Submitted to Trends in Cognitive Sciences

Computational models of retrieval processes in sentence processing

Shravan Vasishth
Department of Linguistics, University of Potsdam, Potsdam, Germany
Bruno Nicenboim
Department of Linguistics, University of Potsdam, Potsdam, Germany
Felix Engelmann
Manchester University, UK
Frank Burchert
Department of Linguistics, University of Potsdam, Potsdam, Germany

August 31, 2019

Abstract

Sentence comprehension requires that the comprehender work out who did what to whom. This process has been characterized as retrieval from memory. This review summarizes the quantitative predictions and empirical coverage of the two existing computational models of retrieval, and shows how the predictive performance of these two competing models can be tested against a benchmark data-set. We also show how computational modeling can help us better understand sources of variability in both unimpaired and impaired sentence comprehension.

Keywords: cue-based retrieval; retrieval interference; comprehension impairments in aphasia; individual differences

Cue-based retrieval and interference in sentence processing

Comprehending a sentence involves an array of cognitive processes, including lexical access of words from memory, incremental structure building, and computing the meaning of the sentence. One important aspect of sentence comprehension is working out who did what to whom. As an example, consider the following sentence:

"The worker was surprised that the resident who was living near the dangerous neighbor was complaining about the investigation."

Correspondence: vasishth@uni-potsdam.de

To fully understand this sentence, the reader must work out who was surprised, what they were surprised about, who was doing the complaining, and what they were complaining about. In sentence processing research, working out these connections between the words is often termed dependency completion. Informally speaking, dependency completion can be thought of as building up an interpretation of the sentence by linking together words or phrases that belong together linguistically.

To interpret the above sentence, among the long-distance dependencies that must be completed is the one involving the verb phrase "was complaining" and the animate noun phrase "the resident", the grammatical subject of the clause that the verb phrase appears in. Over the years, several theories have been developed that specify the process that leads to such dependency completion [1, 2, 3, 4, 5, 6, 7, 8].

Based on linguistic theory [9] and research on memory [10], these theories begin with the assumption that words and phrases are maintained in memory as feature-value bundles. A noun phrase like the grammatical subject "the resident" is assumed to be represented as a bundle of feature-value pairs such as [nominal: yes, subject: yes, singular: yes, animate: yes]. These theories of dependency completion assume that reading the verb phrase triggers a search in memory for a noun that has certain properties or features. For example, the subject of the verb phrase would be searched for using a feature-bundle like [nominal: yes, subject: yes, singular: yes, animate: yes]. This bundle of features, used for searching for a co-dependent item in memory, is referred to collectively as retrieval cues.

A large body of empirical research involving reading studies [5, 8, 11, 12, 13, 14, 15] has demonstrated that when a search is carried out in memory using a set of retrieval cues, increased processing difficulty is observed when multiple nouns have features that match the retrieval cues. This increased processing difficulty is often referred to as retrieval interference [13]. Because of the observed slowdown, it is sometimes also called inhibitory interference [16]. The observed slowdown is always with reference to a baseline condition. This becomes clear when we consider the following pair of sentences:

(1) a. The worker was surprised that the resident$^{+animate}_{+subject}$ who was living near the dangerous neighbor$^{+animate}_{-subject}$ was complaining$\{^{animate}_{subject}\}$ about the investigation.

b. The worker was surprised that the resident$^{+animate}_{+subject}$ who was living near the dangerous warehouse$^{-animate}_{-subject}$ was complaining$\{^{animate}_{subject}\}$ about the investigation.

In the examples above, the feature specifications on the nouns are shown, along with the retrieval cues on the verb phrase "was complaining." In (1a), there are two animate nouns ("resident" and "neighbor") in one phrase, whereas in (1b) there is only one ("resident"). If an animate noun is searched for at the verb, greater difficulty will be experienced in (1a) compared to (1b), because the animacy feature matches two different nouns in (1a), making the nouns difficult to distinguish. This proposal can be made mathematically precise in computationally implemented models [17, 18, 19], and using these models quantitative predictions can derived that can then be tested against data.

This review covers recent developments relating to two computationally implemented models of retrieval processes which aim to explain interference effects: the activation model [17], and the direct-access model [3, 18, 20]. We focus on computationally implemented models because they make quantitative predictions regarding the empirical phenomena of

interest. As we discuss below, quantitative models are valuable because, among other things, they allow us to investigate an important open issue in sentence processing: the sources of individual-level variability in comprehension difficulty in unimpaired and special populations, such as individuals with aphasia.

Two computational models of retrieval

The activation model

The activation model is inspired in part by the push in cognitive psychology and artificial intelligence research to ground cognitive processes in a general theory of information processing [10, 21]. The model is implemented within the general cognitive architecture ACT-R (http://act-r.psy.cmu.edu/), which assumes that all cognitive processes, including language processing, are driven by a common set of constraints on memory and learning.

The activation model's explanation for inhibitory interference is illustrated schematically in the upper half of Figure 1, and refers to example (1).

There are three high-level assumptions in the model, which come from the ACT-R architecture. First, every item in memory undergoes exponential decay in its activation over time. Second, an item in memory with the highest activation will be retrieved, and the higher the activation, the faster and more accurate the retrieval. Third, activation has a Gaussian noise component; this affects which item is retrieval from trial to trial. Given these general assumptions about activation, inhibitory interference arises from some further assumptions about how cue-based retrieval works. If an item in memory has features that match the retrieval cues perfectly, that item will get an activation boost. If another item in memory has features that match a subset of the retrieval cues, the total activation on both the perfectly matching item and the partly matching item will be dampened. This situation is referred to as cue overload [22]: the same retrieval cue is identifying two different items, making it difficult to distinguish between them. This noise component leads to non-deterministic behavior from one trial to the next: sometimes the correct item (which matches the retrieval cue perfectly) will be retrieved, and sometimes the incorrect item (which matches only a subset of the retrieval cues) will be retrieved. Due to the overall dampened activation that occurs when (as in example 1a) multiple items in memory match retrieval cues, mean retrieval time in (1a) is slower compared to (1b), because in the latter case only one item in memory matches the retrieval cues.

The activation model also predicts speedups in processing under certain specific conditions. As an example, consider the sentences below, which have been investigated extensively in reading studies [16, 23, 24, 25]. Also see the lower part of Figure 1 for a schematic illustration.

(2) a. *The bodybuilder$^{-plural}_{+subject}$ who met the trainers$^{+plural}_{-subject}$ were$\{^{plural}_{subject}\}$ ...

b. *The bodybuilder$^{-plural}_{+subject}$ who met the trainer$^{-plural}_{-subject}$ were$\{^{plural}_{subject}\}$ ...

