Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Variability in sentence comprehension in aphasia in German

Dorothea Pregla [*], Paula Lissón, Shravan Vasishth, Frank Burchert, Nicole Stadie

University of Potsdam, Germany

ARTICLE INFO

Keywords:
Aphasia
Sentence Comprehension
Variability
Test-retest reliability
Task demands
Canonicity and interference effects
Object manipulation
Sentence-picture matching
Self-paced listening
Adaptation

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

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, Bosnian, & Stuss, 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, Kretzschmar, Schlesewsky, Bornkessel-Schlesewsky, & Staub, 2018), or fixation proportions (Salverda, Brown, & Tanenhaus, 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, Chen, & Waters, 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

[*] Corresponding author.
E-mail address: pregla@uni-potsdam.de (D. Pregla).

https://doi.org/10.1016/j.bandl.2021.105008
Received 14 September 2020; Received in revised form 4 June 2021; Accepted 10 August 2021
Available online 8 September 2021

deficit rather than a structural deficit (Caplan, Michaud, & Hufford, 2013a; Caplan, Waters, DeDe, Michaud, & Reddy, 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, DeDe, & Michaud, 2006; Caplan, Waters, & Hildebrandt, 1997; Caplan et al., 2007; Caplan et al., 2013a). 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., 2007, 2013a, 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 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

[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, Sandberg, Fedorenko, Bergen, & Kiran, 2016; Warren, Dickey, & Liburd, 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, Vasishth, Engelmann, Caplan, & Burchert, 2018; Patil, Hanne, Burchert, De Bleser, & Vasishth, 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.

[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, McNeil, Rucci-Zimmer, & Cariski, 1982; McNeil, 1983; McNeil, Hageman, & Matthews, 2005; McNeil, 1988).

with respect to brain responses (Martín-Loeches et al., 2017), fixation proportions (Farris-Trimble & McMurray, 2013; Mack, Wei, Gutierrez, & Thompson, 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, Schmah, & Small, 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, McNeil, & Tompkins, 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, Albert, Helm-Estabrooks, & Obler, 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, Odell, & Tseng, 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, McNeil, & Sung, 2007; Murray, 2000; Villard & Kiran, 2015). For example, Villard & 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 & 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 & 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 & 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 (Mack et al., 2016; McNeil et al., 2015; Caplan et al., 2006; Caplan et al., 1997; Caplan et al., 2007; Caplan et al., 2013a; Caplan, Michaud, & Hufford, 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., 2007; Caplan et al., 2013a; Caplan et al., 2015) 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., Caplan et al., 2006; Caplan et al., 2007; Caplan et al., 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, Thiel, Rosemann, Rieger, & Ruigendijk, 2019; e.g., for IWA: *English*: Caramazza & Zurif, 1976; *Greek*: Varlokosta, Nerantzini, Papadopoulou, Bastiaanse, & Beretta, 2014; *Hebrew*: Friedmann, 2008; *Italian*: Garraffa & Grillo, 2008; *Russian*: Friedmann, Reznick, Dolinski-Nuger, & Soboleva, 2010; *Turkish*: Yarbay Duman, Altınok, Özgirgin, & Bastiaanse, 2011). In addition to canonicity effects, we investigate the

(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.	(*non-canonical*)
	here	is	the	hedgehog	**who**$_{acc}$	**the**$_{nom}$	hamster	feeds	

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, Engelmann, & Vasishth, 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.

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, Stadie, Lassotta, Adani, & Burchert, 2017; *declarative sentences:* e.g., Hanne, Sekerina, Vasishth, Burchert, & De Bleser, 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, Bornkessel, & Frisch, 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, Shapiro, & Love, 2018; Sheppard, Walenski, Love, & Shapiro, 2015; Sullivan, Walenski, Love, & Shapiro, 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).

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).

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

(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	

(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	

Table 1
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)	T11 (n = 20)	Aphasia type	AAT[2] 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.

