic discourse. *Journal of Speech, Language, and Hearing Research*, *57*, 966–978.

Broca, P. (1861). Remarques sur le siège de la faculté du langage articulé, suivies d'une observation d'aphémie (perte de la parole). *Bulletins et memoires de la Société Anatomique de Paris*, *36*, 330–357.

Brookshire, R. H., & Nicholas, L. E. (1994). Test-retest stability of measures of connected speech in aphasia. *Clinical Aphasiology*, *22*, 119–133.

Burchert, F., Hanne, S., & Vasishth, S. (2013). Sentence comprehension disorders in aphasia: The concept of chance performance revisited. *Aphasiology*, *27*, 112–125.

Burchert, F., Lorenz, A., Schröder, A., De Bleser, R., & Stadie, N. (2011). *Sätze verstehen. Neurolinguistische Materialien für die Untersuchung von syntaktischen Störungen beim Satzverständnis.* NAT-Verlag.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*, 395–411.

Caplan, D. (1987). *Neurolinguistics and linguistic aphasiology: An introduction.* Cambridge University Press.

Caplan, D. (2001). The measurement of chance performance in aphasia, with specific reference to the comprehension of semantically reversible passive sentences: A note on issues raised by Caramazza, Capitani, Rey, and Berndt (2001) and Drai, Grodzinsky, and Zurif (2001). *Brain and Language*, *76*, 193–201.

Caplan, D. (2010). Task effects on bold signal correlates of implicit syntactic processing. *Language and Cognitive Processes*, *25*, 866–901.

Caplan, D. (2012). Resource reduction accounts of syntactically based comprehension disorders. In R. Bastiaanse & C. K. Thompson (Eds.), *Perspectives on agrammatism* (pp. 48–62). Psychology Press.

Caplan, D., Chen, E., & Waters, G. (2008). Task-dependent and task-independent neurovascular responses to syntactic processing. *Cortex*, *44*, 257–275.

Caplan, D., DeDe, G., & Michaud, J. (2006). Task-independent and task-specific syntactic deficits in aphasic comprehension. *Aphasiology*, *20*, 893–920.

Caplan, D., DeDe, G., Waters, G., Michaud, J., & Tripodis, Y. (2011). Effects of age, speed of processing, and working memory on comprehension of sentences with relative clauses. *Psychology and Aging*, *26*, 439–450.

Caplan, D., & Hildebrandt, N. (1988). *Disorders of syntactic comprehension.* MIT Press.

Caplan, D., Michaud, J., & Hufford, R. (2015). Mechanisms underlying syntactic comprehension deficits in vascular aphasia: New evidence from self-paced listening. *Cognitive Neuropsychology*, *32*, 283–313.

Caplan, D., Michaud, J., & Hufford, R. (2013a). Dissociations and associations of performance in syntactic comprehension in aphasia and their implications for the nature of aphasic deficits. *Brain and Language*, *127*, 21–33.

Caplan, D., Michaud, J., & Hufford, R. (2013b). Short-term memory, working memory, and syntactic comprehension in aphasia. *Cognitive Neuropsychology*, *30*, 77–109.

Caplan, D., & Waters, G. (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory*, *13*, 403–413.

Caplan, D., Waters, G., DeDe, G., Michaud, J., & Reddy, A. (2007). A study of syntactic processing in aphasia I: Behavioral (psycholinguistic) aspects. *Brain and Language*, *101*, 103–150.

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*, 77–94.

Caplan, D., Waters, G. S., & Hildebrandt, N. (1997). Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks. *Journal of Speech, Language, and Hearing Research*, *40*, 542–555.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*, 41–66.

Caramazza, A., Capasso, R., Capitani, E., & Miceli, G. (2005). Patterns of comprehension performance in agrammatic Broca's aphasia: A test of the Trace Deletion Hypothesis. *Brain and Language*, *94*, 43–53.

Caramazza, A., Capitani, E., Rey, A., & Berndt, R. S. (2001). Agrammatic Broca's aphasia is not associated with a single pattern of comprehension performance. *Brain and Language*, *76*, 158–184.