In the examples above, we show the feature specifications on the nouns, as well as the retrieval cues on the auxiliary verb "were". Both (2a) and (2b) are ungrammatical (marked with an asterisk following linguistic convention) because the auxiliary verb "were" initiates a search for a plural-marked subject noun phrase, but the subject, "bodybuilder", is singular

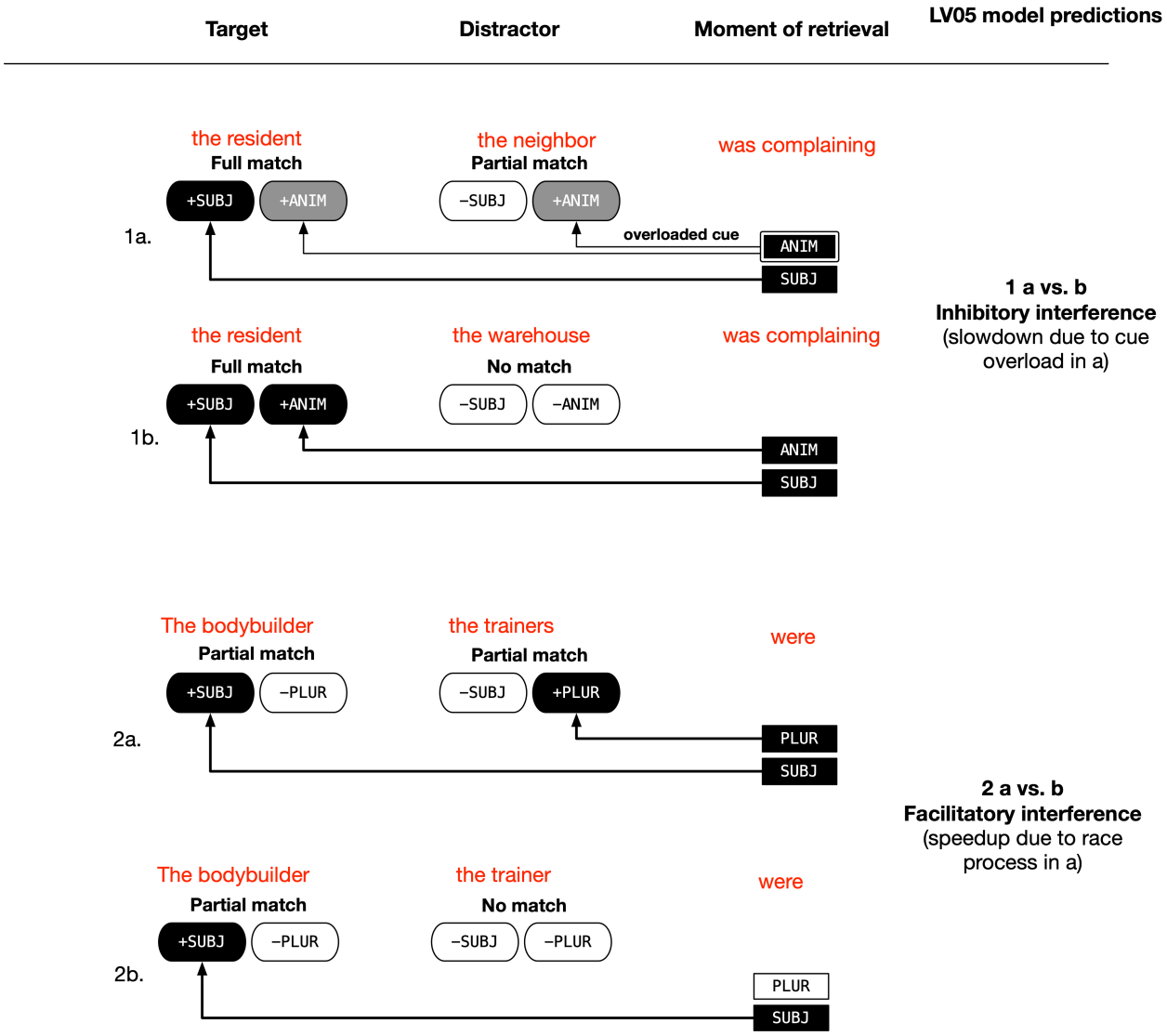

Figure 1. **A schematic illustration of the activation model**. The model predicts that grammatical conditions like (1a) show inhibitory interference effects (slowdowns), and ungrammatical conditions like (2a) show facilitatory interference effects (speedups). The figure is by Engelmann and Vasishth, 2019; available at http://dx.doi.org/10.6084/m9.figshare.9305456 under a CC-BY4.0 license.

marked, so it matches only one retrieval cue (the subject cue). As in example (2a), if another plural marked noun (here, the noun "trainers") matches the other retrieval cue (plural), then faster retrieval time is observed at the auxiliary verb, compared to a baseline condition (2b) where the noun "trainer" has singular marking and therefore doesn't match either of the two retrieval cues. Such speedups have been observed in constructions other than the subject-verb number agreement example above [26].

In the activation model, the underlying cause for the observed speedup is a so-called race process [27]: one process is an attempt to access one item in memory, and the other process is an attempt to access the other item; both these processes unfold in parallel. Whichever finishes first leads to the respective item being retrieved and the search terminating. This is called a parallel self-terminating search [28, 29], in contrast to exhaustive parallel processing, where both (or all) processes must complete [30]. The race process has the effect that on average, a faster reading time is observed compared to a baseline condition where no race occurs. In the above example, the auxiliary verb "were" in (2a) is read faster than in (2b). The reading time distribution that results from a race process is illustrated in Figure 2.

The race process arises within the activation model arises as follows. The assumption in ACT-R is that if a subset of the retrieval cues matches the features on an item in memory, the probability of the item being retrieved increases compared to the case where no retrieval cue matches the item's features. Now, if the features on one item match only some subset of the retrieval cues and the features on a second item match a different subset of retrieval cues, the model predicts a speedup in retrieval time (compared to a baseline condition where only one item matches a subset of the retrieval cues). This speedup is sometimes referred to as a facilitatory interference effect or statistical facilitation [27]. The term facilitatory is not intended to imply that processing is easier, but refers rather to the observed speedup.

The activation model is available in several computational implementations. A simplified version written in the statistical computing language R [31] is available from http://tiny.cc/inter-act, and the full parsing model, written in Lisp and ACT-R, is available from http://tiny.cc/actr-sentence-parser. An implementation in Stan [32] is available from http://tiny.cc/act_model. The activation model has been used to investigate a broad range of phenomena: retrieval interference effects [16, 18, 19, 33, 34, 35, 36, 37, 38, 39, 40, 41]; the relative roles of predictive processing vs. interference effects [42]; impairments in individuals with aphasia [43, 44]; the interaction between oculomotor control and sentence comprehension [45, 46]; and the effect of working memory capacity differences on underspecification (good-enough processing [47]) in sentence comprehension [48]. The activation model has served as a baseline for several new emerging computational models [49, 50, 51, 52].

The direct-access model

The direct-access model was motivated by research from the cognitive psychology literature [3, 11, 53], which shows that accessing items from memory is driven by a content-addressable memory system. That is, as in the activation model, retrieval cues (bundles of feature-value pairs) are used for carry out a search. Sentence processing is assumed to be constrained by the same general memory system that constrains other types of information processing [54].