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; Caplan et al., 2006; Caplan et al., 2007; Caplan et al., 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; Caplan et al., 2007; Caplan et al., 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 & McMurray (2013), Mack et al. (2016), Palmer, Langbehn, Tabrizi, & Papoutsi (2018) and for IWA by Mack et al. (2016), 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

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

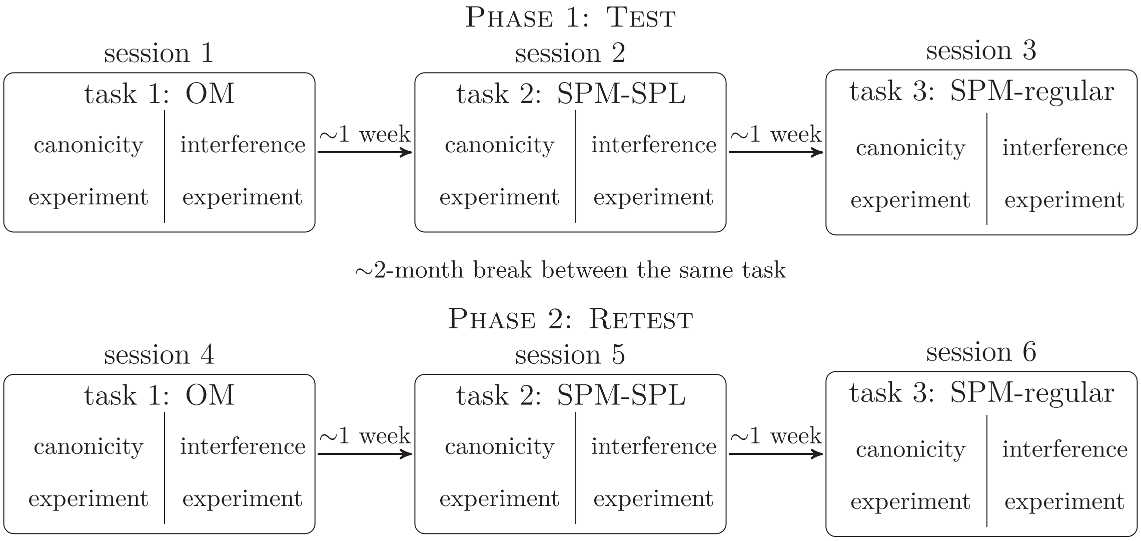

Fig. 1. General procedure of the study. All participants completed an object manipulation task (OM) and two versions of a sentence-picture matching task, in which the sentences were presented at a self-paced speed (SPM-SPL) or at a normal speech rate (SPM-regular). The three response tasks were completed 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.

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 1. 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, Poeck, Weniger, & Willmes, 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, Cholewa, & De Bleser, 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 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

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

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 s. 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 Fig. 2). 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 Fig. 2). 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 Fig. 1. 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 min) 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.

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 min in control participants and 30 min 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 (Harting 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

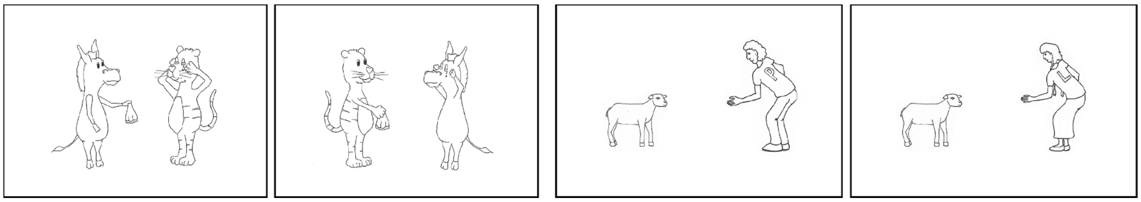

Fig. 2. Sample pictures of sentence-picture matching tasks. For the canonicity experiment (left pair), the canonical sentence *Here comforts the$_{nom}$ tiger the$_{acc}$ donkey* matches the right picture and the left picture is the foil, and conversely, the non-canonical sentence *Here comforts the$_{acc}$ 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.