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*, 572–582.

Chang, F., Janciauskas, M., & Fitz, H. (2012). Language adaptation and learning: Getting explicit about implicit learning. *Language and Linguistics Compass*, *6*, 259–278.

Chomsky, N. (1965). *Aspects of the theory of syntax*. M.I.T. press.

Choy, J. J., & Thompson, C. K. (2005). Online comprehension of anaphor and pronoun constructions in Broca's aphasia: Evidence from eyetracking. *Brain and Language*, *95*, 119–120.

Choy, J. J., & Thompson, C. K. (2010). Binding in agrammatic aphasia: Processing to comprehension. *Aphasiology*, *24*, 551–579.

Christiansen, M. H., Kelly, M. L., Shillcock, R. C., & Greenfield, K. (2010). Impaired artificial grammar learning in agrammatism. *Cognition*, *116*, 382–393.

Ciccone, N. A. (2003). *The measurement of stability in aphasia recovery: Implications for language modelling* (Doctoral dissertation). Curtin University.

Connor, T. L., Albert, M. L., Helm-Estabrooks, N., & Obler, L. (2000). Attentional modulation of language performance. *Brain and Language*, *71*, 52–55.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 84–107.

Cope, T. E., Sohoglu, E., Sedley, W., Patterson, K., Jones, P. S., Wiggins, J., Dawson, C., Grube, M., Carlyon, R. P., Griffiths, T. D., Davis, M. H., & Rowe, J. B. (2017). Evidence for causal top-down frontal contributions to predictive processes in speech perception. *Nature Communications*, *8*, 1–16.

Cupples, L., & Inglis, A. (1993). When task demands induce "asyntactic" comprehension: A study of sentence interpretation in aphasia. *Cognitive Neuropsychology*, *10*, 201–234.

De Bleser, R., Schwarz, W., & Burchert, F. (2006). Quantitative neurosyntactic analyses: The final word? *Brain and Language*, *96*, 143–146.

De Schotten, M. T., & Shallice, T. (2017). Identical, similar or different? Is a single brain model sufficient? *Cortex*, *86*, 172–175.

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20120394.

Dempsey, J., Liu, Q., & Christianson, K. (2020). Convergent probabilistic cues do not trigger syntactic adaptation: Evidence from self-paced reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*, 1906–1921.

Des Roches, C. A., Vallila-Rohter, S., Villard, S., Tripodis, Y., Caplan, D., & Kiran, S. (2016). Evaluating treatment and generalization patterns of two theoretically motivated sentence comprehension therapies. *American Journal of Speech-Language Pathology*, *25*, S743–S757.

Dickey, M. W., Choy, J. J., & Thompson, C. K. (2007). Real-time comprehension of wh-movement in aphasia: Evidence from eyetracking while listening. *Brain and Language*, *100*, 1–22.

Dickey, M. W., & Thompson, C. K. (2009). Automatic processing of wh-and NP-movement in agrammatic aphasia: Evidence from eyetracking. *Journal of Neurolinguistics*, *22*, 563–583.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103.

Dominey, P. F., Hoen, M., Blanc, J.-M., & Lelekov-Boissard, T. (2003). Neurological basis of language and sequential cognition: Evidence from simulation, aphasia, and ERP studies. *Brain and Language*, *86*, 207–225.

Drai, D., & Grodzinsky, Y. (1999). Comprehension regularity in Broca's aphasia? There's more of it than you ever imagined. *Brain and Language*, *70*, 139–143.

Drai, D., & Grodzinsky, Y. (2006). A new empirical angle on the variability debate: Quantitative neurosyntactic analyses of a large data set from Broca's aphasia. *Brain and Language*, *96*, 117–128.

Duncan, E. S., Schmah, T., & Small, S. L. (2016). Performance variability as a predictor of response to aphasia treatment. *Neurorehabilitation and Neural Repair*, *30*, 876–882.