The term direct-access refers to the assumption that search in memory is driven by directly accessing items in memory that match the features used as retrieval cues. The

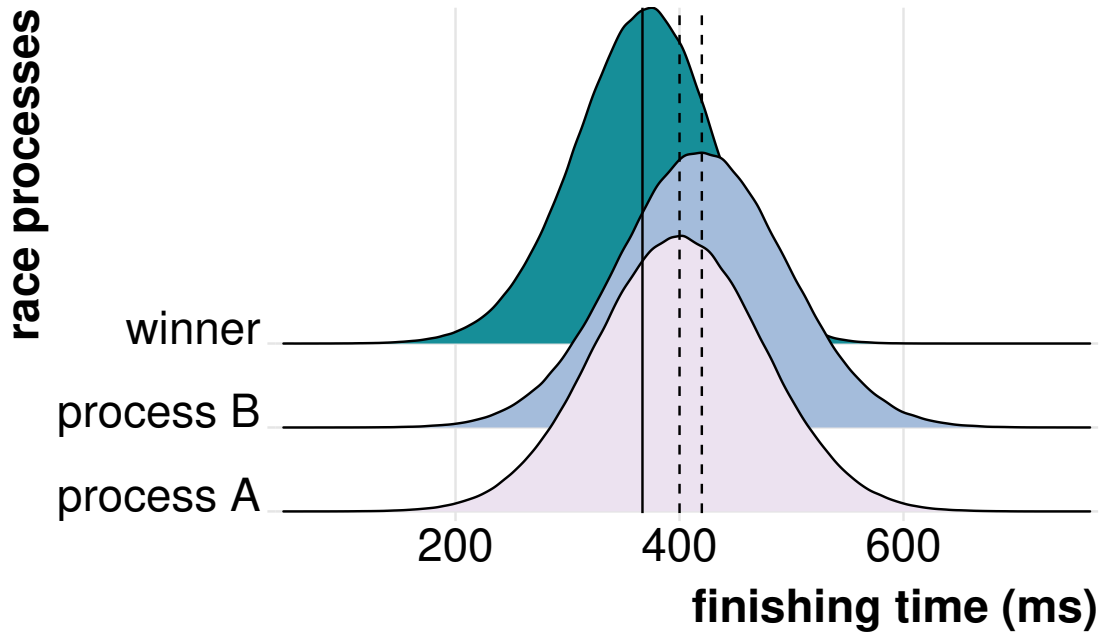

Figure 2. **An illustration of the race process**. Here, we show the resulting reading time distribution that arises from a race between two processes (e.g., one process is an attempt to retrieve one item based on a set of retrieval cues, and the other process, triggered simultaneously, is an attempt to retrieve a different item). The mean of the winning finishing time distribution (the solid vertical line) is smaller than the means of the two processes (broken vertical lines) that are engaged in the race: the winning distribution represents the observed statistical facilitation.

time taken to complete a retrieval is assumed to be constant regardless of when the item was previously encountered. This type of search process is often referred to as a content-addressable cue-based search. The term content-addressable refers to the use of retrieval cues (bundles of feature-value specifications such as [subject: yes, animate: yes]) to search for items in memory using their feature specifications.

The memory-access process assumed in the computationally implemented version of the direct-access model [18] is perhaps most easily understood if we consider how inhibitory interference effects arise in the model. We use example (1) to explain model assumptions. As shown schematically in Figure 3, for sentence (1a) the model assumes that when the retrieval cues [subject: yes, animate: yes] are used to access the subject noun in memory, these cues match the noun "resident", but they also partially match the noun "neighbor" on one feature (animate). This leads to a cue overload as in the activation model. The consequence of this cue overload is that in most trials the subject "resident" will be retrieved, but in some proportion of trials the incorrect noun "neighbor" will be misretrieved. In both cases the time taken to complete the retrieval is the same, say β milliseconds. In trials where the incorrect noun is retrieved, in some proportion of the cases a second retrieval attempt (referred to as reanalysis) is carried out which costs a certain amount of time, say δ ms. In contrast to (1a), in (1b) when the retrieval cues [subject: yes, animate: yes] are used to

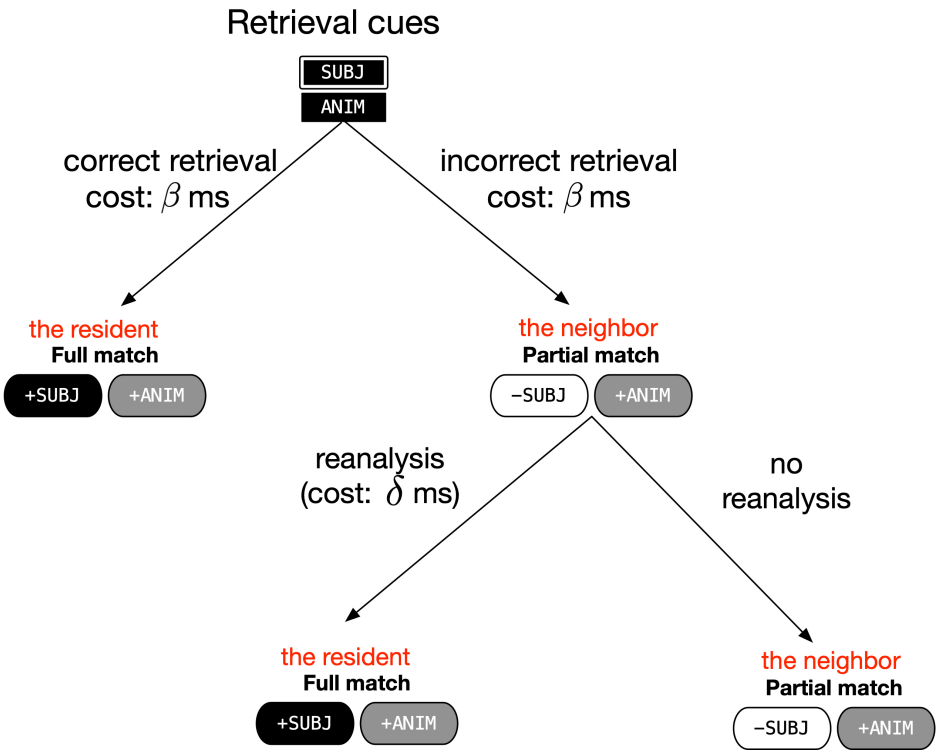

Figure 3. **A schematic illustration of the direct-access model**. For sentences like (1a), the model assumes that once a search is initiated in memory using a set of retrieval cues (here, subject and animate), one of two events can happen. Either the correct item is retrieved from memory, or the incorrect item, which matches some of the retrieval cues, is misretrieved. In the case of a misretrieval, either processing ends with a misretrieval, or a reanalysis step is initiated that leads to a correct retrieval. This reanalysis step costs time, and therefore leads to slowdowns in processing on average. The figure is by Vasishth, 2019; it is available from http://10.6084/m9.figshare.9396515 under a CC-BY4.0 license.

access the subject noun in memory, only one noun ("resident") matches these cues and most of the retrievals succeed immediately. Thus, in (1a) the probability of a misretrieval followed by a reanalysis step is higher than in (1a), and since reanalysis costs δ ms, sentence (1a) takes longer to read than (1b).