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., Fig. 2). 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., Fig. 2). 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.

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 s or shorter than -1 s (i.e. when participants pressed a button more than 1 s 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, Machler, Bolker, & Walker, 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, 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. Furthermore, 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 Fig. A1 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.

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

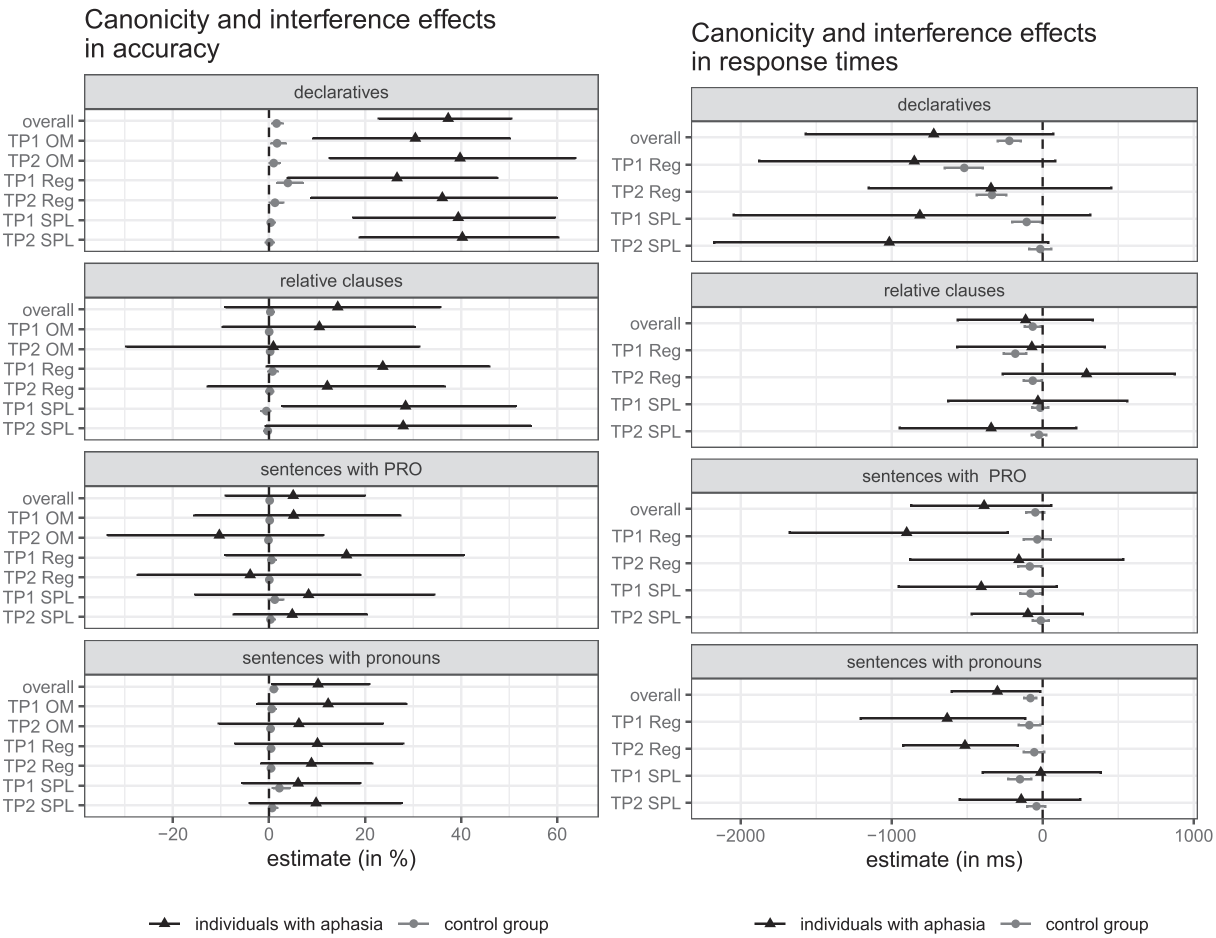

Fig. 3. Canonicity effects in declaratives and relative clauses and interference effects in sentences with a pronoun or PRO in the control group (gray) and the individuals with aphasia (black). Overall effects aggregated across test phases and response tasks, and separate effects in two test phases (TP1, 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).

