



Cognitive Science 45 (2021) e12956


© 2021 The Authors. *Cognitive Science* published by Wiley Periodicals LLC on behalf of Cognitive Science Society (CSS).

All rights reserved.

ISSN: 1551-6709 online

DOI: 10.1111/cogs.12956

## A Computational Evaluation of Two Models of Retrieval Processes in Sentence Processing in Aphasia

Paula Lissón,<sup>a</sup>  Dorothea Pregla,<sup>a</sup> Bruno Nicenboim,<sup>a,b</sup> Dario Paape,<sup>a</sup> Mick L. vanhet Nederend,<sup>c</sup> Frank Burchert,<sup>a</sup> Nicole Stadie,<sup>a</sup> David Caplan,<sup>d</sup> Shravan Vasishth<sup>a</sup>

<sup>a</sup>*Department of Linguistics, University of Potsdam*

<sup>b</sup>*Department of Cognitive Science and Artificial Intelligence, Tilburg University*

<sup>c</sup>*Department of Artificial Intelligence, University of Utrecht*

<sup>d</sup>*Neurology, Massachusetts General Hospital*

Received 22 May 2020; received in revised form 18 January 2021; accepted 24 January 2021

---

### Abstract

Can sentence comprehension impairments in aphasia be explained by difficulties arising from dependency completion processes in parsing? Two distinct models of dependency completion difficulty are investigated, the Lewis and Vasishth (2005) activation-based model and the direct-access model (DA; McElree, 2000). These models' predictive performance is compared using data from individuals with aphasia (IWAs) and control participants. The data are from a self-paced listening task involving subject and object relative clauses. The relative predictive performance of the models is evaluated using k-fold cross-validation. For both IWAs and controls, the activation-based model furnishes a somewhat better quantitative fit to the data than the DA. Model comparisons using Bayes factors show that, assuming an activation-based model, intermittent deficiencies may be the best explanation for the cause of impairments in IWAs, although slowed syntax and lexical delayed access may also play a role. This is the first computational evaluation of different models of dependency completion using data from impaired and unimpaired individuals. This evaluation develops a systematic approach that can be used to quantitatively compare the predictions of competing models of language processing.

**Keywords:** Aphasia; Cue-based retrieval; Sentence processing; Bayesian cognitive modeling; k-fold cross-validation

---

Correspondence should be sent to Paula Lissón, Human Science Faculty, Department of Linguistics, University of Potsdam, 14476 Potsdam, Germany. E-mail: paula.lisson@uni-potsdam.de

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

## 1. Introduction

Understanding a sentence requires the comprehender to access lexical representations of words from memory, link these with upcoming words, build up the structure of the sentence, and compute the meaning of the sentence. Consider example (1):

- (1) The lawyer who came to the office yesterday was looking for some documents.

To understand this sentence, the hearer needs to work out who came to the office, when, and why that happened, and who was looking for what. In the sentence processing literature, the process of linking up words that are linguistically related is known as dependency completion. In sentence (1), an example of a dependency is the one between *lawyer* and *was looking*. A widely held assumption in sentence processing research (Gibson, 2000; Just & Carpenter, 1992; Lewis, Vasishth, & Van Dyke, 2006) is that dependencies require access to the working memory system in order to work out the relationships between words.

Several theories have been developed that spell out the mechanisms that may be involved in the resolution of long-distance dependencies (Gibson, 2000; Just & Carpenter, 1992; McElree, 2000; McElree, Foraker, & Dyer, 2003; Van Dyke & Lewis, 2003). Among these, one class of accounts is referred to as *cue-based retrieval* theory (Engelmann, Jäger, & Vasishth, 2019; Lewis et al., 2006; Vasishth, Nicenboim, Engelmann, & Burchert, 2019). One core assumption here is that words and phrases are stored in memory as a bundle of feature–value pairs. For example, the word *lawyer* is represented in memory as an attribute–value matrix (Pollard & Sag, 1994). Some of the relevant feature–value pairs are shown below:

$$\begin{pmatrix} \text{innominal} - \text{yes} \\ \text{animate} - \text{yes} \\ \text{subject} - \text{yes} \\ \text{singular} - \text{yes} \end{pmatrix}$$

Cue-based retrieval theory assumes that dependencies are resolved via a content-addressable search in memory. For example, in sentence (1), to resolve the dependency between *lawyer* and *was looking*, the comprehender needs to retrieve a mental representation of the noun *lawyer* from memory. Upon encountering the words *was looking*, a retrieval is assumed to be triggered that seeks out a subject noun that has specific features such as [nominal: yes, animate: yes, subject: yes, singular: yes]. We will therefore refer to *was looking* as the retrieval site, that is, the point at which the retrieval of the co-dependent is triggered, and to *the lawyer* as the target of the retrieval. The features that are used to carry out the search and retrieval of a given co-dependent in memory are called retrieval cues. Notice that in this case, it is assumed

that *was looking* is a multi-word unit encoded in memory with a single matrix of feature–value pairs.

One reason that sentence comprehension difficulty arises is when multiple items in memory match the retrieval cues set by the trigger at the retrieval site. This is known as similarity-based interference (Van Dyke & Lewis, 2003; Van Dyke & McElree, 2006). To illustrate, we consider the object relative (OR) clause shown in (2b) along with the baseline condition, the subject relative (SR) clause (2a):

- (2) a. The man who scratched the boy pushed the girl.  
 b. The man who the boy scratched pushed the girl.

In (2b), the comprehender needs to work out who scratched whom, and who pushed whom. When the verb *scratched* is encountered, the retrieval of its corresponding subject (*boy*) is triggered, using retrieval cues such as [nominal: yes, animate: yes, subject: yes, singular: yes]. At the moment of retrieval, there are two nouns available in memory that match these retrieval cues: *man* and *boy*. However, *man* is not the subject of the relative clause (RC) verb *scratched*. Following the literature (Jäger, Engelmann, & Vasishth, 2017), we will refer to *man* as the distractor. Cue-based retrieval theory predicts that processing difficulty increases when both a target and a distractor have features that match the retrieval cues. Processing difficulty arises because these nouns become difficult to distinguish from each other; this phenomenon is called the fan effect in memory research in cognitive psychology (Anderson et al., 2004).

In summary, in (2b) processing is assumed to be more difficult at the verb *scratched* compared to the baseline condition (2a), where the cues used to access the subject of *scratched* only match the subject *man* (Lewis & Vasishth, 2005). Grodner and Gibson (2005) present data from a self-paced reading study on ORs and SRs in English that is consistent with this prediction. The increased processing difficulty at the verb *scratched* is predicted to lead to longer reading times in (2b) versus (2a), and to occasional misretrievals of the incorrect noun from memory. This is the signature effect that is referred to as similarity-based interference (Gordon, Hendrick, Johnson, & Lee, 2006; Jäger et al., 2017; Jäger, Mertzen, Van Dyke, & Vasishth, 2020; Van Dyke & McElree, 2006, 2011; Vasishth et al., 2019).

Two distinct instantiations of cue-based retrieval theory are the Lewis and Vasishth (2005) model of sentence processing (henceforth, LV05) and the direct-access model (henceforth, DA) developed by McElree (2000). The two models share the assumption that retrieval is driven by a cue-based mechanism, and both predict that a distractor disrupts the retrieval of the target when the retrieval cues match the distractor and the target. Despite these similarities, the two models assume fundamentally different underlying processes for the access of representations in memory. In the LV05 model, retrieval time for an item depends on the activation of the item in memory, with reduced discriminability of an item leading to lower activation and therefore longer retrieval times. By contrast, in

the DA model, retrieval time is assumed to be constant, and reduced discriminability only affects the probability of correct retrieval of the target.

Nicenboim and Vasishth (2018) were the first to formally implement these two competing models and compare their relative predictive performance. Using self-paced reading data from a number interference experiment in German (Nicenboim, Vasishth, Engelmann, & Suckow, 2018), Nicenboim and Vasishth implemented the LV05 and DA models in a Bayesian framework. They showed that (a) the DA has better predictive performance than the activation-based model, but (b) the activation-based model yields a comparable performance to the DA when the variance of the retrieval times is allowed to be different for correct and incorrect retrievals. The computational implementations of the two competing models of retrieval make it possible, for the first time, to investigate their relative performance using a broader range of experimental data.

Both LV05 and DA are meant to account for retrieval processes in sentence comprehension in unimpaired populations. An open question is whether these models, which have until now only been investigated in connection with unimpaired processing, can also characterize retrieval difficulty in impaired populations. That is, can the models account for impaired processing through parametric variation? And if they can, what do the changes in the parameters tell us about the impairments? In this paper, we focus on an important and under-studied problem, the underlying nature of retrieval difficulty in individuals with aphasia (IWAs).

Aphasia is an acquired neurological condition caused by brain injury that affects language production and comprehension. One question we seek to answer is: Given the two competing models of retrieval processes, which one better characterizes processing difficulty in IWAs? As data, we use the largest dataset currently in existence on sentence comprehension in IWAs. This dataset, reported in Caplan, Michaud, and Hufford (2015), provides listening times (LTs) and picture-selection accuracies from IWAs and matched unimpaired controls. The full dataset involves a range of syntactic constructions and methods, but in this paper, we focus on self-paced listening data on the SR versus OR clause construction, which is a very well-studied construction in psycholinguistics.

The present paper is structured as follows. We begin by reviewing prior work on modeling retrieval processes in aphasia. Next, we present the data, our implementation of LV05 and DA, the results of the model comparisons, and a Bayes factors (BFs) analysis.

### 1.1. Modeling retrieval processes in aphasia

There are several theories about why language processing deficits arise in IWAs. In this paper we focus on processing deficit theories that can be implemented within the framework of cue-based theory and that are of relevance for our modeling work.<sup>1</sup> In particular, we focus on the following accounts: *delayed lexical access* (Ferrill, Love, Walenski, & Shapiro, 2012), *slow syntax* (Burkhardt, Avrutin, Piñango, & Ruigendijk, 2008), *resource reduction* (Caplan, 2012), and *intermittent deficiencies* (Caplan et al., 2015).

The *delayed lexical access* theory claims that lexical access is delayed in IWAs, and this can cause a slowdown in the formation of a syntactic dependency. Evidence

supporting this theory comes from a series of cross-modal lexical priming studies, which combine a listening comprehension and a lexical decision task. Love, Swinney, Walenski, and Zurif (2008) and Ferrill et al. (2012) (inter alia) found that IWAs showed slower lexical activation relative to controls. Some cross-modal lexical priming studies have also revealed that IWAs build syntactic dependencies at a slower-than-normal speed. This has been taken as support for the *slow syntax* theory (Burkhardt et al., 2008; Burkhardt, Piñango, & Wong, 2003), which posits that a slowdown in syntactic structure building can cause a delayed interpretation or a failure to interpret the sentence. Under this account, the impairment is at the level of syntactic structure formation.

Caplan, Waters, DeDe, Michaud, and Reddy (2007) and Caplan et al. (2015) present online and offline data that support the hypothesis that IWAs have a deficit in the resources used in parsing, what they refer to as *resource reduction* (Caplan, 2012). Complex sentences demand more resources, such as a higher memory load or attention, and therefore, IWAs are more likely to misinterpret complex sentences. Finally, Caplan, Michaud, and Hufford (2013) argue that in addition to a *resource reduction*, IWAs may exhibit intermittent breakdowns in the parsing system, a theory known as *intermittent deficiencies*.

Some of these accounts have been implemented in the framework of LV05. Patil, Hanne, Burchert, De Bleser, and Vasishth (2016) developed several LV05-based models that implement theories of processing deficits in aphasia. They found that IWAs' processing was better characterized by a model that combined the implementation of slowed processing (understood as a "pathological slowdown in the processing system") and intermittent deficiencies, relative to models that included only one of these deficits. Building on the conclusions of Patil et al. (2016), Mätzig, Vasishth, Engelmann, Caplan, and Burchert (2018) investigated variability among IWAs by implementing slowed processing, intermittent deficiencies, and resource reduction within the LV05 model. The range of parameters estimated for IWAs showed a broad variability, whereas the parameters for control participants were closer to the default parameters of the original LV05 model and displayed a smaller range of variability. These results imply that IWAs are very variable in the extent and nature of their deficits along these three hypothesized dimensions (slowed processing, intermittent deficiencies, and resource reduction). The broader conclusion here is that deficits may lie on a continuum, and along different dimensions.

Although Patil et al. (2016) only modeled data from seven IWAs, and Mätzig et al. (2018) modeled offline measures (accuracies), both studies showed that LV05 can account for IWAs' behavior by modifying specific parameters that can be mapped onto theoretically informed assumptions. By doing so, they derived quantitative predictions under the assumptions of theories of deficits in aphasia. However, whether the LV05 model can account for the different hypothesized deficits in a larger dataset with online measures remains to be tested.

As discussed in the previous section, there exists another competing model of retrieval processes, the DA. The crucial difference between these two models is that they assume different underlying mechanisms for the access of items in memory. Yet the relative predictive performance of the activation model and of the DA has never been compared

using data from both unimpaired and impaired populations. By comparing these two models' predictions with data from IWAs, we aim to investigate the following questions: (a) Can the direct-access mechanism of retrieval also account for sentence processing in IWAs? (b) How do the different parameters of these two models relate to theories of processing deficits in IWAs? (c) Which model provides a better fit to data from IWAs and controls? Investigating these questions would provide new insight into the nature of the dependency completion process in impaired and unimpaired populations.

The Caplan et al. dataset makes such a model comparison possible. Below, we begin by revisiting the characteristics of the subset of the Caplan et al. dataset that we use in this paper.