An important difference from the activation model is that cue overload affects only the probability of retrieving an item from memory; the retrieval time per se is constant. In the direct-access model, the increased reading time observed due to cue overload is a consequence of the reanalysis time. By contrast, in the activation model, cue overload redistributes the activation of items, dampening activation for all items that (partly) match the retrieval cues. Since activation affects retrieval accuracy as well as retrieval time, the direct consequence of cue overload is increased retrieval time.

Formally, the direct-access model can therefore be seen as a two-component finite mixture process [56, 57], with some proportion of trials representing a successful retrieval in the first-attempt, and some proportion representing a slower retrieval that is the consequence

of an initially unsuccessful retrieval in the first attempt followed by a subsequent reanalysis step [18]. A formalization in terms of a mixture process is shown in Box 1.

All the published work relating to the direct-access model has focused on inhibitory interference effects [55]. What are the model's predictions regarding facilitatory interference effects? The model assumes that slower reading times occur due to a reanalysis step that results in the correct item being retrieved. However, in ungrammatical sentences, there is no correct item to retrieve. Thus, the model is underspecifcied regarding the processing steps taken in this situation [18]. It is therefore likely that additional assumptions will be needed to account for the speedups discussed in connection with the activation model. This is a potentially interesting topic for future research.

A computational implementation of the direct-access model in Stan [32] is available from http://tiny.cc/da_model.

Comparing the predictions of the direct-access model and the activation model

Since the two computational models discussed above assume different underlying processes which both lead to the same observed behavioral outcome (inhibitory interference), one obvious question arises: which model furnishes a better explanation for inhibitory interference effects? A model comparison was carried out using a relatively large-sample (182 participants) data-set that showed inhibitory interference [41]. The predictive performance of the two models was evaluated using k-fold cross validation [58, 59]. The direct-access model exhibited a better predictive performance compared to the activation model. The reason that the direct-access model outperformed the activation model is that the latter predicts that in inhibitory interference experimental designs, retrievals of the incorrect chunk should be slower than the retrievals of the correct chunk. In the data-set used for model comparison [41], incorrect retrievals had faster reading time than correct retrievals. Box 1 explains this model comparison in more detail.

Modeling and understanding impairments in sentence comprehension

The computational models of retrieval described above are designed to explain sentence comprehension difficulty in unimpaired adult native-speaker populations. If these models represent unimpaired cognitive processes, an obvious question arises: how would impairment (e.g., in individuals with aphasia) arise in these models? In other words, the underlying theoretical constructs in the model should also be able to explain comprehension deficits in impaired populations. If particular constructs within a sentence processing theory can be shown to be related to specific deficits in comprehension, this increases the plausibility of those theoretical constructs. Using a computational model of unimpaired processing to account for deficits in sentence processing has a further advantage: it can also yield new insights into what the underlying causes of the observed deficits might be.

Sentence comprehension difficulties in aphasia are an important example of impairments. Aphasia is an acquired language disorder caused by a brain damage that can affect speech production and comprehension to varying degrees. Several theories have been proposed to explain the nature and sources of comprehension deficits in individuals with aphasia [60].

Attempts have been made to account for processing deficits in aphasia by varying specific parameters of the activation model [43, 44]. This work shows that the behavior of each individual with aphasia can be characterized as being affected to varying degrees by three sources of deficits that have been independently proposed in the literature on aphasia. These are intermittent deficiency, resource reduction, and slowed processing.

Intermittent deficiency [60] is the proposal that there are occasional breakdowns in parsing. The verbal statement of this proposal leaves it underspecified what exactly constitutes a breakdown, but the idea can be transformed into a concrete and quantitatively testable proposal by implementing it within a computational model. Within the activation model, intermittent processing breakdown can be implemented as arising from increased noise in activation; the higher the noise, the greater the fluctuation in activation and the greater the amount of processing difficulty on average, and the higher the probability of parsing failure. This is of course one of several different ways that intermittent deficiency can be implemented.

A second proposed explanation for deficits in comprehension in aphasia is resource reduction [61]; this is the proposal that individuals with aphasia have fewer cognitive resources available compared to unimpaired controls. One way to implement this idea within the activation model is by reducing the total amount of activation (called goal activation in the model) that can be allocated to an item that matches a particular set of retrieval cues—the higher the allocation of the activation, the greater the amount of resources available to the comprehender.

The third proposal is slowed processing [62]; the assumption here is that individuals with aphasia have a slower processing system than unimpaired controls. This proposal can be implemented within the activation model as an increase in the default processing time for each parsing step in the activation model. The activation model assumes by default that each parse step takes 50 ms; this assumption comes from ACT-R. However, this default processing time is a parameter in the model and can be varied at the individual level; larger values will lead to slowed parsing steps that will cause reductions in activation (due to the assumption of activation decay over time in ACT-R), which leads to greater processing difficulty.

These three verbally stated theories of processing deficits in aphasia have been quantitatively investigated using the activation model [43], although that work relied on data from a relatively small-sample study (seven individuals with aphasia, hereafter referred to as IWAs). A larger-sample study [44] implemented these three proposals using self-paced listening data on subjects vs. object relative clauses from 56 IWAs and 46 matched controls [60]. For each participant separately, the numerical parameters in the activation model corresponding to each of the above theoretical proposals were estimated. In other words, for each participant, the noise parameter, the goal activation parameter, and the parsing speed parameter were estimated.

Figure 4 shows the distribution of the best parameter estimates for each participant in subject vs. object relatives. The plots show that compared to controls, IWAs tend to have lower goal activations, higher noise, and slower default action time, suggesting that IWAs experience these deficits to a greater extent than controls, but each of the deficits is present in each IWA to differing degrees. The fact that some controls also show parameter estimates similar to IWAs suggests that the deficits along these three dimensions may all in

principle be candidate theories for characterizing processing difficulty not just in aphasia but also in unimpaired processing.

Some other important future directions in modeling comprehension impairments in aphasia are the following: (a) Expand the cross-linguistic empirical base of the investigation. Previous work [44] has relied on the available data from English, but large-sample data-sets from other languages are needed: the diversity of grammatical constraints, word orders, and morphosyntactic cues available in different languages are known to lead to different processing strategies even in unimpaired populations [63, 64, 65, 66], and such factors may play a role in aphasia as well [67, 68]. (b) Evaluate the relative predictive performance of different (retrieval) theories using data from individuals with aphasia and controls. This kind of model comparison is also needed for other recently developed models [49, 50, 51, 52] that are intended to serve as alternatives to the activation and direct-access models.

Modeling individual differences

It is standard practice in psycholinguistics to focus on the average effect. We usually ask questions like: is the effect of interest present or absent? This is also how most of the research on interference has been framed. If our conclusion from the data is "effect present", the claim is that participants on average show the effect, regardless of the estimated magnitude of the effect and the uncertainty of the effect estimate. If our conclusion is "no effect present", then the claim is that participants on average do not show the effect, again ignoring the magnitude and uncertainty of the effect estimate. However, even ignoring the statistical problems with this reasoning [69], theoretically important individual-level differences can lie behind both significant and non-significant effects. This point has been repeatedly emphasized in the literature [70, 71] but has not been widely appreciated.