Dykiert, D., Der, G., Starr, J. M., & Deary, I. J. (2012). Age differences in intra-individual variability in simple and choice reaction time: Systematic review and meta-analysis. *PLoS ONE*, *7*, e45759.

Engel, S., Shapiro, L. P., & Love, T. (2018). Proform-antecedent linking in individuals with agrammatic aphasia: A test of the intervener hypothesis. *Journal of Neurolinguistics*, *45*, 79–94.

Farris-Trimble, A., & McMurray, B. (2013). Test–retest reliability of eye tracking in the visual world paradigm for the study of real-time spoken word recognition. *Journal of Speech, Language, and Hearing Research*, *56*, 1328–1345.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLoS ONE*, *8*, e77661.

Fine, A. B., Qian, T., Jaeger, T. F., & Jacobs, R. (2010). Is there syntactic adaptation in language comprehension? *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*, 18–26.

Flanagan, J. L., & Jackson, S. T. (1997). Test–retest reliability of three aphasia tests: Performance of non-brain-damaged older adults. *Journal of Communication Disorders*, *30*, 33–43.

Freed, D. B., Marshall, R. C., & Chuhlantseff, E. A. (1996). Picture naming variability: A methodological consideration of inconsistent naming responses in fluent and nonfluent aphasia. *Clinical Aphasiology*, *24*, 193–205.

Friedman, N. P., & Miyake, A. (2017). Unity and diversity of executive functions: Individual differences as a window on cognitive structure. *Cortex*, *86*, 186–204.

Friedmann, N. (2008). Traceless relatives: Agrammatic comprehension of relative clauses with resumptive pronouns. *Journal of Neurolinguistics*, *21*, 138–149.

Friedmann, N., Reznick, J., Dolinski-Nuger, D., & Soboleva, K. (2010). Comprehension and production of movement-derived sentences by Russian speakers with agrammatic aphasia. *Journal of Neurolinguistics*, *23*, 44–65.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). *irr: Various coefficients of interrater reliability and agreement* [R package version 0.84.1]. https://CRAN.R-project.org/package=irr

Garraffa, M., & Grillo, N. (2008). Canonicity effects as grammatical phenomena. *Journal of Neurolinguistics*, *21*, 177–197.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel / hierarchical models*. Cambridge.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, 95–126.

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*, 8051–8056.

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

Goldstein, K. (1948). *Language and language disturbances; aphasic symptom complexes and their significance for medicine and theory of language*. Grune & Stratton.

Goschke, T., Friederici, A. D., Kotz, S. A., & Van Kampen, A. (2001). Procedural learning in Broca's aphasia: Dissociation between the implicit acquisition of spatio-motor and phoneme sequences. *Journal of Cognitive Neuroscience*, *13*, 370–388.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive Science*, *29*, 261–290.

Grodzinsky, Y. (1986). Language deficits and the theory of syntax. *Brain and Language*, *27*, 135–159.

Grodzinsky, Y. (1995). A restrictive theory of trace deletion in agrammatism. *Brain and Language*, *50*, 27–51.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences*, *23*, 1–21.

Grodzinsky, Y., Piñango, M. M., Zurif, E., & Drai, D. (1999). The critical role of group studies in neuropsychology: Comprehension regularities in Broca's aphasia. *Brain and Language*, *67*, 134–147.

Hageman, C. F., McNeil, M. R., Rucci-Zimmer, S., & Cariski, D. M. (1982). The reliability of patterns of auditory processing deficits: Evidence from the Revised Token Test. *Clinical Aphasiology: Proceedings of the Conference 1982*, 230–234.

Hahn, M., & Keller, F. (2018). Modeling task effects in human reading with neural attention. *arXiv preprint arXiv:1808.00054*.

Halai, A. D., Woollams, A. M., & Ralph, M. A. L. (2017). Using principal component analysis to capture individual differences within a unified neuropsychological model of chronic post-stroke aphasia: Revealing the unique neural correlates of speech fluency, phonology and semantics. *Cortex*, *86*, 275–289.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 1–8.

Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2012). *Offline chance performance in sentence comprehension: What single-case analyses reveal about online sentence processing in aphasia.* [Poster presented at the Thirtieth European Workshop on Cognitive Neuropsychology. Bressanone, Italy].

Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2015). Sentence comprehension and morphological cues in aphasia: What eye-tracking reveals about integration and prediction. *Journal of Neurolinguistics*, *34*, 83–111.

Hanne, S., Burchert, F., & Vasishth, S. (2016). On the nature of the subject–object asymmetry in wh-question comprehension in aphasia: Evidence from eye tracking. *Aphasiology*, *30*, 435–462.

Hanne, S., Sekerina, I. A., Vasishth, S., Burchert, F., & De Bleser, R. (2011). Chance in agrammatic sentence comprehension: What does it really mean? evidence from eye movements of German agrammatic aphasic patients. *Aphasiology*, *25*, 221–244.

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory & Cognition*, *46*, 864–877.

Härting, C., Markowitsch, H., Neufeld, H., Calabrese, P., Deisinger, K., & Kessler, J. (2000). Wechsler Memory Scale – Revised Edition, German Edition. *Manual. Bern: Huber*.

Head, H. (1920). Aphasia and kindred disorders of speech (the Linacre Lecture for 1920.) *Brain*, *43*, 87–165.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Hanneforth, T., Geyken, A., & Kliegl, R. (2011). DlexDB–eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, *62*, 10–20.

Henry, N., Hopp, H., & Jackson, C. N. (2017). Cue additivity and adaptivity in predictive processing. *Language, Cognition and Neuroscience*, *32*, 1229–1249.

Huang, Y., & Snedeker, J. (2020). Evidence from the visual world paradigm raises questions about unaccusativity and growth curve analyses. *Cognition*, *200*, 104251.

Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *AAT-Aachener Aphasie Test*. Hogrefe.

Hughlings Jackson, J. (1878). On affections of speech from disease of the brain. *Brain*, *1*, 304–330.

Hula, W. D., & McNeil, M. R. (2008). Models of attention and dual-task performance as explanatory constructs in aphasia. *Seminars in Speech and Language*, *29*, 169–187.

Hula, W. D., McNeil, M. R., & Sung, J. E. (2007). Is there an impairment of language-specific attentional processing in aphasia? *Brain and Language*, *103*, 240–241.

Hultsch, D. F., MacDonald, S. W., & Dixon, R. A. (2002). Variability in reaction time performance of younger and older adults. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *57*, 101–115.

Hultsch, D. F., Strauss, E., Hunter, M. A., & MacDonald, M. A. (2011). Intraindividual variability, cognition, and aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *The Handbook of Aging and Cognition* (3rd ed., pp. 491–556). Psychology Press.

Jäger, L. A., Engelmann, F., & Vasishth, S. (2017). Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language*, *94*, 316–339.

James, A. N., Fraundorf, S. H., Lee, E.-K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions? *Journal of Memory and Language*, *102*, 155–181.

Joanette, Y., & Small, S. (2000). Brain and Language in the millennium. *Brain and Language*, *71*, 1–3.

Johnson, D., & Cannizzaro, M. S. (2009). Sentence comprehension in agrammatic aphasia: History and variability to clinical implications. *Clinical Linguistics & Phonetics*, *23*, 15–37.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329–354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122.

Kaan, E., & Chun, E. (2018). Syntactic adaptation. In K. D. Federmeier & D. G. Watson (Eds.), *Current topics in language*. Academic Press.

Kiran, S., Caplan, D., Sandberg, C., Levy, J., Berardino, A., Ascenso, E., Villard, S., & Tripodis, Y. (2012). Development of a theoretically based treatment for sentence comprehension deficits in individuals with aphasia. *American Journal of Speech-Language Pathology*, *21*, 88–102.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., & Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, *1*, 238.

Kolk, H. H., & Van Grunsven, M. M. (1985). Agrammatism as a variable phenomenon. *Cognitive Neuropsychology*, *2*, 347–384.