## 2. The Caplan et al. dataset: Self-paced listening times in relative clauses

The empirical data we consider here consist of LTs and picture-selection accuracies from 33 IWAs and 46 controls matched by age and years of education. The original dataset reported in Caplan et al. (2015) included 56 IWAs, but we discarded data from eight IWAs because they were in the early post-acute phase (less than 4 months post-stroke), and from 15 other individuals who had been classified as IWAs but showed no symptoms of aphasia in the Boston Diagnostic Aphasia Exam (Goodglass, Kaplan, & Barresi, 2001).

Out of the 11 sentence types in the dataset, we selected the SR and OR constructions (see examples 3a and 3b). This choice was motivated by the fact that RCs have been extensively studied in psycholinguistics, and a great deal is known about RC processing. In English and many other languages, ORs have been uniformly found to be more difficult to process than SRs (Grodner & Gibson, 2005). Moreover, IWAs are known to experience difficulties in the comprehension of OR clauses (Caramazza & Zurif, 1976; Hanne, Sekerina, Vasisht, Burchert, & De Bleser, 2011), especially when the thematic roles of the nouns can be reversed, as in the sentences shown below.

- (3) a. **Subject Relative (SR):** The girl who chased the mother hugged the boy.  
 b. **Object Relative (OR):** The girl who the mother chased hugged the boy.

In the experiment reported by Caplan et al. (2015), participants listened to sentences presented word by word, and pressed a computer key whenever they were ready to hear the next word. This yielded an online measure of comprehension: LTs per segment, in milliseconds. At the end of the sentence, participants had to choose which of two pictures displayed on the screen matched the meaning of the sentence they had just heard. This choice yielded accuracy data (correct/incorrect response). An example of the pictures shown in the picture-selection task is displayed in Fig. 1. These pictures correspond to the sentences (3a) and (3b).

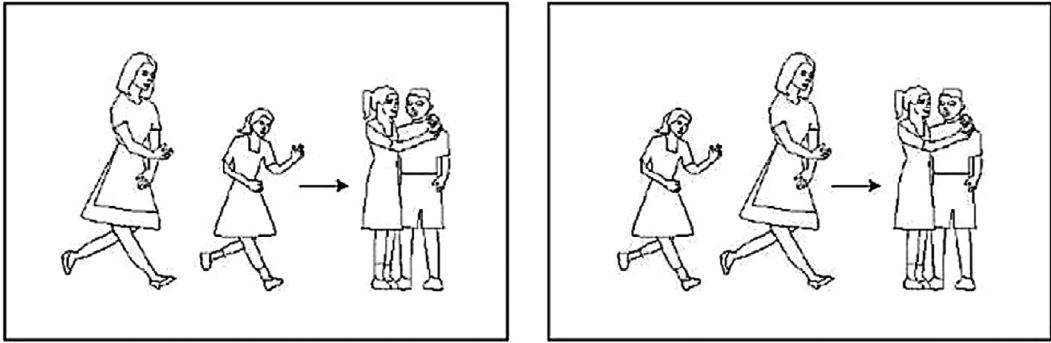


Fig. 1. Example of the images shown in the picture-selection task. In the subject relative condition, the picture on the right is the target, whereas the picture on the left is the foil. In the object relative condition, the picture on the left is the target, and the one on the right is the foil.

Of the 20 items corresponding to the SR and OR conditions in Caplan et al. (2015), we only used items 11–20 for our data analysis and modeling. The modeling is limited to these items because it was only in these items that the pictures in the picture-selection task tested the participant’s understanding of the meaning of the verb inside the RC (e.g., who chased whom in 3a and 3b). For cue-based retrieval theory, in RCs, the retrieval of the agent of the action expressed by the verb within the RC is the first and key retrieval event (Lewis & Vasishth, 2005).

In English, the verb of the subordinate clause (chased in 3a and 3b) does not appear in the same position in SR and OR clauses, and therefore the LTs corresponding to the verb region are not directly comparable. To make the two sentences comparable, we followed the procedure in Traxler, Williams, Blozis, and Morris (2005) and added up the LTs of the noun phrase (“the mother”) and the verb (“chased”) inside the SR/OR clause. Trials with LTs shorter than 200 ms were discarded (around 2% of the data).

In the following section we present descriptive statistics and a Bayesian analysis of the data used for modeling. We analyze the data using the Bayesian framework because this allows us to quantify uncertainty about the estimates of interest (e.g., the difference in LTs for SR and OR clauses). Our statistical inferences are based on 95% credible intervals and means of the estimates; the credible intervals show the range over which plausible values of the parameter lie with 95% probability, given the data and the model.

### 3. Bayesian analysis of the Caplan et al. (2015) relative clause listening time data

The mean accuracy for controls and IWAs across the two conditions is shown in Fig. 2. For controls, accuracy is above 90% in both conditions, whereas for IWAs accuracy in SRs is 75%, and 63% in ORs. Fig. 3 shows the mean LTs across conditions and groups. IWAs are slower than controls in both conditions. For both IWAs and controls, responses in the OR condition are slower relative to responses in the SR condition.

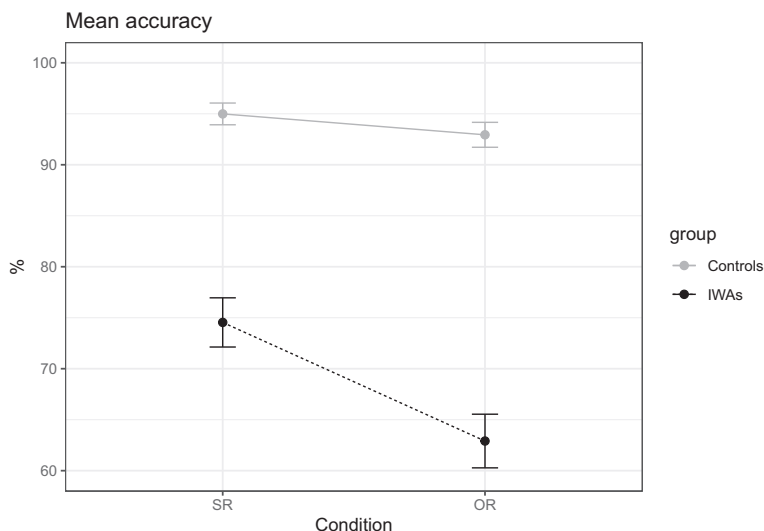


Fig. 2. Mean accuracy across conditions and groups. Error bars show the standard error of the mean.

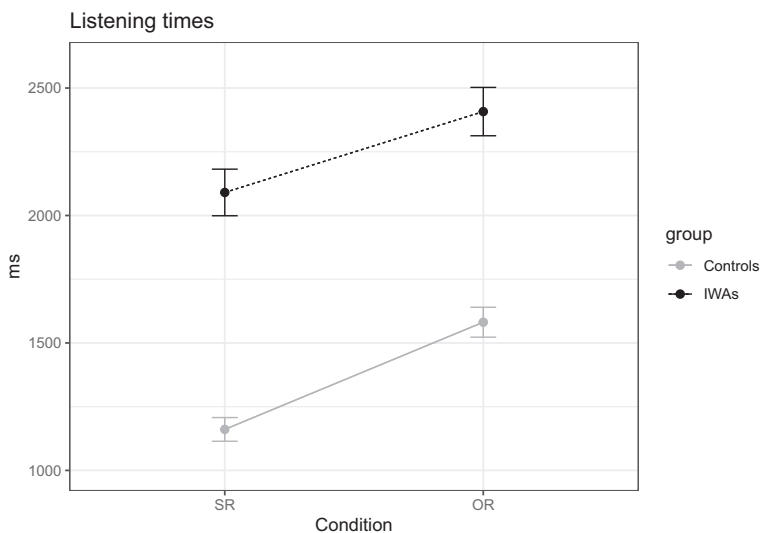


Fig. 3. Mean listening times across conditions and groups. The listening times correspond to the sum of the listening times for the verb and noun phrase of the relative clause. Error bars show the standard error of the mean.

We fit a Bayesian hierarchical model with a lognormal likelihood to the LTs and a Bayesian logistic mixed model to the accuracy data. The analyses were carried out with correct and incorrect trials pooled. We used R (R Core Team, 2020) and the package *brms* (Bürkner, 2017), which is a front-end for Stan (Carpenter et al., 2017). For both models, the factors *group* (controls/IWAs), *condition* (SR/OR), and their interaction were



fit as fixed effects. These factors were sum-coded (Schad, Vasishth, Vasishth, Hohenstein, & Kliegl, 2020): SR were coded as  $-1$  and OR as  $+1$ ; controls as  $-1$  and IWAs as  $+1$ . Random intercepts by subjects and items were included, a slope by item was added to the group effect, and a slope by subject was added to the effect of condition. The varying intercepts and slopes were allowed to be correlated.

We used so-called regularizing priors, which allow a broad range of parameter values but disallow implausible (or impossible) values. The priors for the model of the accuracies, listed in Eq. 1, are on the logit scale, whereas the priors for the LTs model, listed in Eq. 2, are on the log scale. In the prior specification for the residual standard deviation ( $\sigma$ ), the subscript  $+$  in the normal distribution prior stands for a normal distribution truncated at 0 (reflecting the fact that standard deviations can never be less than 0). For the correlation matrix of the random effects, we used the so-called LKJ prior (Lewandowski, Kurowicka, & Joe, 2009) with parameter 2; this parameter disfavors extreme correlations like  $\pm 1$  (Carpenter et al., 2017). The models were fit with four chains and 2,000 iterations, of which 1,000 were warm-up iterations.

$$\begin{aligned}\alpha &\sim \text{normal}(0, 1) \\ \beta_{1,\dots,3} &\sim \text{normal}(0, 0.5) \\ \sigma &\sim \text{normal}_+(0, 0.5)\end{aligned}\quad (1)$$

$$\begin{aligned}\alpha &\sim \text{normal}(7.5, 0.6) \\ \beta_{1,\dots,3} &\sim \text{normal}(0, 0.5) \\ \sigma &\sim \text{normal}_+(0, 0.5)\end{aligned}\quad (2)$$

Fig. 4 shows the posterior distributions of the parameters of interest. In a Bayesian model, the posterior distribution indicates the most likely parameter values given the data and the model. We report the mean estimate for each effect of interest, as well as their corresponding 95% credible interval (CrI). This interval represents the range over which we are 95% certain that the effect lies, given the data and the model.

Fig. 4A shows the posterior distributions of the fixed effects for the analysis of the accuracy data. The data show an effect of group and condition: The estimated effect for group is of  $-24\%$  CrI:  $[-29, -18]$ , indicating that IWAs have more incorrect responses than controls. The effect of condition,  $-5\%$  CrI:  $[-9, -2]$  suggests that more incorrect responses are given in the OR condition. No indication for an interaction is seen,  $-1\%$  CrI:  $[-5, 3]$ .

In LTs, large effects for group and condition were found: ORs yield longer LTs (effect of condition: 323 ms CrI:  $[227, 422]$ ), and IWAs are slower than controls (effect of group: 647 ms CrI:  $[309, 1003]$ ). The interaction ( $-85$  ms CrI:  $[-182, 9]$ ) suggests that the effect of condition could be stronger for controls, but since the CrI overlaps with 0, strong conclusions cannot be drawn from this estimate.

Having summarized the inferences that can be made from the data, we now turn to a description of the two models, and the models' evaluation and comparisons.

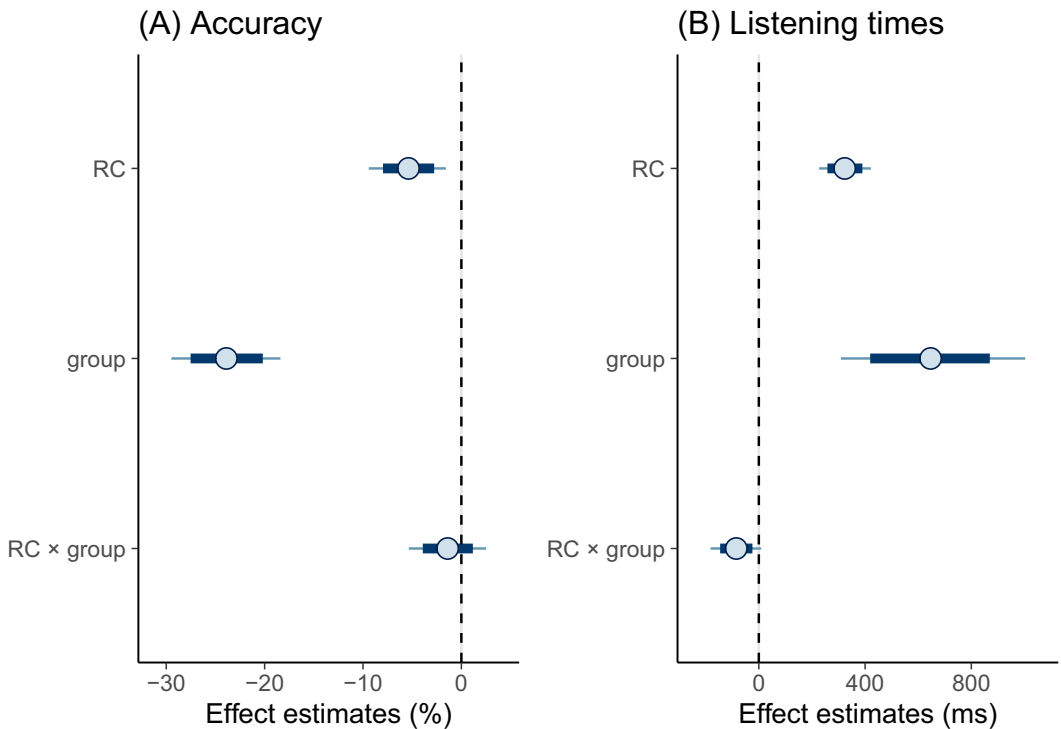


Fig. 4. Posterior probability distributions of the different effect sizes for the effect of group (controls/TWA), condition (SR/OR), and their interaction. The dot corresponds to the mean of the distribution, the thick lines are 80% credible intervals, and the thin lines show 95% credible intervals. The dashed line stands for an effect size of zero.