Early models in sentence processing had in fact paid attention to the question of individual differences. One influential proposal [1] was that individual-level differences in domain-general working memory capacity were driving the observed differences in individual-level behavior in sentence comprehension. Thus, individuals were assumed to have different amounts of capacity for holding or processing information in memory. A persistent problem in this line of research has been how to reliably measure individual differences in capacity among participants. The sentence span task is a commonly used method in sentence processing for measuring capacity differences [72], but it is not clear whether this measure indexes capacity differences [73, 74], and what it relates to in behavioral measures of processing difficulty [75, 76, 77]. Operation span has been argued to be a reliable measure of working memory capacity that is not correlated with reading speed or experience [78, 79], but has yielded rather mixed results in sentence processing research [35, 41, 80, 81], again making it difficult to define any clear link between an independent measure of capacity and model-internal mechanisms. Despite these concerns, there is work suggesting that domain-general working memory capacity differences may play a role in determining individual-level comprehension difficulty [77, 82]. However, this evidence for a capacity-based view only reports statistical associations between behavioral measures and working memory measures which may or may not replicate in confirmatory analyses, especially given the fact that many psycholinguistic studies tend to be underpowered [15, 41, 83, 84, 85].

An interesting alternative to the capacity proposal is that individual differences in sentence comprehension arise not from capacity differences per se, but from differences in

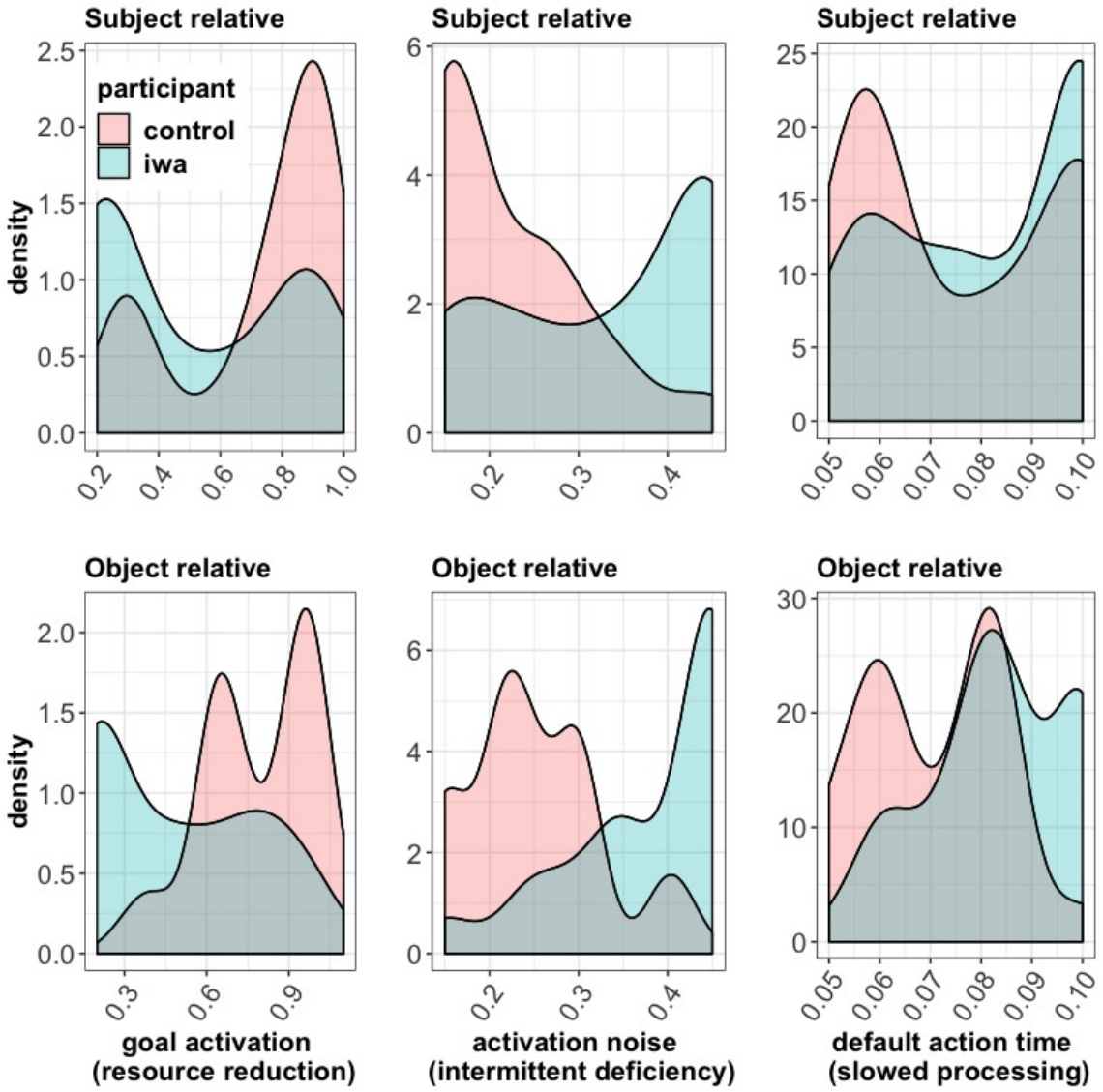

Figure 4. **Distribution of model parameters in the activation model**. Shown are the distributions of the best parameter estimates in the activation model for each individual participant (individuals with aphasia or controls).

the content and quality of mental representations [13, 14]. This proposal can be investigated within the computationally implemented models of retrieval discussed above: differences in individual-level behavior may map onto values of specific numerical parameters. As an example, consider the subject-verb dependencies shown earlier in examples (2a) and (2b). This is the configuration which shows facilitatory interference effects on average. Figure 5 shows the activation model's predicted range of effects, and the estimate of the observed interference effect (95% confidence interval). Also shown are the estimated participant-level interference effects (in milliseconds).

Some interesting observations emerge. First, we see a consistent facilitatory interference effect across participants; however, the magnitudes of the effects are larger in some participants and approach zero in others. At first sight, one might think that this is a trivial consequence of a correlation between individual-level mean reading times and the individual-level interference effect: it can happen that the slower a participant, the larger the effect. If differences in mean reading time among individuals were the reason for the variation in the interference effect across individuals, in the linear mixed model fit to the data the correlation between the intercept and slope adjustments by subject would be positive. However, there is no evidence that the intercept and slope adjustments are correlated: the estimated correlation is -0.14, with 95% credible interval [-0.64,0.44]. In a larger-sample replication attempt of this experiment [85], which had 181 participants instead of the original sample size of 40, the estimated correlation is -0.24, 95% credible interval [-0.71, 0.38]. Of course, this absence of a clear correlation does not imply there is none; it just means that we could not find support for the correlation explanation for the varying magnitudes of the facilitatory interference effect.