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, Wei, Dambacher, Yan, & Zhou, 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, Norman, & Cairney, 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, Lemon, Fellows, & Singh, 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

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 A1 in the appendix. Considering the full data set, control participants had 26% CrI: [19, 34.3] higher accuracies and responded −2082 ms 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

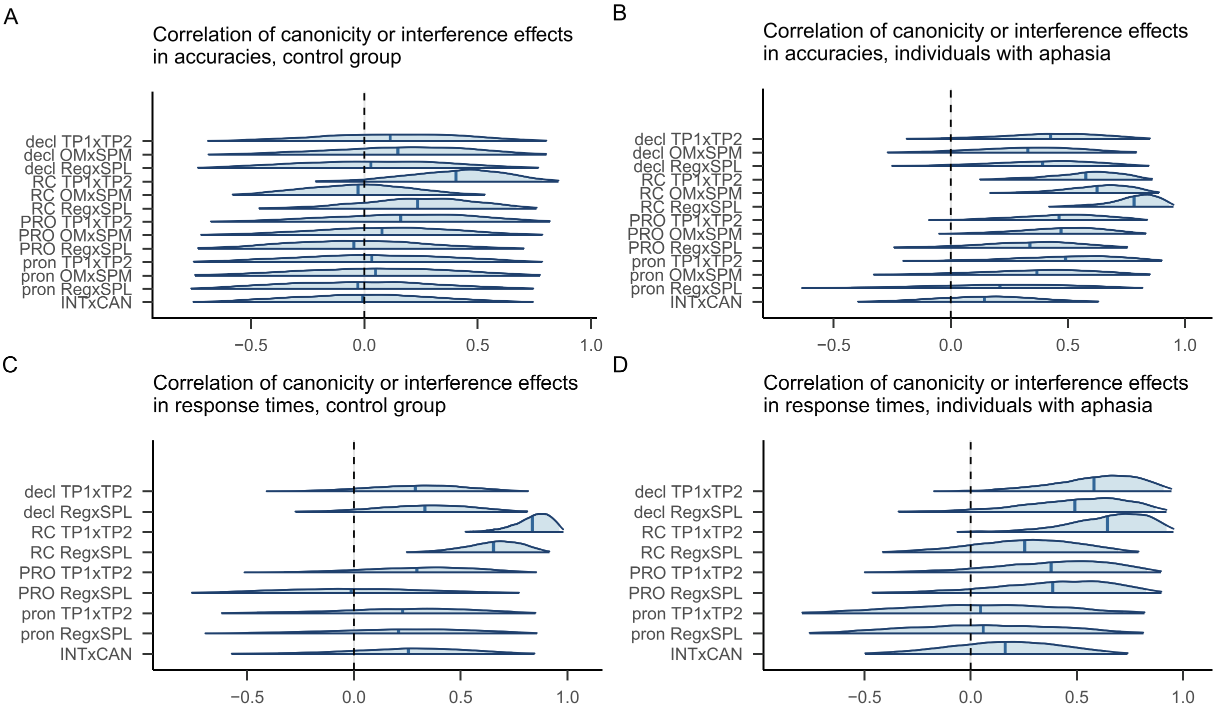

Fig. 4. Correlation of canonicity effects in declarative sentences (decl) and relative clauses (RC) and correlation of interference effects in sentences with pronouns (pron) and PRO in the control group (A, C) and the individuals with aphasia (B, D). The distributions display the posterior estimates 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).

between response tasks, they responded slower in regular than in self-paced sentence-picture matching (control group, test phase: 171 ms CrI: [67, 279], retest phase: 302 ms CrI: [228, 379], IWA, test phase: 504 ms CrI: [-112, 1141], retest phase: 672 ms CrI: [-145, 1499]). Both participant groups answered faster in the retest phase (control group: −146 ms CrI: [-196, −98], IWA: −303 ms 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 Fig. 3 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.