#### 4. The activation-based model

In cognitive psychology, response selection in simple choices is often modeled using accumulation of evidence (Heathcote & Love, 2012; Ratcliff, 1978). Evidence accumulation models assume that when facing a speeded decision, people accumulate noisy samples of information about the different choices that are available, until they have enough evidence to choose one of them (Forstmann, Ratcliff, & Wagenmakers, 2016).

Language processing can be seen as a similar process: When listening to a sentence, the comprehender samples evidence from the linguistic input that unfolds over time. Once the retrieval site is encountered, comprehenders have to retrieve an item from memory. Nicenboim and Vasisht (2018) argued that the retrieval process assumed in LV05 is conceptually similar to a race model (Rouder, Province, Morey, Gomez, & Heathcote, 2015; Usher & McClelland, 2001), in which each choice is represented with an accumulator of evidence. The speed of the process of sampling evidence in a race of accumulators can be equated to the activation in LV05: The item in memory with the faster rate of accumulation (equivalent to the higher activation in LV05) will be the item retrieved, and the rate of accumulation will determine the latency of the retrieval.

In the Caplan et al. (2015) data, the LTs at the RC verb and the second noun phrase serve as a measure of the speed of accumulation of evidence for the retrieval. Because there are two possible interpretations (SR or OR clause), we assume that there are two accumulators racing against each other. For instance, consider again the OR clause (3b), repeated here for convenience as (4):

(4) The girl who the mother chased hugged the boy.

When the comprehender reaches the verb *chased*, they need to retrieve a subject that matches the verb. If the comprehender understands the sentence correctly, they should have retrieved *mother* as the subject of the verb. An alternative possibility is that they accidentally misretrieve *girl* as the subject of the verb. Under these assumptions, the model has two accumulators: One accumulates evidence for the retrieval of the target (which corresponds to the correct OR interpretation in this example), and the other one accumulates evidence for the retrieval of the distractor *girl* (which corresponds to the incorrect SR interpretation in this example). The accumulator that finishes faster represents the interpretation chosen. We also assume that, when selecting one of the pictures during the picture-selection task, participants are choosing the interpretation that corresponds to the chunk retrieved from memory at *chased* (i.e., *mother* or *girl* in 4).

#### 4.1. Implementation of the activation-based model

Following Nicenboim and Vasishth (2018), the activation-based model is implemented as a Bayesian lognormal race of accumulators. The Bayesian framework was chosen for two reasons. First, because modern probabilistic programming languages like Stan (Carpenter et al., 2017) make it possible to flexibly define any assumed generative process while including taking individual differences into account. Second, the Bayesian approach to parameter estimation allows the researcher to directly take the uncertainty of the estimates into account (Lee & Wagenmakers, 2014).

The model was implemented in Stan. For each trial  $i$ , the finishing times  $FT$  for the interpretation of a sentence as SR or OR are sampled from two lognormal distributions with scale  $\sigma$ , see Eq. 3.<sup>2</sup> The noise component ( $\sigma$ ) is assumed to be different for controls and IWAs.<sup>3</sup> The accumulator with the faster (i.e., lower)  $FT$  will represent the winning interpretation, and its sampled value will become the estimated LT for that particular trial  $i$ , as shown in Eq. 4.

*SR accumulator*

$$FT_{SR_i} \sim \text{lognormal}(\mu_{SR}, \sigma) \quad (3)$$

*OR accumulator*

$$FT_{OR_i} \sim \text{lognormal}(\mu_{OR}, \sigma)$$

$$LT_i = \min(FT_{SR_i}, FT_{OR_i}) \quad (4)$$

The complete hierarchical model for the two accumulators is presented in Eq. 5. The terms  $u$  and  $w$  are the by-participant and by-item adjustments to the fixed effects terms; these are the familiar varying intercepts and slopes in linear mixed models (Bates, Maechler, Bolker, & Walker, 2015). All the parameters (which, given the lognormal likelihood, are on the log scale) have regularizing priors, listed in Eq. 6.<sup>4</sup> In the specific context of psycholinguistics, prior specification in hierarchical models is discussed at length in Sorensen, Hohenstein, and Vasishth (2016), Nicenboim and Vasishth (2016), Vasishth, Nicenboim, Beckman, Li, and Kong (2018), and Schad, Betancourt, Betancourt, and Vasishth (2020). The level labeled *group* had contrast coding  $-1$  for controls, and  $+1$  for IWAs; and the level labeled *relative clause type* ( $rc_{type}$ ) was coded such that SRs were represented as  $-1$  and ORs as  $+1$ .

*SR accumulator*

$$\mu_{SR} = \alpha_1 + u_{\alpha_1} + w_{\alpha_1} + (\beta_1 + w_{\beta_1}) \times group \\ + (\beta_3 + u_{\beta_3}) \times rc_{type} + \beta_5 \times group \times rc_{type}$$

*OR accumulator*

$$\mu_{OR} = \alpha_2 + u_{\alpha_2} + w_{\alpha_2} + (\beta_2 + w_{\beta_2}) \times group \\ + (\beta_4 + u_{\beta_4}) \times rc_{type} + \beta_6 \times group \times rc_{type} \quad (5)$$

*Noise parameter*

$$\sigma = \sigma_0 + \beta_7 \times group$$

$$\alpha_{1,2} \sim normal(7.5, 0.6) \\ \beta_{1,\dots,7} \sim normal(0, 0.5) \\ \sigma_0 \sim normal_+(0, 0.5) \quad (6)$$

The varying intercepts and slopes for subject,  $\mathbf{u} = \langle u_{\alpha_1}, u_{\alpha_2}, u_{\beta_3}, u_{\beta_4} \rangle$ , come from a multivariate normal distribution with four dimensions, abbreviated as  $MVN_4$ ; and the varying intercepts and slopes for items,  $\mathbf{w} = \langle w_{\alpha_1}, w_{\alpha_2}, w_{\beta_1}, w_{\beta_2} \rangle$ , also come from a multivariate normal distribution with four dimensions,  $MVN_4$ . In the equations below,  $\mathbf{0}$  is a column vector of zeros with the four (participants) or four (items) dimensions. The  $\Sigma$  are the variance–covariance matrices of the multivariate normal distributions.

$$\mathbf{u} = MVN_4(\mathbf{0}, \Sigma_u) \quad (7)$$

$$\mathbf{w} = MVN_4(\mathbf{0}, \Sigma_w) \quad (8)$$

The fixed effects  $\beta$  have the following interpretations:

- $\beta_1, \beta_3, \beta_5$  are the effects of group, RC type, and the group  $\times$  RC type interaction, respectively, in the accumulator for the SR interpretation.
- $\beta_2, \beta_4, \beta_6$  are the effects of group, RC type, and the group  $\times$  RC type interaction, respectively, in the accumulator for the OR interpretation.
- $\beta_7$  is the effect of group in the  $\sigma$  parameter.

Of interest in this model are the distributions of finishing times in the SR and OR accumulators, in the SR and OR conditions, and in the different population groups (controls vs. IWAs). These are generated in milliseconds once the posterior distributions of all the parameters in the model are estimated. The finishing times for each one of the accumulators in each condition and for each group are estimated taking into account the abovementioned terms  $\beta_{1,\dots,7}$  and the adjustments by item and by participant listed in Eq. 5.

#### 4.2. Predictions

In the activation-based model the parameter  $\sigma$  and the finishing times of the accumulators have a theoretically meaningful interpretation. We expect these parameters to show different patterns across groups. The different  $\sigma$  reflect the assumption that for IWAs, the rate of accumulation of evidence can be noisier. A larger estimated  $\sigma$  for IWAs would be consistent with the *intermittent deficiencies* theory (Caplan et al., 2007), which claims that there are intermittent breakdowns in the parsing system of IWAs. However, the effects of crucial interest are on the finishing times: When the mean finishing time of the incorrect interpretation is similar to the finishing time of the correct interpretation, misretrievals become more likely. We therefore expect that compared to controls, IWAs should have more similar mean finishing times in the two accumulators; controls should have a bigger difference between the mean finishing times of the two accumulators. We also expect both accumulators to be slower for IWAs than for controls because IWAs may need more time than controls to retrieve items from memory and to build the dependency. Such a slowdown could be due to a *lexical access deficit* (Love et al., 2008) and/or to *slow syntax* (Burkhardt et al., 2008).

### 5. The direct-access model

The DA (McElree, 2000) assumes that items (i.e., traces of words or phrases, such as *the girl*) in memory are accessed via a content-based, direct-access mechanism. That is, the cues set at the retrieval site enable direct access to matching items in memory. The retrieval process is subject to interference and decay: Increasing distance between the target and the retrieval site, or competing items in the sentence can lower the quality of the representation of the target item in memory. In the DA, the probability of retaining a memory representation at the retrieval site is known as the *availability* of a given item. Crucially, proponents of the DA argue that interference and decay have an impact on the availability of items in memory, but not on retrieval latencies. That is, whereas the

probability of retrieving an item decreases as a function of the complexity of a sentence, complexity does not affect retrieval times. The DA has been developed and tested within the speed-accuracy tradeoff paradigm (SAT) by McElree and colleagues (Martin & McElree, 2008, 2011; McElree et al., 2003), inter alia. They consistently found that the asymptote of the SAT function (which assesses successful retrieval of the target and/or quality of the retrieved representation) decreased as a function of sentence complexity. By contrast, the intercept and the rate of the SAT function (which assess processing speed) did not show a significant effect of complexity. Based on these findings, McElree and colleagues argue that interference and/or decay affect the probability of retrieving the target, but not the retrieval speed. In addition, it is assumed that low availability can cause a failure in parsing or the retrieval of a distractor item. On some trials, this initial failure could be followed by a reanalysis process (Martin & McElree, 2008; McElree, 1993; McElree et al., 2003; Van Dyke & McElree, 2011).

### 5.1. Implementation of the direct-access model

We follow Nicenboim and Vasishth (2018) by implementing the DA as a two-component Bayesian mixture model. The key assumptions of the DA are thus that retrieval cues enable direct access to the item's memory representation at the retrieval site, and that the retrieval of an item takes an average time  $t_{da}$ . Differences in availability can lead to an initial incorrect retrieval of the distractor item. McElree and colleagues assume that on a certain proportion of trials, after a failure in parsing, comprehenders could engage in a "costly reanalysis process" (Martin & McElree, 2008). We formalize this assumption with two main parameters:  $P_b$ , which is the probability of backtracking (what McElree and colleagues call *reanalysis*), and  $\delta$ , which is the extra time needed for backtracking. This extra time is independent of the retrieval time  $t_{da}$ . Notice that these two parameters ( $P_b$  and  $\delta$ ) are not part of the SAT paradigm, and they constitute an implementation of McElree and colleagues' assumption of reanalysis. The model is shown schematically in Fig. 5.

The parameter  $\theta$  is the probability of correctly retrieving an item on the first retrieval attempt. This probability is allowed to vary across conditions, as it is assumed by McElree et al. (2003) that sentence complexity can have an impact on the availability of the items, and therefore on their retrieval probability. If an initial misretrieval or failure in parsing occurs at the retrieval site, a backtracking process is initiated with probability  $P_b$  that, by assumption, always results in correct retrieval of the target (McElree, 1993).

There are four fixed-effects parameters that have to be estimated in this model. For the parameter  $\theta$  we define varying intercepts by participants and by items, and varying slopes for the effect of RC type (by participants) and group type (by items). The parameter  $\mu$  represents the estimated log mean LTs at the critical region. Since the DA assumes that the retrieval time of an item takes on average  $t_{da}$  log ms and is not affected by sentence complexity, RC type was not included as a fixed effect for the parameter  $\mu$ . However, we assume that IWAs, given their impairment, could have a higher  $\mu$  compared to controls and therefore add a main effect of group. That is, we assume that IWAs may differ in

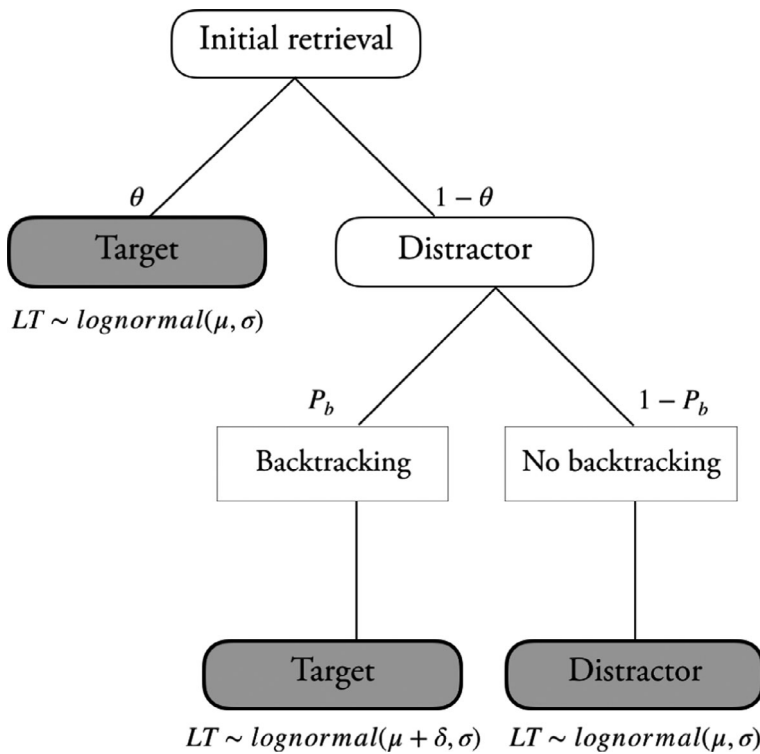


Fig. 5. Graphical representation of how retrieval probabilities work in the direct-access model. An initially wrong retrieval can lead to backtracking (with probability  $P_b$ ), and backtracking leads to the retrieval of the target.