A possible alternative explanation for the differing magnitudes across participants is that, within the activation model's framework, different participants could have different weights for each cue [34, 55, 89, 90]. This possibility holds equally for both the activation and direct-access models, because both models assume that retrieval cues are used to carry out a search in memory. Some participants may weight the subject cue in (2) higher than the number cue; these participants can efficiently target the subject without experiencing interference from the distractor noun. Others show large facilitatory patterns, which are consistent with equal cue-weighting. The activation model's cue-weighting parameter can explain this variation (Box 2). The variation in cue-weighting between individuals could in principle be independently motivated by measuring, for example, the linguistic proficiency (acquired through varying degrees of exposure) of the comprehender [73]. One important open problem here may be obtaining reliable measures of proficiency [91].

Thus, a rich open area for research is using computational models to explain patterns of individual differences [92]. Such a differentiated, individual-differences account cannot be developed if we focus only on average effects. The average is just an abstraction that leads to generalizations masking theoretically important variation at the individual level [93]. In future work, much theoretical insight can be gained by investigating individual-level effects in the context of quantitative model predictions.

Concluding remarks

In closing, although it is clear that retrieval processes are not the only factor that determine processing difficulty in sentence comprehension [71, 94, 95, 96, 97, 98, 99, 100], there

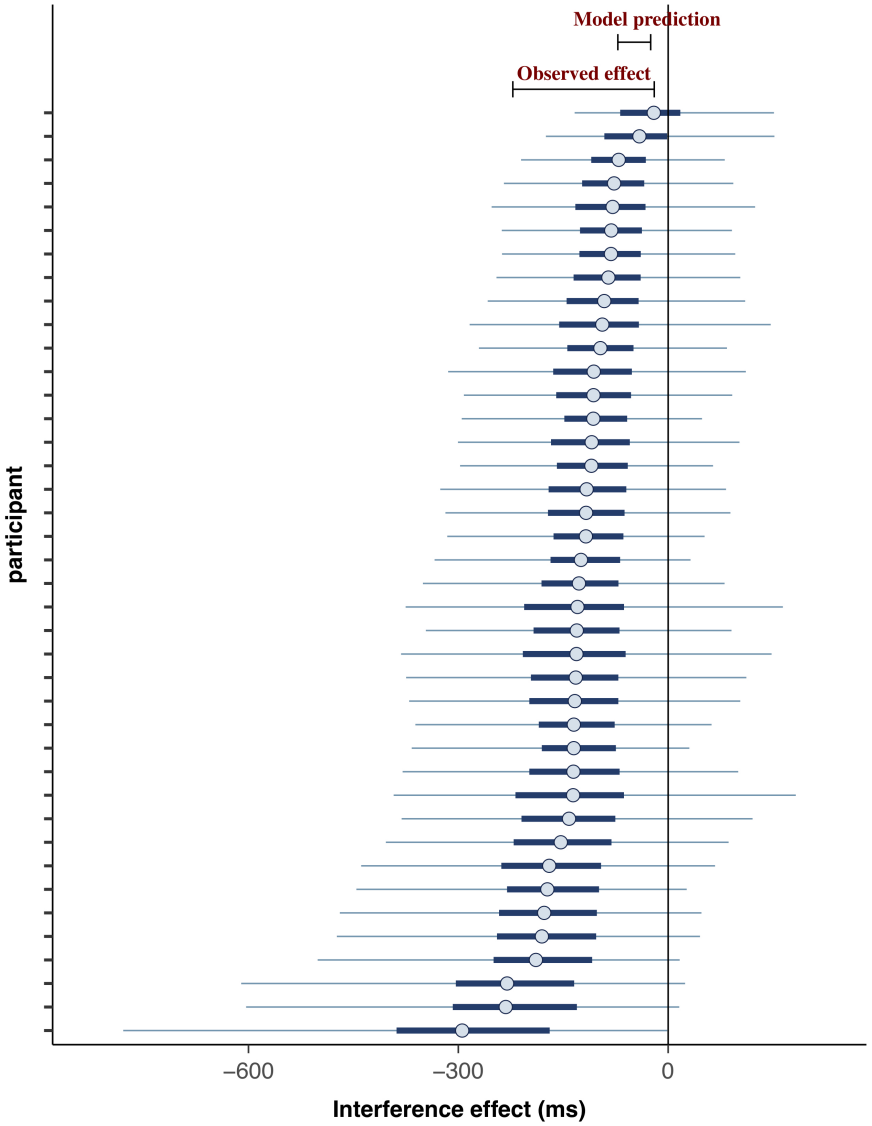

Figure 5. **Dillon Expt 1: Ungrammatical agreement by participants**. Shown are the individual-level facilitatory interference effects (circles show means, the thick lines 80% credible intervals, and the thin lines 95% credible intervals) in subject-verb agreement configurations; the data are from a recent study [16]. The participant-level estimates of the interference effect were computed from a hierarchical Bayesian model [86]. The average facilitatory interference effect is shown as a 95% credible interval (labeled Observed effect), and the range of predicted values from the activation model are also shown as a 95% credible interval (labeled Model prediction); the model predictions are computed using Approximate Bayesian Computation [85, 87, 88]. We see the predicted facilitatory interference effect even at the individual participant level, but different participants show varying magnitudes. The activation model can account for this variation in effect magnitude as a function of cue-weighting; see Box 2.

is considerable evidence supporting a role for retrieval in completing linguistic dependencies. The relative performance of the two existing computational models of retrieval has been evaluated against only one data-set; future work should carry out a more extensive evaluation using publicly available data-sets on interference effects. One possible outcome of such an evaluation could be that a hybrid of the two models shows the best performance; such a unification of the two models could significantly advance our theoretical understanding of retrieval processes.

This review covered the main recent empirical and theoretical developments in this research area, and discussed two important, related avenues that computational modeling opens up for future research: modeling impairments in sentence comprehension, and modeling individual differences.

Acknowledgements

The research reported here was partly funded by the Volkswagen Foundation through grant 89 953; and the Deutsche Forschungsgemeinschaft (German Science Foundation), Collaborative Research Center - SFB 1287, project number 317633480 (*Limits of Variability in Language*) through project B2 (PIs: Shravan Vasishth, Frank Burchert and Nicole Stadie) and B3 (PIs: Ralf Engbert and Shravan Vasishth). We are grateful to Garrett Smith, Dorothea Pregla, Dario Paape, Paula Lissón, Pavel Logačev, Sol Lago, and Daniela Mertzen for comments on earlier drafts of this paper. Thanks go to Brian Dillon for generously releasing his published data.

References

[1] Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*(1), 122–149.

[2] Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*(1), 1–76.

[3] McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, *29*(2), 111–123.

[4] Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain* (Chap. 5, pp. 95–126). Cambridge, MA: MIT Press.

[5] 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 misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–316.

[6] Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentential complexity. *Cognitive Science*, *29*, 261–291.

[7] 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.

[8] Van Dyke, J. A., & McElree, B. (2011). Cue-dependent interference in comprehension. *Journal of Memory and Language*, *65*(3), 247–263.