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 −220 ms CrI: [-299, −144]; IWA: 37.3% CrI: [22.9, 50.4] and −721 ms 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 −66 ms CrI: [-118, −14]; IWA: 14.3% CrI: [-9.1, 35.6] and −113 ms 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 −49 ms CrI: [-107, 11]; IWA: 5% CrI: [-9.1, 19.9] and −388 ms 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 −81 ms CrI: [-120, −43]; IWA: 10.2% CrI: [0.7, 20.9] and −300 ms 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: 89 ms CrI: [34, 144], −2.9% CrI: [-5.8, −0.7], relative clauses: 25 ms CrI: [-7, 58], −0.7% CrI: [-1.5, 0.2], pronouns: 47 ms CrI: [-10, 104], −2.3% CrI: [-8.5, 1.4], PRO: 8 ms 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

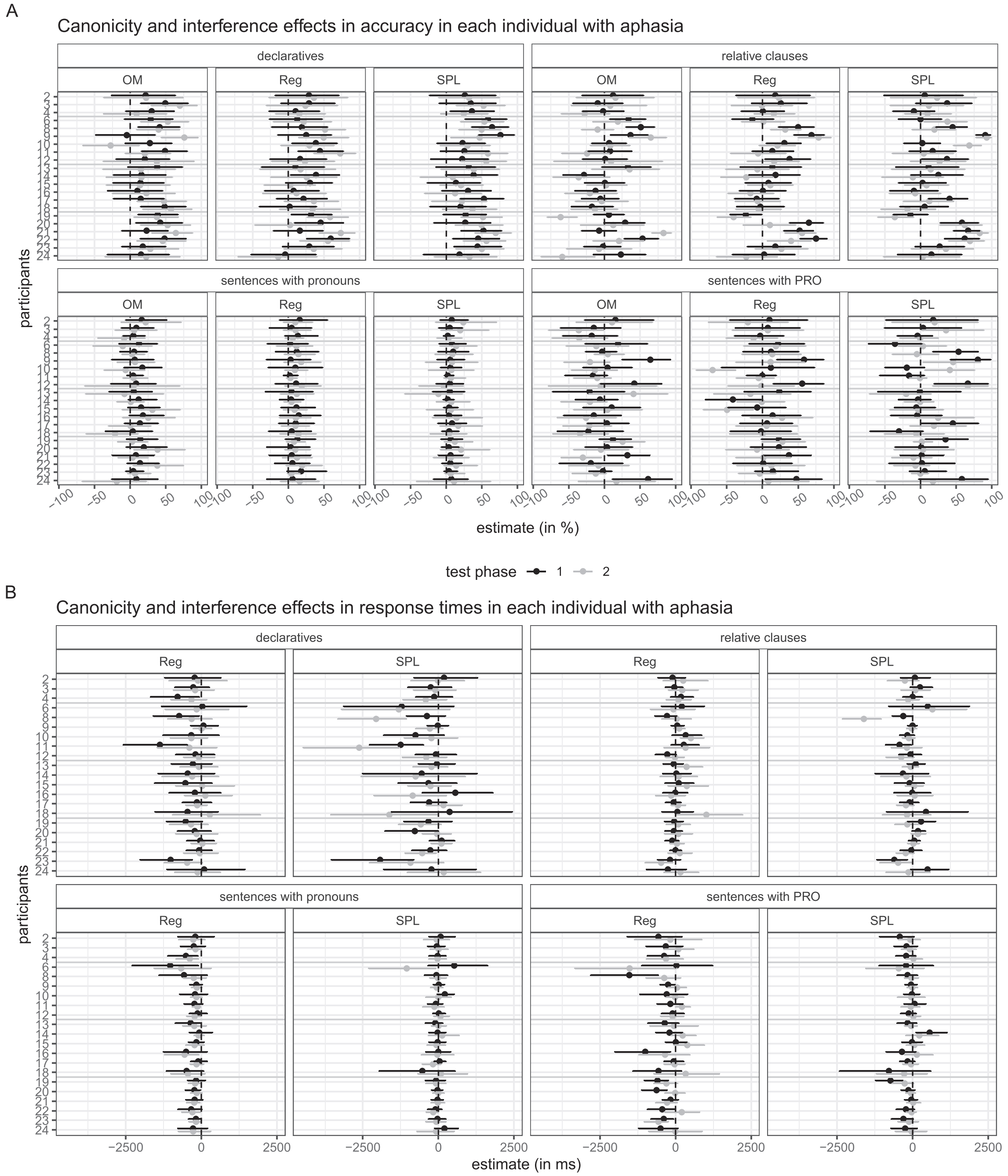

Fig. 5. Canonicity effects in declaratives and relative clauses and interference effects in sentences with pronouns or PRO in accuracy (A) and response times (B) of each individual with aphasia. Each participant completed three response tasks, object manipulation (OM), regular (Reg) and self-paced (SPL) 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).

retest phase (declaratives: −33 ms CrI: [-176, 108], 3.5% CrI: [0, 7.3], relative clauses: −8 ms CrI: [-87, 71], 0.1% CrI: [-2, 2.2], pronouns: −108 ms CrI: [-248, 27], 0.3% CrI: [-3.4, 4], PRO: 163 ms 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: −193 ms CrI: [-256, −132], 0.6% CrI: [0.2, 1.1], relative clauses: −50 ms CrI: [-84, −16], 0.2% CrI: [0.1, 0.5], pronouns: 19 ms CrI: [-34, 74], −0.1% CrI: [-0.8,