the average time they need to process the critical region relative to controls. Notice that  $t_{da}$  is a latent variable that is part of  $\mu$ , since we cannot directly compute  $t_{da}$  from the observed LTs. The probability of backtracking,  $P_b$  is also not assumed to vary across conditions, and thus only has an adjustment for group and a varying intercept by-participants because we assume that IWAs could have a different  $P_b$  relative to controls. The parameter  $\delta$  is the cost of backtracking, that is, the time (in log ms) that the backtracking process takes, and has an adjustment for group. The standard deviation  $\sigma$  also has a main effect of group. As in the activation-based model, the terms  $u$  and  $w$  are the by-participant and by-item adjustments to the fixed effects terms. As with the activation-based model, all the parameters (which are on the logit scale for probabilities and on the log scale for LTs) have regularizing priors, listed in Eq. 11. The level group had contrast coding  $-1$  for controls, and  $+1$  for IWAs; and  $rc_{type}$  was coded  $-1$  for SR clauses and  $+1$  for ORs. The complete hierarchical model for all the parameters is shown in Eqs. 9 and 10. The mixture process is shown in Eq. 9, and the parameters and their priors are defined in Eq. 10.

$$LT \sim \begin{pmatrix} \text{lognormal}(\mu, \sigma), & \text{retrieval succeeds, probability } \theta \\ \text{lognormal}(\mu + \delta, \sigma), & \text{retrieval fails initially, probability } 1 - \theta \end{pmatrix} \quad (9)$$

$$\begin{aligned} \mu &= \mu_0 + u_{\mu_0} + w_{\mu_0} + \beta_1 \times \text{group} \\ \theta &= \alpha + u_{\alpha} + w_{\alpha} + (\beta_2 + u_{\beta_2}) \times rc_{\text{type}} \\ &\quad + (\beta_3 + w_{\beta_3}) \times \text{group} + \beta_4 \times \text{group} \times rc_{\text{type}} \\ P &= \gamma + u_{\gamma} + \beta_5 \times \text{group} \\ \delta &= \delta_0 + \beta_6 \times \text{group} \\ \sigma &= \sigma_0 + \beta_7 \times \text{group} \end{aligned} \quad (10)$$

$$\begin{aligned} \alpha &\sim \text{normal}(1, 0.5) \\ \beta_{1,\dots,7} &\sim \text{normal}(0, 0.5) \\ \mu_0 &\sim \text{normal}(7.5, 0.6) \\ \gamma &\sim \text{normal}(-1, 0.5) \\ \delta_0 &\sim \text{normal}(0, 0.1) \\ \sigma_0 &\sim \text{normal}(0, 0.5) \end{aligned} \quad (11)$$

In Eq. 10, the varying intercepts and slopes for subject,  $\mathbf{u} = \langle u_{\mu_0}, u_{\alpha}, u_{\beta_2}, u_{\gamma} \rangle$ , come from a multivariate normal distribution with four dimensions, abbreviated as  $MVN_4$ ; and the varying intercepts and slopes for items,  $\mathbf{w} = \langle w_{\mu_0}, w_{\alpha}, w_{\beta_3} \rangle$ , come from an  $MVN_3$  distribution. In the equations below, 0 is a column vector of zeros with the four (participants) or three (items) dimensions.

$$\mathbf{u} = MVN_4(0, \Sigma_u) \quad (12)$$

$$\mathbf{w} = MVN_3(0, \Sigma_w) \quad (13)$$

The fixed effects  $\beta$  have the following interpretations:

- $\beta_1$  is the effect of group on the average time needed to listen to the critical region.
- $\beta_2, \beta_3, \beta_4$  are the effects of RC, group, and the group  $\times$  RC interaction, respectively, on the probability of a first correct retrieval.
- $\beta_5$  and  $\beta_6$  are the effect of group on the probability of backtracking and on the estimated backtracking time, respectively.
- $\beta_7$  is the effect of group on  $\sigma$ .

Consider the three possible scenarios according to the DA, and their corresponding paths shown in Fig. 5.

**Case (i):** The target is retrieved through a direct-access mechanism based on the cues set at the retrieval site, with probability  $\theta$ . In this case, LTs are assumed to be drawn from a lognormal distribution with mean  $\mu$  and standard deviation  $\sigma$ :



$$LT \sim \text{lognormal}(\mu, \sigma).$$

**Case (ii):** The distractor is initially retrieved, but backtracking leads to the target being retrieved, with probability  $(1 - \theta) \times P_b$ . Once  $\theta$  (the probability of initial correct retrieval) has been estimated,  $(1 - \theta)$  yields the probability of an initial incorrect retrieval. The probability of backtracking is assumed to be independent of  $\theta$ . Thus, multiplying  $P_b$  with  $(1 - \theta)$  yields the probability of correctly retrieving the target after an initial misretrieval and subsequent backtracking. In this case, the LTs are drawn from a lognormal distribution with mean  $\mu + \delta$ , which is the cost of backtracking, and standard deviation  $\sigma$ :

**Case (iii):** The distractor is initially retrieved and there is no backtracking, with probability  $(1 - \theta) \times (1 - P_b)$ . In this case, we multiply the probability that the first retrieval is incorrect with the probability that there is no backtracking. Here, the LTs are drawn from a lognormal distribution with mean  $\mu$  and standard deviation  $\sigma$ :  $LT \sim \text{lognormal}(\mu, \sigma)$ , and a misretrieval is predicted.

Notice that incorrect answers without backtracking in case (iii) are expected to have similar LTs to correct answers without backtracking, case (i), whereas in case (ii), longer LTs should be observed due to the extra time needed for backtracking. As such, in this model, the distribution of LTs associated with correct responses is a mixture of initially retrieved targets (i), and initial misretrievals plus backtracking (ii).

## 5.2. Predictions

The parameters  $\theta$ ,  $\mu$ ,  $P_b$ ,  $\delta$ , and  $\sigma$  have a group adjustment because they are expected to differ between controls and IWAs. We present here a short theoretical explanation of the interpretation of these parameters.

We expect a lower estimate of the probability of correct initial retrieval,  $\theta$ , for IWAs, in OR clauses. This would be in line with *resource reduction*. Complex sentences are assumed to require more processing resources, because additional linguistic operations need to be carried out and more material has to be kept in working memory (Caplan, 2012). This suggests that IWAs should show a lower probability of initial correct retrieval in ORs relative to SRs. The different  $\mu$  for controls and IWAs reflect the assumption that IWAs may need more time for parsing. This assumption can be linked to *slowed processing* theories, which would explain the slowdown in terms of lexical access (Love et al., 2008) or syntactic processing (Burkhardt et al., 2008). We expect IWAs to have a lower probability of backtracking: If the model predicts IWAs to backtrack, but not as often as controls, this could also be in line with the *resource reduction* hypothesis (Caplan, 2012). In unimpaired sentence comprehension, the DA model assumes that backtracking is a mechanism used on a certain proportion of trials when the initial interpretation of the sentence fails. If IWAs show a lower probability of backtracking, this could mean that even though they can backtrack, they do not do it as often as controls because the mechanism is disrupted. Alternatively, the  $P_b$  parameter could also be linked to *intermittent deficiencies*, because the process of backtracking could be intermittently disrupted. In addition, we expect the cost of backtracking,  $\delta$  to be higher for IWAs. This would reflect delayed

syntactic processing (Burkhardt et al., 2008). Finally, a larger  $\sigma$  would imply more noise in the retrieval mechanism for IWAs. This would be consistent with the *intermittent deficiency* hypothesis (Caplan et al., 2007) that postulates that IWAs suffer from intermittent reductions in the resources used in parsing.

## 6. Results

### 6.1. Results of the activation-based race model

We used the *rstan* package (Stan Development Team, 2020) to fit the models, with three chains, 6,000 iterations, and a warm-up of 3,000.<sup>5</sup> The chains were plotted and visually inspected for convergence. An additional metric of convergence is the so-called Rhat statistic (the ratio of between-to-within chain variance); when the sampler has converged, the Rhat statistic is close to 1 (Gelman et al., 2014). We checked that Rhats were always near 1. Two tuning parameters, delta and the tree depth,<sup>6</sup> were adapted when necessary for achieving convergence. Following Gelman et al. (2014), we also made sure that the parameters of the model could be recovered using simulated data (see the online supplementary materials).

The activation-based model assumes that for each trial, LTs are drawn from the two accumulators, and the accumulator with the fastest LT wins the race. The two distributions of finishing times (i.e., the finishing time of each one of the accumulators for each trial) can be plotted against each other, so as to assess the precise predictions of the model. For example, Fig. 6 shows the distribution of finishing times for the correct and the incorrect interpretation for each of the two groups, and across the two conditions. Fig. 6a,b displays the accumulators for controls, while 6c and 6d stand for IWAs' accumulators.

Fig. 6a displays the distribution of finishing times associated with the accumulator for the correct interpretation (SR) in dark gray, and for the incorrect interpretation (here OR) in light gray, for controls. The distribution for SR is clearly faster: The mean of the finishing times for the SR accumulator is 1,204 ms, whereas the mean finishing time for the OR accumulator is around 4,000 ms. In Fig. 6b, finishing times for the correct interpretation (OR, in light gray) are faster on average (1,655 ms) than the finishing times for the incorrect interpretation (SR, in dark gray, 4,647 ms). Therefore, Fig. 6a,b indicates that controls tend to choose the correct interpretation, since the distributions associated with the correct interpretations have faster finishing times.

Fig. 6c shows that IWAs also tend to choose the right interpretation in SRs. The mean of the accumulator for SR in the SR condition is 2,694 ms, whereas the mean of the OR accumulator is 4,717 ms. However, Fig. 6d indicates that it is difficult for IWAs to differentiate between the two interpretations in the OR condition (6d), where the two distributions show greater overlap. On average, the accumulator for the correct interpretation is faster: The estimated mean for the OR accumulator in the OR condition is 3,573 ms, whereas the estimated mean for the SR accumulator in the OR condition is 4,553 ms. But

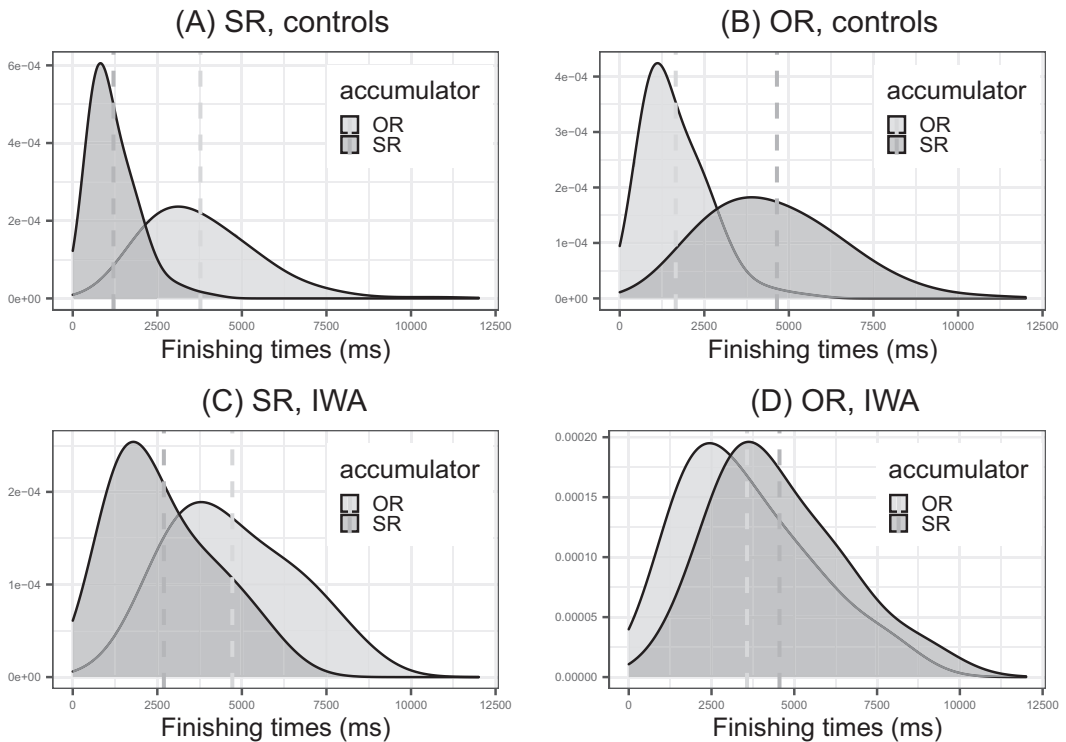


Fig. 6. Accumulators of evidence. The figure presents the distribution of finishing times associated with each accumulator in the activation-based model, across groups and conditions. The x-axis stands for finishing times (in milliseconds). The dashed lines represent the mean finishing time for the object relative clause interpretation (in light gray) and the subject relative clause interpretation (in dark gray).

the overlap between the two distributions shows that the accumulator for the incorrect interpretation is sometimes as fast as the one for the correct interpretation. Therefore, the model predicts a difficulty for IWAs in distinguishing between the correct and interpretation in ORs.

Fig. 6 shows that the model exhibits the predicted patterns: The means for the finishing times across conditions are slower for IWAs than for controls. For IWAs, the mean finishing times of the accumulator in the OR condition are more similar than for controls. We also predicted IWAs to have a higher  $\sigma$  because we assumed that their rate of accumulation could be noisier, and the model estimates reflect this prediction, as displayed in Fig. 7.