[9] Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Cambridge Press.

[10] 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), 1036–60.

[11] Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*(2), 157–166.

[12] Van Dyke, J. A. (2007). Interference effects from grammatically unavailable constituents during sentence processing. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *33*(2), 407–430.

[13] Van Dyke, J. A., & Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, *6*(4), 193–211.

[14] Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*(3), 373–403.

[15] 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.

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

[17] Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419.

[18] 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.

[19] Engelmann, F., Jäger, L. A., & Vasishth, S. (2019). The effect of prominence and cue association in retrieval processes: A computational account. *Cognitive Science*. Accepted pending minor revisions.

[20] McElree, B. (2003). Accessing recent events. *Psychology of Learning and Motivation*, *46*, 155–200.

[21] Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283–308). New York: Academic Press.

[22] Watkins, O. C., & Watkins, M. J. (1975). Buildup of proactive inhibition as a cue-overload effect. *Journal of Experimental Psychology: Human Learning and Memory*, *104*(4), 442–452.

[23] Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237.

[24] Lago, S., Shalom, D. E., Sigman, M., Lau, E. F., & Phillips, C. (2015). Agreement processes in Spanish comprehension. *Journal of Memory and Language*, *82*, 133–149.

[25] Tucker, M. A., Idrissi, A., & Almeida, D. (2015). Representing number in the real-time processing of agreement: Self-paced reading evidence from Arabic. *Frontiers in Psychology*, *6*(347).

[26] Cunnings, I., & Sturt, P. (2018). Retrieval interference and sentence interpretation. *Journal of Memory and Language*, *102*, 16–27.

[27] Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*(5 Series II), 574–590.

[28] Colonius, H., & Vorberg, D. (1994). Distribution inequalities for parallel models with unlimited capacity. *Journal of Mathematical Psychology*, *38*(1), 35–58.

[29] Logačev, P., & Vasishth, S. (2016a). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266–298.

[30] Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. CUP Archive.

[31] R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

[32] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

[33] Vasishth, S., Bruessow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*(4).

[34] Parker, D., & Phillips, C. (2017). Reflexive attraction in comprehension is selective. *Journal of Memory and Language*, *94*, 272–290.

[35] Nicenboim, B., Logačev, P., Gattei, C., & Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory differences in unbounded dependencies. *Frontiers in Psychology*, *7*, 280.

[36] Kush, D., & Phillips, C. (2014). Local anaphor licensing in an SOV language: Implications for retrieval strategies. *Frontiers in Psychology*, *5*(1252).

[37] Patil, U., Vasishth, S., & Lewis, R. L. (2016a). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*, *7*(329).

[38] Parker, D., & Phillips, C. (2016). Negative polarity illusions and the format of hierarchical encodings in memory. *Cognition*, *157*, 321–339.

[39] Jäger, L. A., Engelmann, F., & Vasishth, S. (2015). Retrieval interference in reflexive processing: Experimental evidence from Mandarin, and computational modeling. *Frontiers in Psychology*, *6*(617).

[40] Patil, U., Vasishth, S., & Lewis, R. L. (2016b). Retrieval interference in syntactic processing: The case of reflexive binding in English. *Frontiers in Psychology*, *7*, 329.

[41] 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.

[42] Boston, M., Hale, J., Vasishth, S., & Kliegl, R. (2011). Parallel processing and sentence comprehension difficulty. *Language and Cognitive Processes*, *26*(3), 301–349.

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

[44] 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. Allen Newell Best Student-Led Paper Award at MathPsych/ICCM 2017.

[45] Engelmann, F., Vasishth, S., Engbert, R., & Kliegl, R. (2013). A framework for modeling the interaction of syntactic processing and eye movement control. *Topics in Cognitive Science*, *5*(3), 452–474.

[46] Dotlačil, J. (2018). Building an ACT-R reader for eye-tracking corpus data. *Topics in Cognitive Science*, *10*(1), 144–160.

[47] Ferreira, F., Ferraro, V., & Bailey, K. G. D. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, *11*, 11–15.

[48] Engelmann, F. (2016). *Toward an integrated model of sentence processing in reading* (Doctoral dissertation, University of Potsdam, Potsdam, Germany).

[49] Rasmussen, N. E., & Schuler, W. (2018). Left-corner parsing with distributed associative memory produces surprisal and locality effects. *Cognitive Science*, *42*, 1009–1042.

[50] Smith, G., Franck, J., & Tabor, W. (2018). Semantic features unpack notional plurality in pseudopartitive agreement: A self-organizing approach. *Cognitive Science*, *42*, 1043–1074.

[51] Parker, D. (2019). Cue combinatorics in memory retrieval for anaphora. *Cognitive Science*, *43*(3), e12715.

[52] Hammerly, C., Staub, A., & Dillon, B. (2019). The grammaticality asymmetry in agreement attraction reflects response bias: Experimental and modeling evidence. *Cognitive Psychology*, *110*, 70–104.

[53] McElree, B. (2006). Accessing recent events. In B. H. Ross (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 46, pp. 155–200). San Diego, CA: Elsevier.

[54] McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language*, *32*, 536–571.

[55] Parker, D., Shvartsman, M., & Van Dyke, J. A. (2017). The cue-based retrieval theory of sentence comprehension: New findings and new challenges. *Language processing and disorders*, 121–144.

[56] McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

[57] Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

[58] Vehtari, A., Ojanen, J. et al. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228.

[59] 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.

[60] 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.

[61] Caplan, D. (2012). Resource reduction accounts of syntactically based comprehension disorders. In C. K. Thompson & R. Bastiannse (Eds.), *Perspectives on agrammatism* (pp. 34–48). Psychology Press.

[62] 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.

[63] Husain, S., Vasishth, S., & Srinivasan, N. (2014). Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, *9*(7), 1–14.

[64] Safavi, M. S., Husain, S., & Vasishth, S. (2016). Dependency resolution difficulty increases with distance in Persian separable complex predicates: Implications for expectation and memory-based accounts. *Frontiers in Psychology*, *7*.

[65] Vasishth, S., Suckow, K., Lewis, R. L., & Kern, S. (2010). Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures. *Language and Cognitive Processes*, *25*(4), 533–567.

[66] Frank, S. L., Trompenaars, T., & Vasishth, S. (2015). Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, *40*, 554–578.

[67] 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.

[68] Hanne, S., Burchert, F., & Vasishth, S. (2016). On the nature of the subject-object asymmetry in wh-question comprehension in aphasia: Evidence from eye-tracking. *Aphasiology*, *30*(4), 435–462.

[69] Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133.

[70] Farmer, T. A., Misyak, J. B., & Christiansen, M. H. (2012). Individual differences in sentence processing. *Cambridge handbook of psycholinguistics*, 353–364.

[71] Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–169.

[72] Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *19*, 450–466.

[73] Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (8888). Experience and sentence comprehension: Statistical learning, and individual differences. *Cognitive Psychology*.

[74] Friedman, N. P., & Miyake, A. (2004). The reading span test and its predictive power for reading comprehension ability. *Journal of Memory and Language*, *51*(1), 136–158.