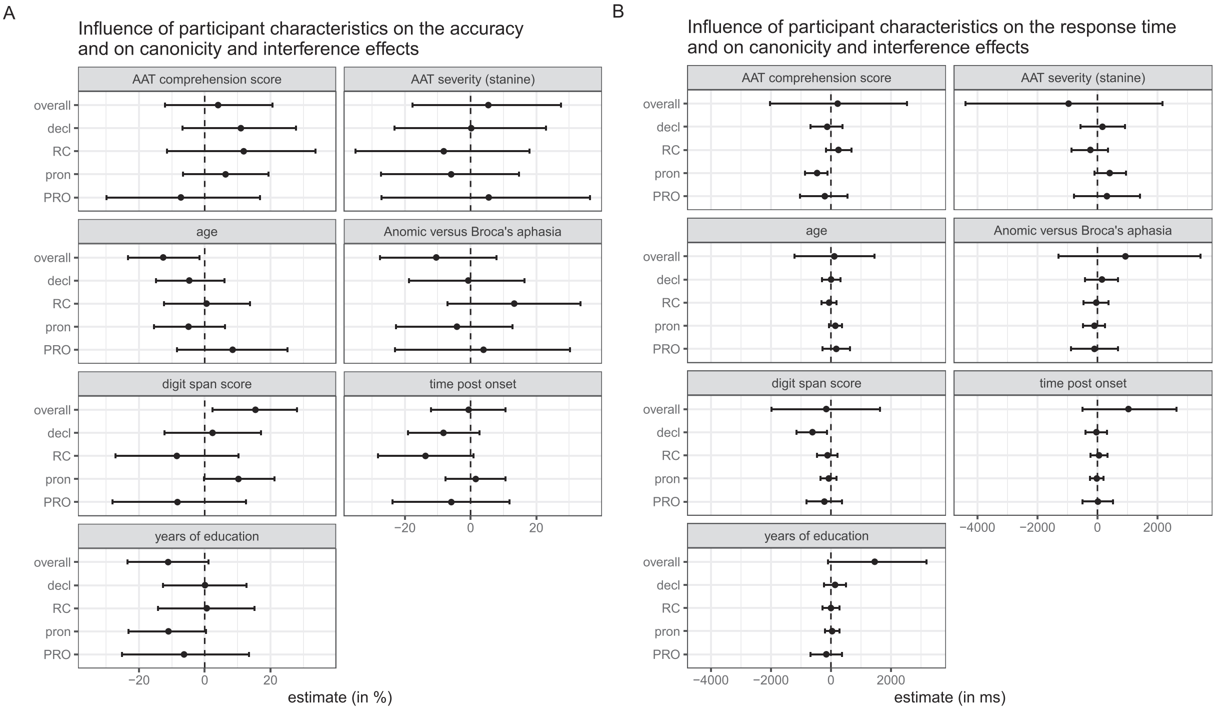

Fig. 6. Mean estimates (dots) and 95% credible intervals (solid lines) of the interaction of different participant characteristics with the overall accuracy (A) and response times (B) and with canonicity effects in declarative sentences (decl) and relative clauses (RC) and interference effects in sentences 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.

0.4], PRO: −7 ms 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: 52 ms CrI: [-35, 140], −3.5% CrI: [-8.2, 0.9], relative clauses: 41 ms CrI: [-10, 92], −0.3% CrI: [-2.7, 2], pronouns: −131 ms CrI: [-223, −41], −1.1% CrI: [-5.6, 3.2], PRO: −55 ms 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. Fig. 4 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 Fig. 4A), 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 Fig. 4B) 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 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 Fig. 4C), 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 Fig. 4D) displayed correlation estimates that were uninformative in all sentence types.

To sum up, only the correlation estimates of 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.