### 6.1.1. Posterior predictive checks

In order to evaluate the performance of the model, we compared the empirical data against the posterior predictive distributions estimated by the model (Gelman et al., 2014), a procedure that is known as posterior predictive checks (PPCs). We present the PPCs graphically, with violin plots, where the dots represent the mean of the empirical

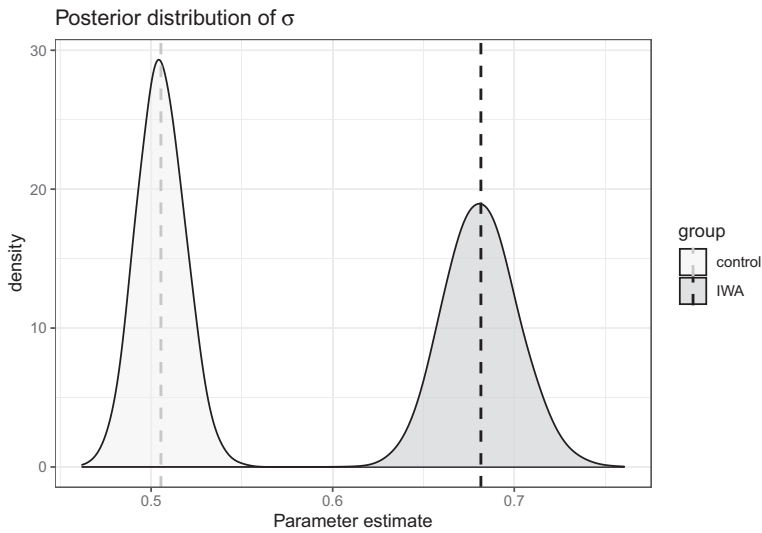


Fig. 7. Posterior distribution of the  $\sigma$  parameter for both groups, in log scale. The dashed lines show the mean of the distributions.

data. This is a way to inspect whether the data could have been generated by the models: If the mean of the empirical data is predicted by the model, that is, if the dot lies within the violin plots, the model could have generated the data. If the model is unable to reflect the distribution of the data, that implies a bad fit.

Fig. 8 shows the PPCs for the activation-based model in the picture-selection accuracies. In general, the activation-based model predicts the observed accuracies for both groups and conditions. Fig. 9 shows the PPCs corresponding to the LTs. The model can correctly estimate the LT distribution of the data across conditions and groups, although it tends to overestimate the LT for controls in incorrect responses.

## 6.2. Results of the direct-access model

The DA was fit with three chains and 7,000 iterations, and a warm-up of 3,500. The chains were visually inspected, and we verified that all the Rhats were close to 1. Delta and the tree depth parameters were adapted when necessary, and we made sure that the parameters of the model could be recovered using simulated data.

The DA model has three critical parameters: the probability of initial correct retrieval,  $\theta$ , the probability of backtracking if the initial retrieval is not correct,  $P_b$ , and  $\delta$ , which is the time taken for backtracking. We turn now to assess the posterior distributions for these parameters across groups and conditions.

The posterior distribution of  $\theta$  (Fig. 10a) indicates that in SRs, controls initially retrieve the target 83% of the time, whereas IWAs have a lower probability of initial correct retrieval, 69%. However, in ORs, the probability of initial correct retrieval is 41% for controls, and 53% for IWAs. We discuss this surprising outcome below.

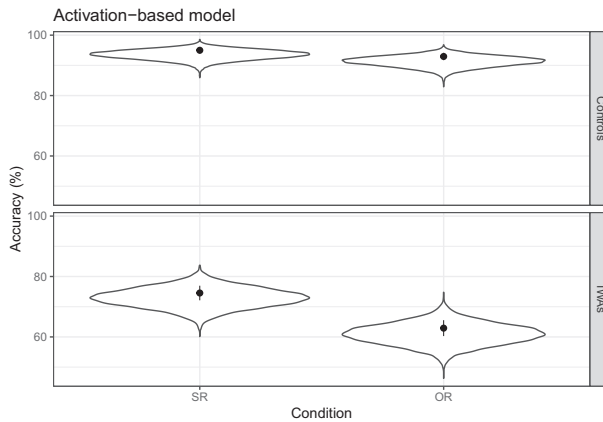


Fig. 8. Violin plots depicting the PPCs for the activation-based model corresponding to the accuracy responses split by group and condition. The black dots represent the mean proportions of responses in the data and the corresponding error bars show 95% confidence intervals, and the violin plots display the posterior predicted distributions from the model. Note that the controls' confidence intervals are not visible because variability is low in this group.

Regarding the probability of backtracking, the posterior distribution of the parameter  $P_b$  (Fig. 10b) indicates that controls perform backtracking around 82% if they initially retrieve the distractor, whereas IWAs backtrack 21% of the time. Notice that the parameters  $\theta$  (Fig. 10a) and  $P_b$  (Fig. 10b) are interrelated, and they should be interpreted together. The interpretation of both parameters shows that:

- a. Controls initially carry out a retrieval that leads to the correct interpretation most of the time in SRs (83%), and 41% in ORs. If the first retrieval was incorrect, they backtrack and get the correct interpretation in 82% of the cases.
- b. IWAs are estimated to retrieve the correct interpretation without backtracking for SRs about 69% of the time and for ORs 53% of the time. However, IWAs backtrack only 21% after an incorrect first retrieval. Therefore, misretrievals are more likely for IWAs than controls, especially in ORs.

Fig. 11 shows the estimated time needed for backtracking. The posterior of  $\delta$  shows that backtracking takes less time for controls, with a mean centered around 546 ms. By contrast, IWAs' estimate for  $\delta$  is higher, around 678 ms.

We predicted IWAs to have a lower probability of backtracking relative to controls, and Fig. 11 shows that the model confirms our prediction. We also predicted controls to have higher values for  $\mu$  and  $\sigma$ . The model estimates are in line with these predictions (see Fig. 12). However, the model's estimates contradict our prediction about  $\theta$ : We had assumed that due to *resource reductions*, IWAs should have a lower probability of initial correct retrieval in ORs. This surprising outcome in the DA is an inherent shortcoming of the model, at least under the assumptions made here. We discuss alternative explanations in Section 9.

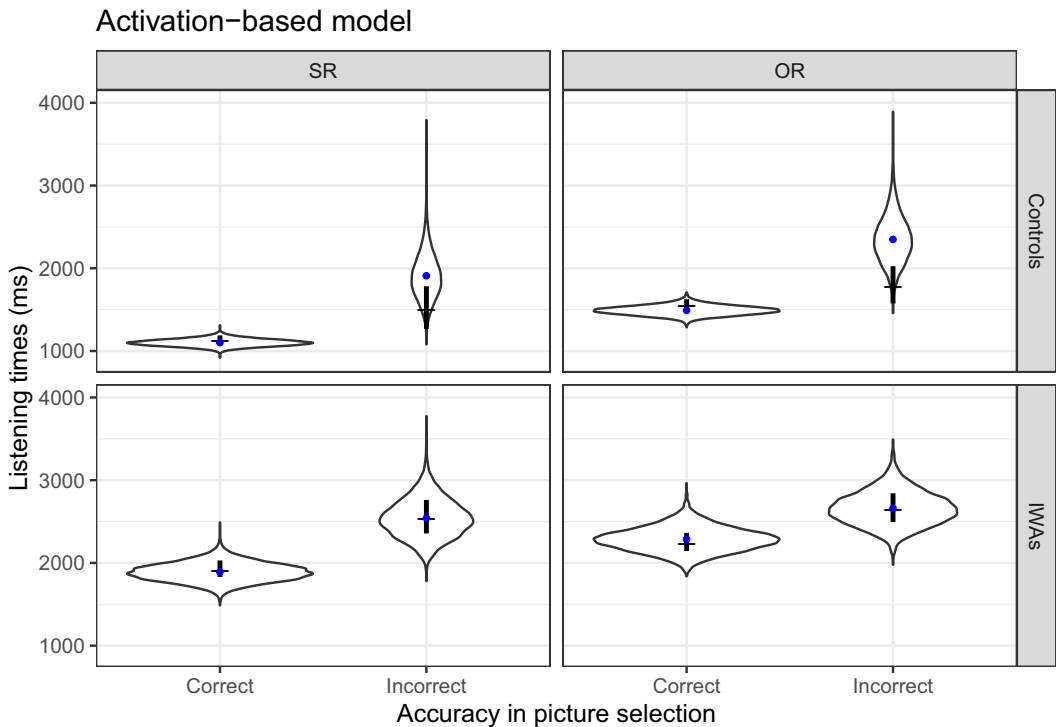


Fig. 9. Violin plots depicting the PPCs of the activation-based model for listening times split by group and condition. The listening times correspond to the sum of the listening times for the verb of the subordinate clause plus the listening times of the second noun phrase. The horizontal bars represent the mean of the data and the vertical bars are the standard error of the mean. The dots represent the mean of the posterior distribution.

### 6.2.1. Posterior predictive checks

As with the activation-based model, we graphically compare the distribution of the empirical data with the estimated posteriors of the model. Fig. 13 shows the PPCs corresponding to the picture-selection accuracies. In general, the model correctly predicts the qualitative pattern of the observed accuracies. Fig. 14 shows that the model estimates the LTs across conditions and groups, but it tends to underestimate the LTs for incorrect responses, and overestimate the correct responses in the SRs condition for IWAs.

## 7. Quantitative comparison of the activation model and the direct-access model

Although PPCs offer a visual way to assess the descriptive adequacy of the models, a more quantitative way of model assessment is required, in order to measure which model fits the data better. We compared the predictive accuracy of the models using 10-fold cross-validation (Vehtari, Gelman, & Gabry, 2017). Cross-validation in the Bayesian

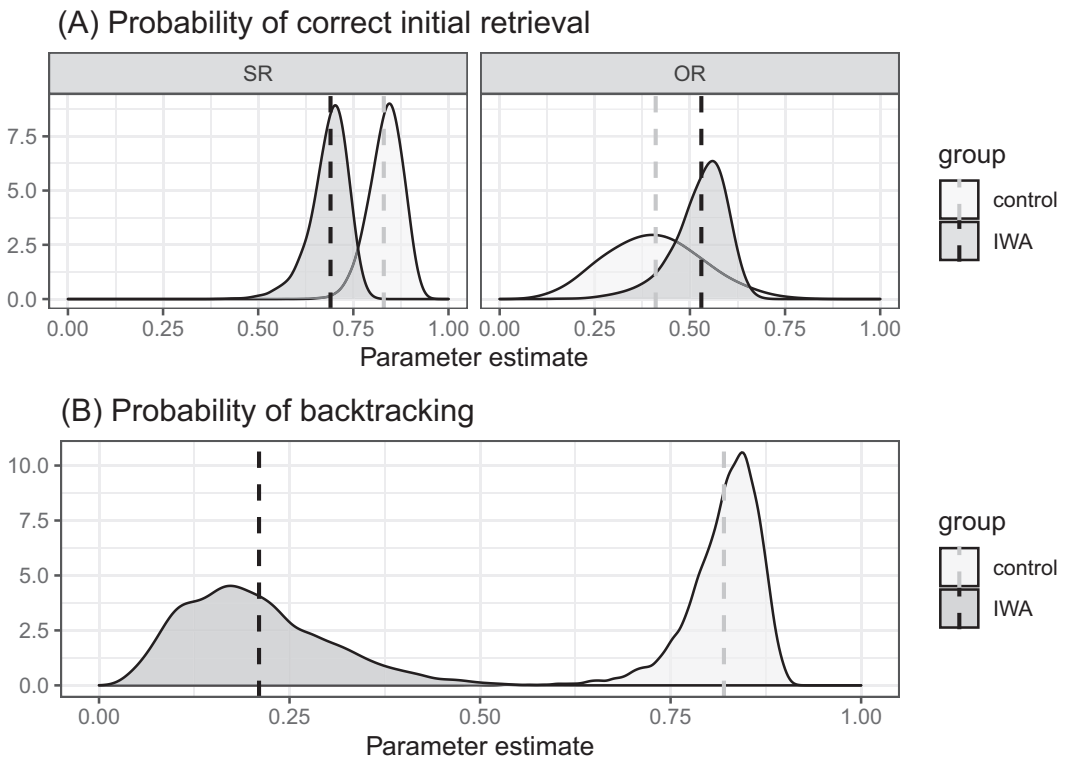


Fig. 10. Posterior distribution for the probability of initial correct retrieval and backtracking in the direct-access model. This figure shows the estimated probability of initial correct retrieval across groups and conditions in the upper panel, and the estimated probability of backtracking across groups in the lower panel. The dashed lines stand for the means of their respective distributions.

framework allows for comparisons of models that assume different generative processes for the data, such as the two models in this study. A 10-fold cross-validation involves splitting the dataset into 10 subsets of balanced data (balanced here means that each participant contributes approximately the same amount of data). One of the subsets is held out, and the model is fit to the nine remaining subsets. The posterior distributions of the parameters of this model are used to compute predictive accuracy on the subset of held-out data. This procedure is then repeated 10 times, one for each subset of held-out data. The difference between predicted and observed held-out data points is used to compute a measure of predictive accuracy: the *expected log point-wise predictive density*, or  $\widehat{elpd}$ . When comparing two models, the model with the higher  $\widehat{elpd}$  value is the model that represents a better fit to the data. The standard deviation of the sampling distribution of  $\widehat{elpd}_{diff}$ , the difference in  $\widehat{elpd}$ , can also be computed, and has the standard frequentist interpretation:  $\widehat{elpd}_{diff} \pm 2 \times SE$  can be interpreted as a 95% confidence interval.