Kroczek, L. O., & Gunter, T. C. (2017). Communicative predictions can overrule linguistic priors. *Scientific Reports*, *7*, 1–9.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*, 155–177.

Kwon, N., & Sturt, P. (2016). Processing control information in a nominal control construction: An eye-tracking study. *Journal of Psycholinguistic Research*, *45*, 779–793.

Laures, J. S. (2005). Reaction time and accuracy in individuals with aphasia during auditory vigilance tasks. *Brain and Language*, *95*, 353–357.

Levelt, W. J. (2001). Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences*, *98*, 13464–13471.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*, 1989–2001.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

Lissón, P., Pregla, D., Nicenboim, B., Paape, D., van het Nederend, M. L., Burchert, F., Stadie, N., Caplan, D., & Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science*, *45*, e12956.

Luzzatti, C., Toraldo, A., Guasti, M. T., Ghirardi, G., Lorenzi, L., & Guarnaschelli, C. (2001). Comprehension of reversible active and passive sentences in agrammatism. *Aphasiology*, *15*, 419–441.

Mack, J. E., & Thompson, C. K. (2017). Recovery of online sentence processing in aphasia: Eye movement changes resulting from treatment of underlying forms. *Journal of Speech, Language, and Hearing Research*, *60*, 1299–1315.

Mack, J. E., Wei, A. Z.-S., Gutierrez, S., & Thompson, C. K. (2016). Tracking sentence comprehension: Test-retest reliability in people with aphasia and unimpaired adults. *Journal of Neurolinguistics*, *40*, 98–111.

Martín-Loeches, M., Ouyang, G., Rausch, P., Stürmer, B., Palazova, M., Schacht, A., & Sommer, W. (2017). Test–retest reliability of the N400 component in a sentence-reading paradigm. *Language, Cognition and Neuroscience*, *32*, 1261–1272.

Mätzig, P., Vasishth, S., Engelmann, F., Caplan, D., & Burchert, F. (2018). A computational investigation of sources of variability in sentence comprehension difficulty in aphasia. *Topics in Cognitive Science*, *10*, 161–174.

McMurray, B. (2020). *I'm not sure that curve means what you think it means: Toward a [more] realistic understanding of the role of eye-movement generation in the visual world paradigm.* Retrieved October 19, 2020, from https://psyarxiv.com/pb2c6/

McNeil, M. R. (1983). Aphasia: Neurological considerations. *Topics in Language Disorders*, *3*, 1–20.

McNeil, M. R. (1988). Aphasia in the Adult. In N. J. Lass, L. V. McReyolds, J. L. Northern, & D. E. Yoder (Eds.), *Handbook of Speech-Language Pathology and Audiology* (pp. 738–786). B.C. Decker Inc.

McNeil, M. R., Dionigi, C. M., Langlois, A., & Prescott, T. E. (1989). A measure of revised token test ordinality and intervality. *Aphasiology*, *3*, 31–40.

McNeil, M. R., & Doyle, P. J. (2000). Reconsidering the hegemony of linguistic explanations in aphasia: The challenge for the beginning of the millennium. *Brain and Language*, *71*, 154–156.

McNeil, M. R., Hageman, C., & Matthews, C. (2005). CAC classics: Auditory processing deficits in aphasia evidenced on the Revised Token Test: Incidence and prediction of across subtest and across item within subtest patterns. *Aphasiology*, *19*, 179–198.

McNeil, M. R., Odell, K., & Tseng, C.-H. (1991). Toward the integration of resource allocation into a general theory of aphasia. *Clinical Aphasiology*, *20*, 21–39.

McNeil, M. R., Odell, K., & Campbell, T. F. (1982). The frequency and amplitude of fluctuating auditory processing in aphasic and nonaphasic brain-damaged persons. *Clinical Aphasiology*, 220–229.

McNeil, M. R., & Pratt, S. R. (2001). Defining aphasia: Some theoretical and clinical implications of operating from a formal definition. *Aphasiology*, *15*, 901–911.