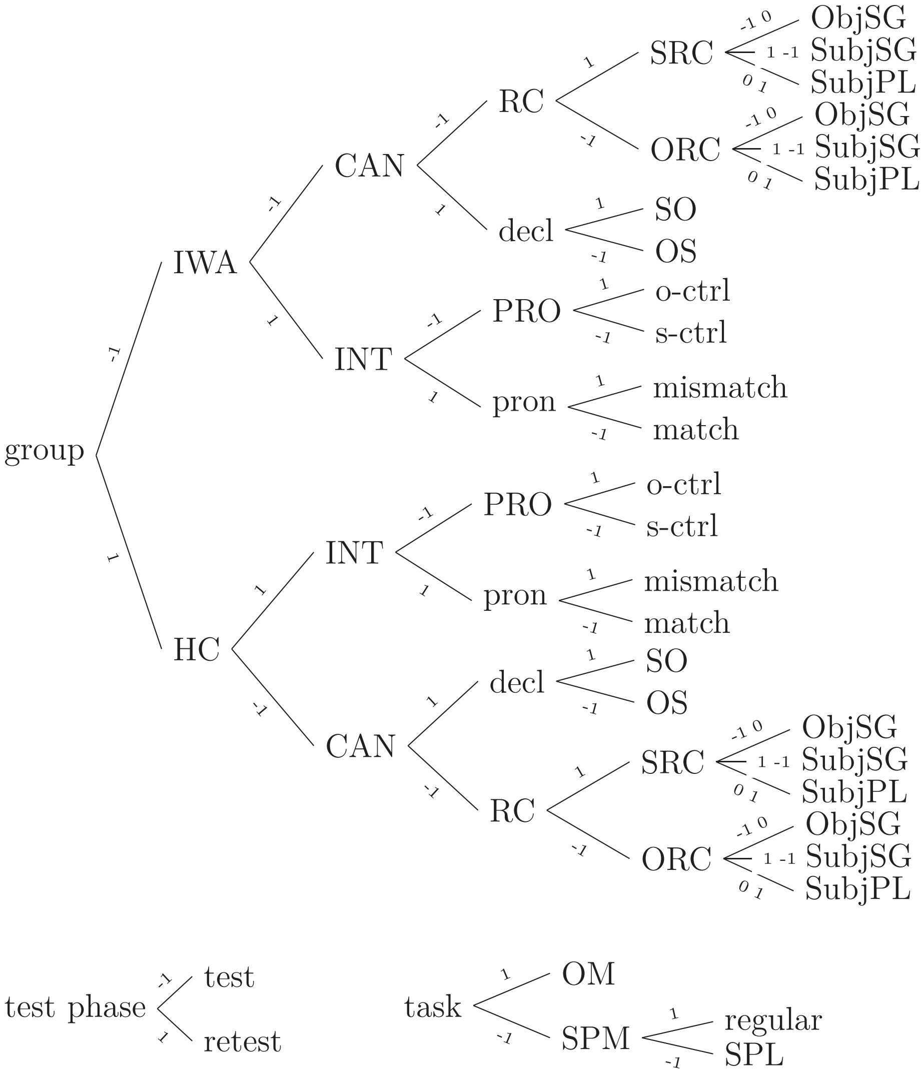

Fig. A1. Visualization of the nestings and contrast codings of the model. CP = control particiapnts, IWA = individuals with aphasia, INT = interference, CAN = canonicity, pron = pronoun, decl = declarative, RC = relative clause, SRC/ORC = subject/object relative, SO/OS = subject-before-object/object-before-subject, 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.

Table A1
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	mismatch	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.

These are represented in Table A2 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. Fig. 5 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

Table A2
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 individuals with aphasia								
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.

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.

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). Fig. 6 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 (-461 ms CrI: [-862, −115]) and canonicity effects in declarative sentences decreased with higher digit span scores (-615 ms 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.

4. Discussion

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 presentation 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; Caplan et al., 2007; Caplan et al., 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 Fig. 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, DeDe, Waters, Michaud, & Tripodis, 2011; Caplan, Michaud, & Hufford, 2013b). Similarly, canonicity 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 extra-linguistic 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 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

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.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We would like to thank all participants who volunteered to contribute to this study as well as three anonymous reviewers for their

[7] An anonymous reviewer pointed out to us that the fact that we did not apply the test battery *Sätze Verstehen* (Burchert, Lorenz, Schröder, De Bleser, & Stadie, 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.

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).

Appendix A

A.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.

A.2. Contrast coding

Fig. A1

A.3. Descriptive statistics

Table A1

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

Table A2

References

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.

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.

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using Ime4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01.

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/.

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal, 10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017.

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*(2), 439.

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*(2), 77–109.

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

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

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

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.

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.

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.

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.