The  $\widehat{elpd}$  values yielded a difference of 115 ( $SE = 69$ ) in favor of the activation-based model. This suggests that the activation-based model shows a somewhat better fit for our

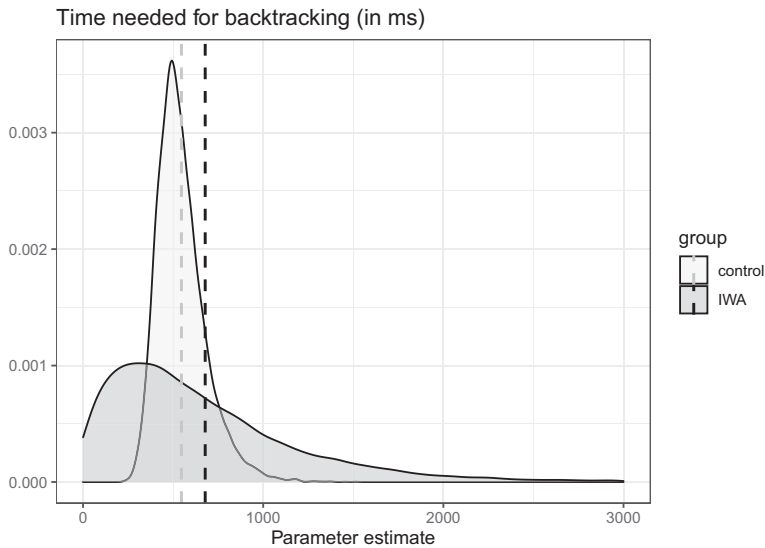


Fig. 11. Posterior distributions of parameters representing the effect of backtracking (in milliseconds). The figure shows the posterior distribution of estimated time needed for backtracking across groups.

data ( $\widehat{elpd}_{act} = -12,515$ ,  $SE = 49$  and  $\widehat{elpd}_{DA} = -12,630$ ,  $SE = 52$ ). However, the relatively large standard error means that the difference in the predictive performance of the models is not decisive. Table 1 details the difference in  $\widehat{elpd}$  by condition and group, and their corresponding  $SE$ . Although the activation-based model consistently shows an advantage across conditions and groups, the standard errors indicate that the differences are not decisive.

In this section, the relative performance of the models was assessed. We turn now to assess the relative importance that the individual parameters within each model have, in terms of explaining the data from IWAs.

## 8. Model evaluation using Bayes factors

The estimates from the activation-based model and the DA show that IWAs behave differently from controls. As discussed in the previous sections, given our linking assumptions, the different parameter estimates for the two groups can tell us whether the deficits that we link to the different parameters can explain IWAs' data. For instance, the larger  $\sigma$  that IWAs have in both models (relative to controls) indicates that *intermittent deficiencies* may be one of the causes of IWAs' processing difficulties.

One question that arises is, to which extent is there evidence that these deficits are playing a role in IWAs' sentence comprehension? By assumption, both models had group adjustments in all of the parameters. These adjustments reflect the difference between IWAs and controls. However, if the group adjustment of a given parameter does not



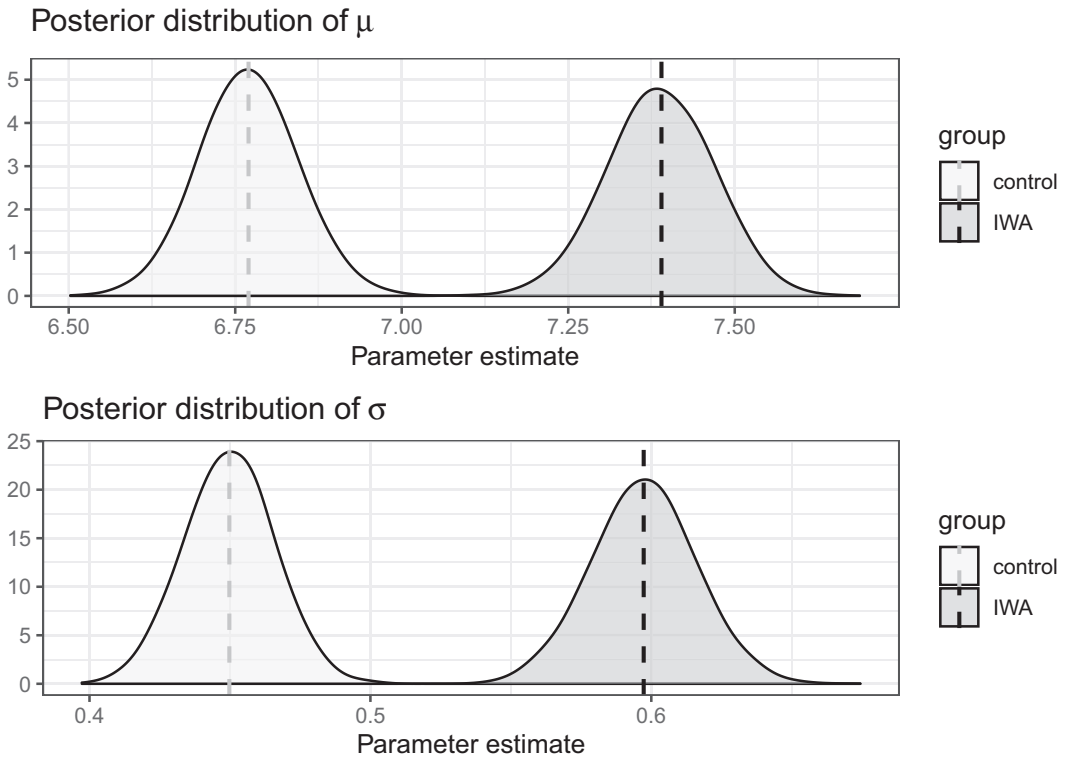


Fig. 12. Parameter distributions for the  $\mu$  and  $\sigma$  parameters across groups, on the log scale. The dashed lines stand for the means of their respective distributions.

improve the model fit (i.e., the model would perform better if no difference was assumed between IWAs and controls), this could mean that the processing deficit we are linking to this parameter may not be playing a role in impaired sentence comprehension. One way to assess whether the group adjustments improve the models' fit is to compute a series of BFs.

The BF quantifies the evidence against or in favor of a null model ( $M_0$ ) that does not assume an effect of group (no  $\beta$  adjustment for the group factor), relative to a model that assumes a group effect ( $M_1$ ). The BF is a ratio of marginal likelihoods (as shown in Eq. 14), and it indicates how likely it is that the data have been generated by one model relative to the other one. In Eq. 14, the subscript in  $BF_{10}$  stands for the order of the models: Evidence of  $M_1$  over  $M_0$ .

$$BF_{10} = \frac{P(Data|Model1)}{P(Data|Model0)} \tag{14}$$

The interpretation of BF is done in terms of relative odds. For instance, a  $BF_{10}$  of 5 means that the odds are 5:1 in favor of  $M_1$ . A BF closer to 1 is inconclusive, whereas a  $BF_{10}$  larger than 1 indicates evidence in favor of  $M_1$ , and  $BF_{10}$  below 1 indicates

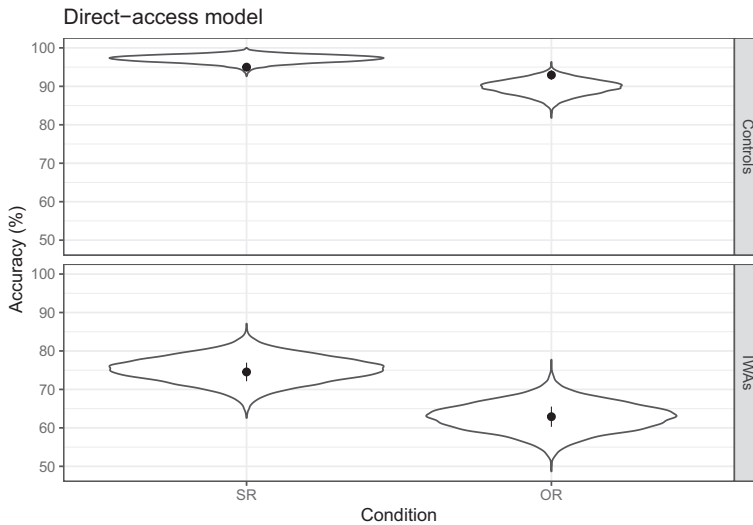


Fig. 13. Violin plots depicting the PPCs for the DA model corresponding to the accuracy responses split by group and condition. The black dots represent the mean proportions of responses in the data, whereas the violin plots display the posteriors estimated by the model.

evidence in favor of  $M_0$ . The BF has a continuous scale (meaning the higher the  $BF_{10}$ , the stronger the evidence for  $M_1$ ). There is no specific cutoff for the interpretation of the strength of the evidence in favor of a model over the other one, but guidelines have been proposed (Jeffreys, 1939/1998). In general, a  $BF_{10}$  larger than 100 is considered as strong evidence in favor of  $M_1$ . Conversely, a  $BF_{10}$  of  $1/100$  or smaller is considered as strong evidence in favor of  $M_0$ .

BF and cross-validation are two different ways to perform model comparisons. Cross-validation is well suited for comparing models with different generative processes (such as the activation-based model vs. the DA), but cross-validation may be problematic with models that make very similar predictions. In this case, the estimated standard error might be biased (Sivula, Magnusson, & Vehtari, 2020). Since our model evaluation at the parameter level involves comparing nested models that are likely to make similar predictions, in this section we use BFs instead of cross-validation. In what follows we perform a BF analysis for each parameter of the two models that has an adjustment for the group factor. For instance, for the  $\sigma$  parameter in both models, the  $M_0$  (null model) and  $M_1$  would be as shown in Eq. 15.

$$\begin{aligned} M_{0\sigma} &: \sigma_0 \\ M_{1\sigma} &: \sigma_0 + \beta \times \text{group} \end{aligned} \quad (15)$$

Because BF is known to be sensitive to the choice of priors (Rouder, Haaf, & Vandekerckhove, 2018), we ran  $M_1$  with three different standard deviations for the prior of the  $\beta$  of interest (the adjustment for group) in order to show how the BF changes as a function

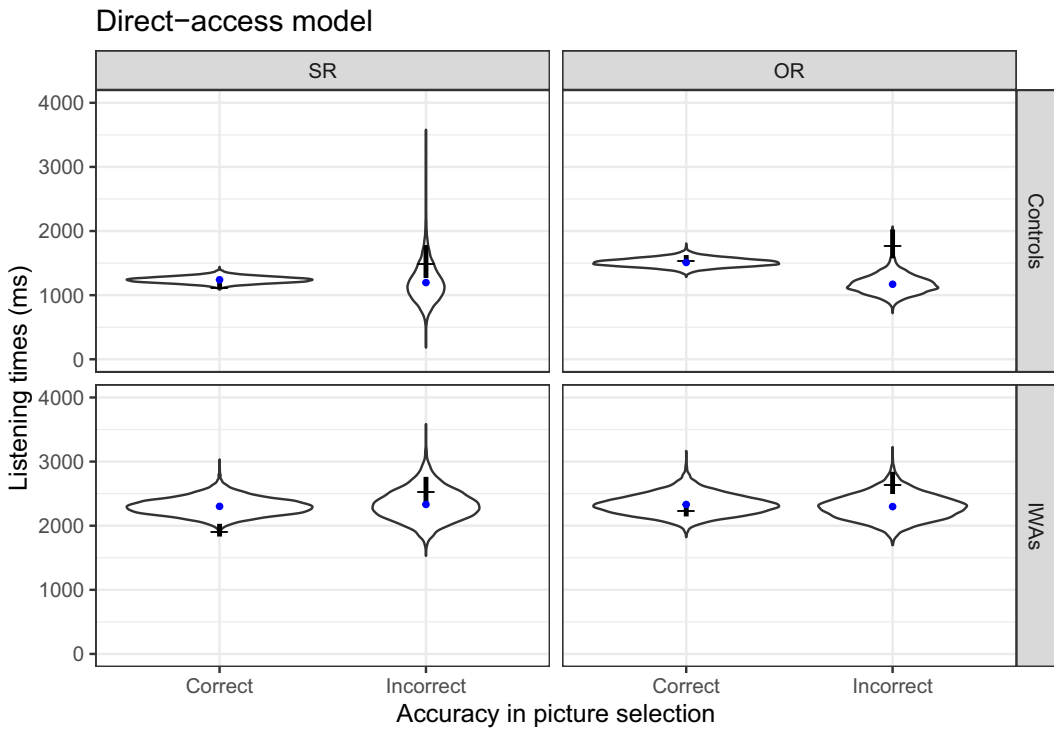


Fig. 14. Violin plots depicting the PPCs of the direct-access model for listening times split by group and condition. The listening times correspond to the sum of the listening times for the verb of the subordinate clause plus the listening times of the second noun phrase. The horizontal bars represent the mean of the data, and the vertical bars are the standard error of the mean. The blue dots represent the mean of the posterior distribution.