McNeil, M. R., Pratt, S. R., Szuminsky, N., Sung, J. E., Fossett, T. R., Fassbinder, W., & Lim, K. Y. (2015). Reliability and validity of the Computerized Revised Token Test: Comparison of reading and listening versions in persons with and without aphasia. *Journal of Speech, Language, and Hearing Research*, *58*, 311–324.

McNeil, M. R., & Prescott, T. E. (1978). *Revised Token Test*. University Park Press.

Meyer, A. M., Mack, J. E., & Thompson, C. K. (2012). Tracking passive sentence comprehension in agrammatic aphasia. *Journal of Neurolinguistics*, *25*, 31–43.

Mirman, D. (2014). *Growth curve analysis and visualization using R* (1st ed.). CRC Press.

Miyake, A., Carpenter, P. A., & Just, M. A. (1994). A capacity approach to syntactic comprehension disorders: Making normal adults perform like aphasic patients. *Cognitive Neuropsychology*, *11*, 671–717.

Murray, L. L. (2000). The effects of varying attentional demands on the word retrieval skills of adults with aphasia, right hemisphere brain damage, or no brain damage. *Brain and Language*, *72*, 40–72.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, *53*, 695–699.

Navarro, D. (2013). *Learning statistics with R: A tutorial for psychology students and other beginners: Version 0.5*. University of Adelaide Adelaide, Australia.

Nespoulous, J.-L. (2000). Invariance vs variability in aphasic performance. An example: Agrammatism. *Brain and Language*, *71*, 167–171.

Nicenboim, B., Schad, D. J., & Vasishth, S. (2022). *Introduction to Bayesian data analysis for cognitive science*. Under contract with Chapman; Hall/CRC Statistics in the Social; Behavioral Sciences.

Nickels, L., Howard, D., & Best, W. (2011). On the use of different methodologies in cognitive neuropsychology: Drink deep and from several sources. *Cognitive Neuropsychology*, *28*, 475–485.

Nozari, N., Mirman, D., & Thompson-Schill, S. L. (2016). The ventrolateral prefrontal cortex facilitates processing of sentential context to locate referents. *Brain and Language*, *157*, 1–13.

Odell, K. H., Hashi, M., Miller, S. B., & McNeil, M. R. (1995). A critical look at the notion of selective impairment. *Clinical Aphasiology*, *23*, 1–8.

Oldfield, R. C., et al. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*, 97–113.

Palmer, C. E., Langbehn, D., Tabrizi, S. J., & Papoutsi, M. (2018). Test–retest reliability of measures commonly used to measure striatal dysfunction across multiple testing sessions: A longitudinal study. *Frontiers in Psychology*, *8*, 2363.

Park, G. H., McNeil, M. R., & Tompkins, C. A. (2000). Reliability of the Five-Item Revised Token Test for individuals with aphasia. *Aphasiology*, *14*, 527–535.

Patil, U., Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2016). A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science*, *40*, 5–50.

Perez Naranjo, N., Del Río Grande, D., & González Alted, C. (2018). Individual variability in attention and language performance in aphasia: A study using Conner's Continuous Performance Test. *Aphasiology*, *32*, 436–458.

Porch, B. E. (1971). *The Porch Index of Communicative Ability*. Consulting Psychologists Press.

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*.

Pregla, D., Lissón, P., Vasishth, S., Burchert, F., & Stadie, N. (2021). Variability in sentence comprehension in aphasia in German. *Brain and Language*, *222*, 105008.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Reichle, E. D., Warren, T., & McConnell, K. (2009). Using EZ Reader to model the effects of higher level language processing on eye movements during reading. *Psychonomic Bulletin & Review*, *16*, 1–21.

Rohde, D. (2003). *Linger: A flexible platform for language processing experiments [computer program]. version 2.94*. Retrieved February 24, 2018, from http://tedlab.mit.edu/~dr/Linger/

Salis, C., & Edwards, S. (2009). Tests of syntactic comprehension in aphasia: An investigation of task effects. *Aphasiology*, *23*, 1215–1230.