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.

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

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.

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

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

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

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*, 230–234.

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

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.

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

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.

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

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

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.

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.

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

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*.

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.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 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.

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.

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, … Vasishth, S. (2021). A computational evaluation of two models of retrieval processes in sentence processing in aphasia. *Cognitive Science, 45*, Article e12956.

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*(1), 161–174.

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., & 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. *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., 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.

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., Bedirian, 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.

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

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, I. L., Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2016). A computational evaluation of sentence processing deficits in aphasia. *Cognitive Science, 40*(1), 5–50.

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

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

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. *Ada Psychologica, 137*, 172–180.

Schlesewsky, M., Bornkessel, I., & Frisch, S. (2003). The neurophysiological basis of word order variations in German. *Brain and Language, 86*, 116–128.

Shammi, P., Bosnian, E., & Stuss, D. T. (1998). Aging and variability in performance. *Aging, Neuropsychology, and Cognition, 5*, 1–13.

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.

Stiebels, B., McFadden, T., Schwabe, K., Solstad, T., Kellner, E., Sommer, L., & Stoltmann, K. (2018). *ZAS database of clause-embedding predicates*. Mannheim: Institut fur Deutsche Sprache.

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.

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.

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.

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. In *Proceedings of the 41th Annual Conference of the Cognitive Science Society* (pp. 3036–3042).

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.

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.

Yarbay Duman, T., Altınok, N., Özgirgin, N., & Bastiaanse, R. (2011). Sentence comprehension in Turkish Broca's aphasia: An integration problem. *Aphasiology, 25*, 908–926.