Table 1

$\widehat{elpd}$  differences between the activation-based and the direct-access model across conditions and groups. A positive difference indicates an advantage for the activation-based model

	$\widehat{elpd}$ Difference	SE
SR, Controls	31	37
OR, Controls	42	36
SR, IWAs	28	33
OR, IWAs	14	33

of the prior standard deviation. The prior was always centered at 0 and the standard deviations were 0.1, 0.3, and 0.5. In addition, we included the following constraints:

$$\begin{aligned}
 \text{Controls: } & \text{Intercept} + (-1) \times \beta = \text{Intercept} - \beta \\
 \text{IWAs: } & \text{Intercept} + (+1) \times \beta = \text{Intercept} + \beta
 \end{aligned}
 \tag{16}$$

- i. For the parameter  $\mu$  in both models, and  $\delta$  in DA, the group  $\beta$  in M1 was constrained to be positive. These parameters reflect the mean LTs and the time needed for backtracking, respectively. Therefore, according to theory, due to *slow syntax* and/or *delayed lexical access*, IWAs should be slower than controls. Because the contrast coding is +1 for IWAs and -1 for controls, a positive  $\beta$  would indicate that controls are faster than IWAs, as shown in Eq. 16.

$$LT \sim \text{lognormal}(\mu + \delta, \sigma).$$

- ii. Similarly, for the parameter  $\sigma$  in both models, the group  $\beta$  was also constrained to be positive, since according to *intermittent deficiencies*, IWAs should have more noise in the processing system.
- iii. We assumed that the probability of initial correct retrieval and the probability of backtracking could be linked to the *resource reduction* hypothesis. Therefore, IWAs should show a lower  $\theta$  and  $P_b$  estimate, and the group  $\beta$  was thus constrained to be negative. Since IWAs are contrast coded +1, a negative  $\beta$  would imply a lower estimated probability for IWAs.
- iv. In the activation-based model, a condition  $\times$  group interaction is assumed on the  $\mu$  parameter. The priors for the effect of this interaction should be vague because there is no prediction about the direction of the effect. One could assume that (a) IWAs are more affected by the condition manipulation than controls, or (b) IWAs are less affected by the condition manipulation than controls, because IWAs perform poorly in both conditions. Therefore, the  $\beta$  for the interaction did not have any constraint. And similarly, the  $\beta$  for the interaction in the  $\theta$  parameter in DA was not constrained either.

A summary of the models that were run and their corresponding prior *SD* is shown in Table 2. All the BF were computed using the *bridgesampling* R package (Gronau, Singmann, & Wagenmakers, 2017) after running the models for 40,000 iterations. In addition, some of the models were run three times in order to confirm that the number of iterations was high enough to produce stable BF. Notice that for all parameters, M0 is the model that has no adjustment for the group effect. Three versions of M1 were run, each with a different prior *SD* for the group adjustment, as shown in Table 2. In the case of parameters with an interaction, nine versions of M1 were run, one for each possible combination of the prior *SD* of the two adjustments (the  $\beta$  for the group effect and the  $\beta$  for the interaction condition  $\times$  group).

## 8.1. Results

### 8.1.1. Activation-based model

In the activation-based model there are two  $\mu$  parameters, one for each accumulator of evidence,  $\mu_{SR}$  and  $\mu_{OR}$ .  $M0_{\mu}$  does not include any adjustment for the effect of group or

Table 2

Summary of the BF analysis for both models. This table shows the priors used, the theories that map to each parameter, and the BF results

Model	Param.	Group SD	Inter. SD	Theory	BF <sub>10</sub>
ACT	$\mu$	0.1, 0.3, 0.5, +	0.1, 0.3, 0.5	Slow syntax, DLA	1/3 to 1
ACT	$\sigma$	0.1, 0.3, 0.5, +		Intermittent deficiencies	>100
DA	$\mu$	0.1, 0.3, 0.5, +		Slow syntax, DLA	>100
DA	$\theta$	0.1, 0.3, 0.5, -	0.1, 0.3, 0.5	Resource reduction	>100
DA	$P_b$	0.1, 0.3, 0.5, -		Resource reduction	2 to >100
				Intermittent deficiencies	
DA	$\delta$	0.1, 0.3, 0.5, +		Slow syntax	1/3 to 1/11
DA	$\sigma$	0.1, 0.3, 0.5, +		Intermittent deficiencies	>100

ACT stands for the activation-based model, and DLA stands for delayed lexical access theory. The columns “Group SD” and “Inter. SD” show the different prior SD of the  $\beta$  adjustments to the effect of group and the interaction group  $\times$  condition, respectively. In the “Group SD” column, a plus sign indicates that the  $\beta$  for the group adjustment was constrained to be positive, and a minus sign indicates that the  $\beta$  was constrained to be negative. No constraints were applied to the  $\beta$  of the interactions. The column “BF<sub>10</sub>” summarizes the range of BF results for the priors shown in the table

the interaction group  $\times$  condition, for any of the two accumulators.  $M1_\mu$  includes an adjustment for the effect of group and another adjustment for the interaction, for both accumulators. This is shown in more detail in Eq. 17.

$$\begin{aligned}
 &M0_\mu \\
 &\mu_{SR} = \alpha_1 + u_{\alpha_1} + w_{\alpha_1} + (\beta_1 + u_{\beta_1}) \times rc_{type} \\
 &\mu_{OR} = \alpha_2 + u_{\alpha_2} + w_{\alpha_2} + (\beta_2 + u_{\beta_2}) \times rc_{type} \\
 &M1_\mu \\
 &\mu_{SR} = \alpha_1 + u_{\alpha_1} + w_{\alpha_1} + (\beta_1 + w_{\beta_1}) \times group \\
 &\quad + (\beta_3 + u_{\beta_3}) \times rc_{type} + \beta_5 \times group \times rc_{type} \\
 &\mu_{OR} = \alpha_2 + u_{\alpha_2} + w_{\alpha_2} + (\beta_2 + w_{\beta_2}) \times group \\
 &\quad + (\beta_4 + u_{\beta_4}) \times rc_{type} + \beta_6 \times group \times rc_{type}
 \end{aligned}
 \tag{17}$$

The BF results are summarized in Table 2. The BFs for  $\mu$  in the activation-based model are either inconclusive or yield anecdotal evidence in favor of the model that does not assume a difference between controls and IWAs ( $M0_\mu$ ).<sup>7</sup> In contrast, the BF results for  $\sigma$  yield strong evidence in favor of  $M1_\sigma$ : The model with a group adjustment for  $\sigma$  provides a better fit. This suggests that the group adjustment in  $\sigma$  could be sufficient to explain the differences between the two groups. Given our linking assumption, this means that the activation-based model estimates intermittent deficiencies to be the main source of processing deficits in IWAs.

### 8.1.2. Direct-access model

In the DA, the  $\theta$  parameter also has a  $\beta$  for the interaction group  $\times$  condition in addition to the  $\beta$  for the group effect. For the BF analysis, the  $M0_\theta$  does not have any of these  $\beta$ , whereas the  $M1_\theta$  has both, as shown in Eq. 18.

$$\begin{aligned}
 &M0_\theta \\
 &\theta = \alpha + u_\alpha + w_\alpha + (\beta_2 + u_{\beta_2}) \times rc_{type} \\
 &M1_\theta \\
 &\theta = \alpha + u_\alpha + w_\alpha + (\beta_2 + u_{\beta_2}) \times rc_{type} \\
 &\quad + (\beta_3 + w_{\beta_3}) \times group + \beta_4 \times group \times rc_{type}
 \end{aligned} \tag{18}$$

Nine versions of  $M1_\theta$  models were run (see Table 2), such that all possible combinations of prior  $SD$  for both adjustments could be considered. All BF for  $\theta$  yield strong evidence in favor of  $M1$ , the model that assumes that IWAs have a lower probability of initial correct retrieval relative to controls (due to resource reductions). Irrespective of the prior  $SD$ , the BFs for  $\mu$  and  $\sigma$  yield strong evidence in favor of  $M1$ . The BF for  $P_b$  yields anecdotal to strong evidence in favor of  $M1$  depending on the priors. In general, all of these parameters benefit from a group adjustment.

By contrast, the BF for  $\delta$  yields some evidence in favor of  $M0$ , suggesting that the group adjustment is not needed. Recall that  $\delta$  is the time needed for backtracking, and that estimated LTs for trials with backtracking are drawn from  $(\mu + \delta, \sigma)$ . The BF for  $\delta$  could indicate that the group  $\beta$  is redundant because  $\mu$  and  $\sigma$  (with their corresponding group adjustments) already explain the differences between controls and IWAs. This means that IWAs may not have an impairment in the mechanism of backtracking. That is, IWAs perform backtracking less often than controls (as estimated in  $P_b$ ), but when they do backtrack, the mechanism is not disrupted. These results suggest that the DA accounts for *slow syntax* and/or *delayed lexical access* in  $\mu$  (mean LTs), but not in  $\delta$  (time needed for backtracking).

In conclusion, the BF analyses at the individual parameter level revealed that in the activation-based model, an increased noise value for IWAs can explain the processing differences between IWAs and controls, which speaks in favor of the *intermittent deficiencies* theory. The model could also be in line with *slow syntax* and/or *delayed lexical access*, but the BF for the parameter linked to these theories was inconclusive, so the role of these deficits in the activation-based model remains unclear. By contrast, the DA is in line with a mixture of *slow syntax* and/or *delayed lexical access*, *resource reduction*, and *intermittent deficiencies*.

## 9. Discussion

In this study we presented a Bayesian implementation of two models of cue-based retrieval: the activation-based model and the DA. We linked the parameters of these

models to major theories of processing deficits in sentence comprehension in aphasia, namely *slow syntax*, *delayed lexical access*, *resource reduction*, and *intermittent deficiencies*. The predictive performance of the two models was assessed with 10-fold cross-validation, and the quantitative and qualitative predictions of the models concerning data from IWAs and controls have been discussed. A BF analysis was performed, in order to quantify the evidence that the models had with respect to the different processing deficits that were evaluated. In what follows we discuss some unexpected aspects of the DA, we compare our findings to prior computational modeling work in the field of aphasia, and we point out some limitations of the present work as well as future directions.

### 9.1. Unexpected behavior of the direct-access model

The DA estimates IWAs to have a higher probability of initial correct retrieval in ORs relative to controls, which is surprising, since ORs are generally more difficult to process for IWAs than for controls (Caramazza & Zurif, 1976). However, this prediction would be in line with studies showing that unimpaired controls have an agent-first preference: Unimpaired controls tend to interpret the first NP of a clause as the agent, which clashes with the actual thematic relations in some constructions (Hanne, Burchert, De Bleser, & Vasishth, 2015; Mack, Wei, Gutierrez, & Thompson, 2016). For instance, in an eye-tracking experiment involving a sentence–picture matching task with active and passive sentences such as (5a) and (5b), Mack et al. (2016) found that unimpaired controls showed initial agent-first processing followed by a thematic reanalysis. That is, in passive sentences, controls tended to initially look at the image in which the first noun phrase was the agent. After hearing the region that contained the disambiguating morphological information (i.e., the verb: *visiting/visited*), controls started fixating the target picture. This implies that controls, after processing the morphological cues, had to reanalyze the initial agent-first interpretation. By contrast, in the study of Mack et al. (2016), IWAs did not show signs of agent-first processing: They looked at the target and distractor pictures equally prior to the arrival of the disambiguating information.

- (5) a. **Active:** The man was visiting the woman.  
 b. **Passive:** The man was visited by the woman.

Previous studies where controls showed an agent-first bias used eye-tracking and the visual world paradigm, but our modeling suggests that the agent-first bias could also be detected in a self-paced listening experiment. In our data, if unimpaired controls experienced an initial agent-first bias in ORs, they would initially parse the sentence as an SR. Consider sentence (6). Once they hear the disambiguating region (e.g., second noun phrase in sentence 6), they would have to backtrack on a high proportion of trials to end up with the right thematic interpretation. In this regard, the estimates for controls in ORs

would be in line with an initial agent-first strategy. However, a replication of these estimates would be needed, ideally with visual-world eye-tracking data, as in Hanne et al. (2015) or Mack et al. (2016).

(6) **OR:** The girl who the mother chased hugged the boy.

Finally, a major issue for the DA model is the fact that the data show longer LTs for incorrect responses. This pattern contradicts the core assumptions of the model, because correct responses are expected, on average, to take longer due to the cost of backtracking.<sup>8</sup> Intuitively, IWAs' incorrect responses may be associated with longer LTs because after backtracking IWAs may not be able to retain the retrieved representation. The *slow syntax* and the *resource reduction* hypotheses would be compatible with this view. However, the data show that the incorrect responses of unimpaired controls are also associated with longer LTs relative to correct responses. Therefore, the assumption that backtracking leads to the retrieval of the target (McElree, 1993) seems incompatible with our data.

## 9.2. Comparison with previous computational modeling work on aphasia

Taken together, the higher  $\widehat{elpd}$  value in favor of the activation-based model, plus the fact that the DA underestimates the LTs for incorrect responses in ORs, suggests that the activation-based model is better at characterizing the processing of RCs in IWAs and controls. The BF analysis for the activation-based model highlights the role that intermittent deficiencies may be playing in an activation-based mechanism of retrieval, but *slow syntax* and/or *delayed lexical access* should not be ruled out, since the BFs for the parameters associated with these theories were rather inconclusive.

Our results are consistent with previous sentence processing modeling work on aphasia. For example, Patil et al. (2016) found that the LV05 model that included slowed processing (understood as a slowdown in the parsing mechanism) and intermittent deficiencies showed the best fit to data from IWAs, relative to models that included only one of these deficits. It is also possible that IWAs may exhibit different degrees of these deficits, as suggested by Mätzig et al. (2018), who modeled the accuracies of the Caplan et al. (2013) dataset estimating ACT-R parameters at the individual level. Interestingly, their modeling also revealed that *intermittent deficiencies* was the deficit that affected most of the IWAs. Out of the 56 IWAs, 53 showed a higher noise value (relative to the default noise value in ACT-R) in OR clauses. Unfortunately, we do not have enough LT data to get robust parameter estimates at the individual level, but our modeling suggests that on average, IWAs are more subject to intermittent deficiencies than to *slow syntax* and/or *delayed lexical access*.