Salverda, A. P., Brown, M., & Tanenhaus, M. K. (2011). A goal-based perspective on eye movements in visual world studies. *Acta Psychologica*, *137*, 172–180.

Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, *110*, 104038.

Schlesewsky, M., Bornkessel, I., & Frisch, S. (2003). The neurophysiological basis of word order variations in German. *Brain and Language*, *86*, 116–128.

Schroeder, S. (2007). *Interaktion gedächtnis-und erklärungs-basierter Verarbeitungsprozesse bei der pronominalen Auflösung. Analyse der Effekte von impliziten Kausalitäts- und Gender-Informationen durch die Modellierung von Reaktionszeitverteilungen.* (Doctoral dissertation). Universität zu Köln.

Schuchard, J., Nerantzini, M., & Thompson, C. K. (2017). Implicit learning and implicit treatment outcomes in individuals with aphasia. *Aphasiology*, *31*, 25–48.

Schuchard, J., & Thompson, C. K. (2014). Implicit and explicit learning in individuals with agrammatic aphasia. *Journal of Psycholinguistic Research*, *43*, 209–224.

Schumacher, R., Cazzoli, D., Eggenberger, N., Preisig, B., Nef, T., Nyffeler, T., Gutbrod, K., Annoni, J.-M., & Müri, R. M. (2015). Cue recognition and integration-eye tracking evidence of processing differences in sentence comprehension in aphasia. *PLoS ONE*, *10*, e0142853.

Schwartz, M. F., & Dell, G. S. (2010). Case series investigations in cognitive neuropsychology. *Cognitive Neuropsychology*, *27*, 477–494.

Shallice, T. (2015). Cognitive neuropsychology and its vicissitudes: The fate of Caramazza's axioms. *Cognitive Neuropsychology*, *32*, 385–411.

Shammi, P., Bosman, E., & Stuss, D. T. (1998). Aging and variability in performance. *Aging, Neuropsychology, and Cognition*, *5*, 1–13.

Sharma, S., Kim, H., Harris, H., Haberstroh, A., Wright, H. H., & Rothermich, K. (2021). Eye tracking measures for studying language comprehension deficits in aphasia: A systematic search and scoping review. *Journal of Speech, Language, and Hearing Research*, *64*, 1008–1022.

Sheppard, S. M., Walenski, M., Love, T., & Shapiro, L. P. (2015). The auditory comprehension of wh-questions in aphasia: Support for the intervener hypothesis. *Journal of Speech, Language, and Hearing Research*, *58*, 781–797.

Stadie, N., Cholewa, J., & De Bleser, R. (2013). *LEMO 2.0: Lexikon modellorientiert: Diagnostik für Aphasie, Dyslexie und Dysgraphie*. NAT-Verlag.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts. *Journal of Memory and Language*, *42*, 423–443.

Stiebels, B. (2007). Towards a typology of complement control. *ZAS Papers in Linguistics*, *47*, 1–58.

Stiebels, B., McFadden, T., Schwabe, K., Solstad, T., Kellner, E., Sommer, L., & Stoltmann, K. (2018). *ZAS database of clause-embedding predicates, release 1.0 (january, 2018)*. Institut für Deutsche Sprache, Mannheim. http://www.owid.de/plus/zasembed

Stone, K., von der Malsburg, T., & Vasishth, S. (2020). The effect of decay and lexical uncertainty on processing long-distance dependencies in reading. *PeerJ*, *8*, e10438.

Stowe, L. A., Kaan, E., Sabourin, L., & Taylor, R. C. (2018). The sentence wrap-up dogma. *Cognition*, *176*, 232–247.

Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales: A practical guide to their development and use* (5th ed.). Oxford University Press.

Sturt, P. (2003). The time-course of the application of binding constraints in reference resolution. *Journal of Memory and Language*, *48*, 542–562.

Sullivan, N., Walenski, M., Love, T., & Shapiro, L. P. (2017). The comprehension of sentences with unaccusative verbs in aphasia: A test of the intervener hypothesis. *Aphasiology*, *31*, 67–81.