One caveat that applies to Patil et al. (2016), Mätzig et al. (2018), and our own work, is that the models cannot distinguish between *slow syntax* and *delayed lexical access*. In our implementation of the activation-based model and the DA, one possibility would be to include a shift parameter (Rouder, 2005) that accounts for lexical access, as implemented in Nicenboim and Vasishth (2018). Ideally, this parameter should have a group



adjustment (to assess whether there is a delay in lexical access in IWAs on average, taking the estimate for controls as reference), and an individual adjustment, to assess to which extent each individual is affected by this deficit. Unfortunately, such parameter could not be fit due to data sparsity.

Another issue to consider is that our modeling is limited to sentence comprehension. There is important modeling work in the aphasia literature that focuses on lexical processing (Evans, Hula, & Starns, 2019; Mirman, Yee, Blumstein, & Magnuson, 2011), the interface between lexical access and word production (Dell, Lawler, Harris, & Gordon, 2004), and word production (Walker, Hickok, & Fridriksson, 2018), among others. Ideally, a model of impairments in IWAs should account for both aphasic comprehension and production, and disentangle the difficulties that arise from lexical and syntactic processes. However, as we show in this study, there is no single parameter that can account for aphasic impairments, and it is very unlikely that a computational model, even with a larger number of parameters, could account for all the particularities of aphasic performance, which is variable in nature. Nevertheless, we believe that more computational modeling is needed in the field of aphasia, in order to better understand the underlying nature of language impairments in IWAs. Computational models require researchers to formalize hypotheses and assumptions, which is essential for theory development (Guest & Martin, 2020).

### 9.3. *Some limitations of the present work and future directions*

An important limitation of the present work is that even though the Caplan et al. (2015) dataset on IWAs and age-matched controls is the largest currently in existence, the data are still relatively sparse compared to standard datasets used for similar model comparisons in psycholinguistics, both in terms of the number of items (10) and participants (33 IWAs and 46 controls). For example, Nicenboim and Vasishth (2018) compared the predictive performance of the activation-based model and DA from reading time data from some 180 participants. It would be useful to revisit these model comparisons with larger datasets in the future. Another important step will be to test the two models against new experimental designs and with different experimental paradigms. This would allow for a more comprehensive evaluation of the differences between the models, as well as an assessment of their predictive ability when modeling interference effects in different tasks, languages, and conditions. We are currently compiling a comprehensive database containing several tasks and conditions of data from IWAs and unimpaired controls in (Pregla, Lissón, Vasishth, Burchert, & Stadie, 2020). In future work, we intend to use this database to further evaluate the models discussed here.

## 10. Conclusion

We compared the predictive performance of two competing models of cue-based retrieval using data from IWAs and age-matched controls. We tested whether the two

models—the activation-based model and the DA—could account for experimental data from both IWAs and controls. This is the first study where competing models of cue-based retrieval have been tested against data from impaired populations. We also investigated the relative importance of the various parameters in both models using BFs. The BF analyses show that in the activation-based model, *intermittent deficiencies* (Caplan et al., 2015) best explains the behavioral data from IWAs, although *slow syntax* (Burkhardt et al., 2008) and *delayed lexical access* (Ferrill et al., 2012) may also play a role. In the DA, the behavior of IWAs is best explained in terms of a combination of *slow syntax*, *delayed lexical access*, *resource reduction* (Caplan, 2012), and *intermittent deficiencies*. The model comparisons show that both models have a similar performance for out-of-sample predictions (assessed with 10-fold cross-validation), with a slight advantage for the activation-based model.

In closing, we have presented the first-ever computational evaluation of different models of dependency completion, using the largest available database from IWAs and unimpaired controls that currently exists. Our work lays out a systematic workflow that can be used to quantitatively compare the predictions of competing models of language processing.

## Acknowledgments

This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project number 317633480 – SFB 1287, project B02 (PIs: Shrivastava Vasishth, Frank Burchert, and Nicole Stadie).

## Open Research badges



This article has earned Open Data and Open Materials badges. Data and materials are available at <https://osf.io/kdjqz/>.

## Notes

1. For alternative accounts, see, for example, Grodzinsky (1995), Grillo (2009), or Engel et al. (2018). A complete summary of the theories of processing deficits in aphasia can be found in Caplan et al. (2015).
2. As a reviewer points out, it seems a bit confusing to talk about the interpretation of a sentence, since we are modeling the critical region, and there are several upcoming words that still have to be processed in order to finish the whole sentence. What is meant here by interpretation of a sentence as SR is that the first noun

phrase has been retrieved (i.e., the first noun phrase is interpreted as the agent) versus the interpretation of a sentence as OR, where the second noun phrase is retrieved as the agent.

3. We also fit a model with different variances for correct and incorrect responses, as introduced in Nicenboim and Vasissth (2018). However, the quantitative difference in predictive performance between the model with a single variance and the model with two variances was negligible. Both models show a comparable quantitative fit to the data. Here, we report the model with a single variance for correct and incorrect responses.
4. The prior distributions of the main parameters are plotted in the online supplementary materials.
5. The code for both the activation-based and the direct-access models is available at <https://bit.ly/3lda7Qj>.
6. The adjustment of these tuning parameters (`adapt_delta`, `max_treedepth`) leads to the whole posterior distribution of the parameters being correctly explored by the Hamiltonian Monte Carlo algorithm used in Stan. See the Stan manual or the short guide on warnings for more information (<https://mc-stan.org/misc/warnings>).
7. A series of tables and plots showing the BF as a function of the priors for all of the parameters in both models is available in the online supplementary materials.
8. Notice, however, that due to random noise the model estimates slower incorrect responses in some trials, as shown in the tails of the distribution for incorrect responses in Fig. 14.

## References

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*(4), 1036–1060.
- Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48.
- Burkhardt, P., Avrutin, S., Piñango, M. M., & Ruigendijk, E. (2008). Slower-than-normal syntactic processing in agrammatic broca's aphasia: Evidence from Dutch. *Journal of Neurolinguistics*, *21*(2), 120–137.
- Burkhardt, P., Piñango, M. M., & Wong, K. (2003). The role of the anterior left hemisphere in real-time sentence comprehension: Evidence from split intransitivity. *Brain and Language*, *86*(1), 9–22.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.
- Caplan, D. (2012). Resource reduction accounts of syntactically based comprehension disorders. In C. K. Thompson & R. Bastianse (Eds.), *Perspectives on agrammatism* (pp.34–48). New York, NY: Psychology Press.
- Caplan, D., Michaud, J., & Hufford, R. (2013). Dissociations and associations of performance in syntactic comprehension in aphasia and their implications for the nature of aphasic deficits. *Brain and Language*, *127*(1), 21–33.
- Caplan, D., Michaud, J., & Hufford, R. (2015). Mechanisms underlying syntactic comprehension deficits in vascular aphasia: New evidence from self-paced listening. *Cognitive Neuropsychology*, *32*(5), 283–313.

- Caplan, D., Waters, G., DeDe, G., Michaud, J., & Reddy, A. (2007). A study of syntactic processing in aphasia: Behavioral (psycholinguistic) aspects. *Brain and Language*, *101*(2), 103–150.
- Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and Language*, *3*(4), 572–582.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32.
- Dell, G. S., Lawler, E. N., Harris, H. D., & Gordon, J. K. (2004). Models of errors of omission in aphasic naming. *Cognitive Neuropsychology*, *21*(2–4), 125–145.
- 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.
- Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*, *43*(12), e12800.
- Evans, W. S., Hula, W. D., & Starns, J. J. (2019). Speed–accuracy trade-offs and adaptation deficits in aphasia: Finding the “sweet spot” between overly cautious and incautious responding. *American Journal of Speech-Language Pathology*, *28*(1S), 259–277.
- Ferrill, M., Love, T., Walenski, M., & Shapiro, L. P. (2012). The time-course of lexical activation during sentence comprehension in people with aphasia. *American Journal of Speech-Language Pathology*, *21*(2), S179.
- Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: Advantages, applications, and extensions. *Annual Review of Psychology*, *67*, 641–666.
- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: CRC Press.
- Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. *Image, Language, Brain*, *2000*, 95–126.
- Goodglass, H., Kaplan, E., & Barresi, B. (2001). *Bdae-3: Boston diagnostic aphasia examination*. Philadelphia, PA: Lippincott Williams & Wilkins.
- Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-based interference during language comprehension: Evidence from eye tracking during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1304–1321.
- Grillo, N. (2009). Generalized minimality: Feature impoverishment and comprehension deficits in agrammatism. *Lingua*, *119*(10), 1426–1443.
- Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–290.
- Grodzinsky, Y. (1995). A restrictive theory of agrammatic comprehension. *Brain and Language*, *50*(1), 27–51.
- Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridge sampling: An R package for estimating normalizing constants. *arXiv Preprint arXiv:1710.08162*.
- Guest, O., & Martin, A. E. (2020). How computational modeling can force theory building in psychological science. *PsyArXiv preprint*. <https://doi.org/10.31234/osf.io/rybh9>.
- Hanne, S., Burchert, F., De Bleser, R., & Vasishth, S. (2015). Sentence comprehension and morphological cues in aphasia: What eye-tracking reveals about integration and prediction. *Journal of Neurolinguistics*, *34*, 83–111.
- Hanne, S., 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*(2), 221–244.
- Heathcote, A., & Love, J. (2012). Linear deterministic accumulator models of simple choice. *Frontiers in Psychology*, *3*, 292.
- 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.

- Jäger, L. A., Merten, D., Van Dyke, J. A., & Vasishth, S. (2020). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*, *111*, 104063.
- Jeffreys, H. (1939/1998). *The theory of probability*. Oxford, UK: Oxford University Press.
- Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.
- Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*. Cambridge, UK: Cambridge University Press.
- Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.
- Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.
- Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447–454.
- Love, T., Swinney, D., Walenski, M., & Zurif, E. (2008). How left inferior frontal cortex participates in syntactic processing: Evidence from aphasia. *Brain and Language*, *107*(3), 203–219.
- 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.
- Martin, A. E., & McElree, B. (2008). A content-addressable pointer mechanism underlies comprehension of verb-phrase ellipsis. *Journal of Memory and Language*, *58*(3), 879–906.
- Martin, A. E., & McElree, B. (2011). Direct-access retrieval during sentence comprehension: Evidence from sluicing. *Journal of Memory and Language*, *64*(4), 327–343.
- 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.
- McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, *32*(4), 536–571.
- McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.
- McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*(1), 67–91.
- Mirman, D., Yee, E., Blumstein, S. E., & Magnuson, J. S. (2011). Theories of spoken word recognition deficits in aphasia: Evidence from eye-tracking and computational modeling. *Brain and Language*, *117*(2), 53–68.
- Nicenboim, B., & Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas – Part II. *Language and Linguistics Compass*, *10*(11), 591–613.
- Nicenboim, B., & Vasishth, S. (2018). Models of retrieval in sentence comprehension: A computational evaluation using Bayesian hierarchical modeling. *Journal of Memory and Language*, *99*, 1–34.
- Nicenboim, B., Vasishth, S., Engelmann, F., & Suckow, K. (2018). Exploratory and confirmatory analyses in sentence processing: A case study of number interference in German. *Cognitive Science*, *42*, 1075–1100.
- Patil, U., 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.
- Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago, IL: University of Chicago Press.
- Pregla, D., Lissón, P., Vasishth, S., Burchert, F., & Stadie, N. (2020). Variability in sentence comprehension in aphasia in German. *PsyArXiv preprint*. <https://doi.org/10.31234/osf.io/7hfpX>
- R Core Team. (2020). *R: A language and environment for statistical computing* (version 4.0.2). Vienna, Austria: R Foundation for Statistical Computing.
- Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

- Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time? *Psychometrika*, 70(2), 377–381.
- Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian inference for psychology, part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1), 102–113.
- Rouder, J. N., Province, J. M., Morey, R. D., Gomez, P., & Heathcote, A. (2015). The lognormal race: A cognitive-process model of choice and latency with desirable psychometric properties. *Psychometrika*, 80(2), 491–513.
- Schad, D. J., Betancourt, M., & Vasishth, S. (2020). Toward a principled Bayesian workflow in cognitive science [Accepted for publication in Psychological Methods]. *arXiv Preprint* arXiv:1904.12765.
- Schad, D. J., Vasishth, S., Hohenstein, S., & Kliegl, R. (2020). How to capitalize on a priori contrasts in linear (mixed) models: A tutorial. *Journal of Memory and Language*, 110, 104038.
- Sivula, T., Magnusson, M., & Vehtari, A. (2020). Unbiased estimator for the variance of the leave-one-out cross-validation estimator for a Bayesian normal model with fixed variance. *arXiv Preprint* arXiv:2008.10859.
- Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *The Quantitative Methods for Psychology*, 12(3), 175–200.
- Stan Development Team. (2020). RStan: The R interface to Stan (Version 2.21.2).
- Traxler, M. J., Williams, R. S., Blozis, S. A., & Morris, R. K. (2005). Working memory, animacy, and verb class in the processing of relative clauses. *Journal of Memory and Language*, 53(2), 204–224.
- Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, 108(3), 550–592.
- Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalysed ambiguities. *Journal of Memory and Language*, 49(3), 285–316.
- Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, 55(2), 157–166.
- Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, 65(3), 247–263.
- Vasishth, S., Nicenboim, B., Beckman, M. E., Li, F., & Kong, E. J. (2018). Bayesian data analysis in the phonetic sciences: A tutorial introduction. *Journal of Phonetics*, 71, 141–161.
- Vasishth, S., Nicenboim, B., Engelmann, F., & Burchert, F. (2019). Computational models of retrieval processes in sentence processing. *Trends in Cognitive Sciences*, 23, 968–982.
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432.
- Walker, G. M., Hickok, G., & Fridriksson, J. (2018). A cognitive psychometric model for assessment of picture naming abilities in aphasia. *Psychological Assessment*, 30(6), 809–826.