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

Tesak, J., & Code, C. (2008). *Milestones in the history of aphasia: Theories and protagonists*. Psychology Press.

Thompson, C. K., Dickey, M. W., & Choy, J. J. (2004). Complexity in the comprehension of wh-movement structures in agrammatic Broca's aphasia: Evidence from eye-tracking. *Brain and Language*, *91*, 124–125.

Toraldo, A., & Luzzatti, C. (2006). Which variability? *Brain and Language*, *96*, 154–156.

Vadinova, V., Buivolova, O., Dragoy, O., van Witteloostuijn, M., & Bos, L. S. (2020). Implicit-statistical learning in aphasia and its relation to lesion location. *Neuropsychologia*, *147*, 107591.

Varkanitsa, M., & Caplan, D. (2018). On the association between memory capacity and sentence comprehension: Insights from a systematic review and meta-analysis of the aphasia literature. *Journal of Neurolinguistics*, *48*, 4–25.

Varlokosta, S., Nerantzini, M., Papadopoulou, D., Bastiaanse, R., & Beretta, A. (2014). Minimality effects in agrammatic comprehension: The role of lexical restriction and feature impoverishment. *Lingua*, *148*, 80–94.

Vasishth, S., Schad, D. J., Bürki, A., & Kliegl, R. (2022). *Linear mixed models for linguistics and psychology: A comprehensive introduction*. Under contract with Chapman; Hall/CRC Statistics in the Social; Behavioral Sciences.

Villard, S., & Kiran, S. (2015). Between-session intra-individual variability in sustained, selective, and integrational non-linguistic attention in aphasia. *Neuropsychologia*, *66*, 204–212.

Villard, S., & Kiran, S. (2018). Between-session and within-session intra-individual variability in attention in aphasia. *Neuropsychologia*, *109*, 95–106.

Vogelzang, M., Thiel, C. M., Rosemann, S., Rieger, J., & Ruigendijk, E. (2019). Cognitive abilities to explain individual variation in the interpretation of complex sentences by older adults. *Proceedings of the 41th Annual Conference of the Cognitive Science Society*, 3036–3042.

Warren, T., Dickey, M. W., & Lei, C.-M. (2016). Structural prediction in aphasia: Evidence from either. *Journal of Neurolinguistics*, *39*, 38–48.

Warren, T., Dickey, M. W., & Liburd, T. L. (2017). A rational inference approach to group and individual-level sentence comprehension performance in aphasia. *Cortex*, *92*, 19–31.

Weigl, E., & Bierwisch, M. (1970). Neuropsychology and linguistics: Topics of common research. *Foundations of Language*, *6*, 1–18.

Weiss, A. F., Kretzschmar, F., Schlesewsky, M., Bornkessel-Schlesewsky, I., & Staub, A. (2018). Comprehension demands modulate re-reading, but not first-pass reading behavior. *Quarterly Journal of Experimental Psychology*, *71*, 198–210.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.

Wendt, D., Brand, T., & Kollmeier, B. (2014). An eye-tracking paradigm for analyzing the processing time of sentences with different linguistic complexities. *PLoS ONE*, *9*, e100186.

Wernicke, C. (1874). *Der aphasische Symptomencomplex: Eine psychologische Studie auf anatomischer Basis*. Cohn & Weigert.

Yarbay Duman, T., Altınok, N., Özgirgin, N., & Bastiaanse, R. (2011). Sentence comprehension in Turkish Broca's aphasia: An integration problem. *Aphasiology*, *25*, 908–926.

Zakariás, L., & Lukács, Á. Between-session intraindividual variability in phonological, lexical, and semantic processing in post-stroke aphasia: A pilot study [Poster presented at the Academy of Aphasia 59th Annual Meeting.]. In: Poster presented at the Academy of Aphasia 59th Annual Meeting. Held online, 2021.

Zimmerer, V. C., Cowell, P. E., & Varley, R. A. (2014). Artificial grammar learning in individuals with severe aphasia. *Neuropsychologia*, *53*, 25–38.