[75] Kuperman, V., & Van Dyke, J. A. (2011). Effects of individual differences in verbal skills on eye-movement patterns during sentence reading. *Journal of Memory and Language*, *65*(1), 42–73.

[76] Caplan, D., Michaud, J., & Hufford, R. (2013). Short-term memory, working memory, and syntactic comprehension in aphasia. *Cognitive Neuropsychology*, *30*(2), 77–109.

[77] Varkanitsa, M., & Caplan, D. (2018). On the association between memory capacity and sentence comprehension: Insights from a systematic review and meta-analysis of the aphasia literature. *Journal of Neurolinguistics*, *48*, 4–25.

[78] Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of memory and language*, *28*(2), 127–154.

[79] Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review*, *12*(5), 769–786.

[80] Nicenboim, B., Vasishth, S., Kliegl, R., Gattei, C., & Sigman, M. (2015). Working memory differences in long distance dependency resolution. *Frontiers in Psychology*, *6*, 312.

[81] von der Malsburg, T., & Vasishth, S. (2013). Scanpaths reveal syntactic underspecification and reanalysis strategies. *Language and Cognitive Processes*, *28*(10), 1545–1578.

[82] Fedorenko, E., Gibson, E., & Rohde, D. (2006). The nature of working memory capacity in sentence comprehension: Evidence against domain specific resources. *Journal of Memory and Language*, *54*, 541–553.

[83] Gelman, A., & Carlin, J. (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651.

[84] Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175.

[85] Jäger, L. A., Mertzen, D., Van Dyke, J. A., & Vasishth, S. (2019). Interference patterns in subject-verb agreement and reflexives revisited: A large-sample study. *Journal of Memory and Language*. Under review.

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

[87] Palestro, J. J., Sederberg, P. B., Osth, A. F., Van Zandt, T., & Turner, B. M. (2018). *Likelihood-free methods for cognitive science*. Springer.

[88] Kangasrääsiö, A., Jokinen, J. P., Oulasvirta, A., Howes, A., & Kaski, S. (2019). Parameter inference for computational cognitive models with Approximate Bayesian Computation. *Cognitive Science*, *43*(6), e12738.

[89] Kush, D. (2013). *Respecting relations: Memory access and antecedent retrieval in incremental sentence processing* (PhD thesis, University of Maryland, College Park, MD).

[90] Cunnings, I., & Sturt, P. (2014). Coargumenthood and the processing of reflexives. *Journal of Memory and Language*, *75*, 117–139.

[91] Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

[92] Haaf, J. M., & Rouder, J. N. (2018). Some do and some don't? Accounting for variability of individual difference structures. *Psychonomic Bulletin and Review*, *26*(3), 1–18.

[93] Spiegelhalter, D., & Blastland, M. (2013). *The Norm chronicles: Stories and numbers about danger*. Profile Books.

[94] Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(6), 1411–1423.

[95] Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*(5), 425–430.

[96] Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*, 97–114.

[97] 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.

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

[99] Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

[100] Logačev, P., & Vasishth, S. (2016b). Understanding underspecification: A comparison of two computational implementations. *Quarterly Journal of Experimental Psychology*, *69*(5), 996–1012.

Box 1: Comparing the activation and direct-access model The direct-access model can be defined as a finite mixture model as follows. Let y be the reading time in milliseconds, and β the mean time in log milliseconds taken for a successful retrieval, with standard deviation σ. Such a successful retrieval happens with probability p. Retrieval is assumed to fail with probability $(1-p)$, and the extra cost of re-attempting and successfully carrying out retrieval is δ log ms. For the full, hierarchically specified model, see [18].

$$y \sim \begin{cases} LogNormal(\beta, \sigma^2), & \text{retrieval succeeds, probability } p \\ LogNormal(\beta + \delta, \sigma^2), & \text{retrieval fails initially, probability } 1-p \end{cases} \quad (1)$$

We can now determine whether the observed data are underlyingly coming from a two-component mixture (the direct-access model) or from the activation model, and whether a mixture distribution yields better predictions with respect to the data. Figure I shows a comparison of the relative predictive fits from the hierarchical Bayesian models implementing the activation model, and McElree's direct-access model as a finite mixture process. The violin plots show posterior predictive distributions from the model; their width represents the density of the predicted mean reading times. The black circles show the empirically observed mean reading times. The four types of reading times refer to four different kinds of question responses that the participants gave in a self-paced reading task [41]. The activation model overestimates the reading times in the incorrect responses, compared to the direct-access model. Although not shown here, this overestimation is due to the activation model assuming a single variance component for both correct and incorrect responses. When that assumption is relaxed, both models show similar predictive accuracy [18].

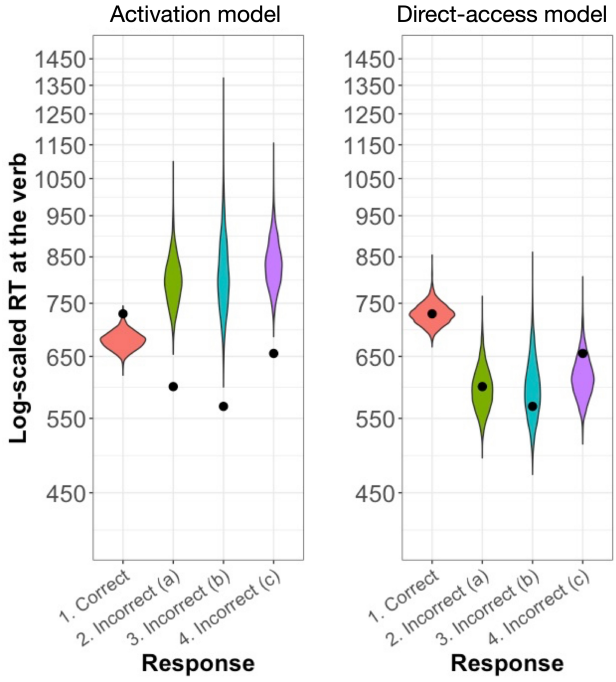

Fig. I. **A comparions of the activation and direct-access models**. A comparison of observed sample means with the posterior predictive distributions of the activation and direct-access models.

Box 2: Cue-weighting in the activation model Figure I shows the quantitative predictions for different values of three parameters in the model: the latency factor, which is a scaling factor that maps retrieval time to reading times; maximum associative strength (mas), which specifies the strength between retrieval cues and features of items in memory as well as the total pool of activation available that can be allocated to items in memory; and cue-weighting, which controls whether a particular cue has a higher weight than another cue (expressed as a ratio).

For example, suppose that there are two retrieval cues [number: plural, subject: yes], as in example (2). The default assumption in the activation model is that both the cues have an equal effect on the activation of the item in memory; this can be represented as 1:1 weighting. However, if the subject cue is weighted higher (for example, it could have twice or four times the weight of number), then the subject cue will dominate in affecting the activation. As the figure shows, when the cue-weighting ratio is 2:1 or 4:1, the interference effects (inhibitory and facilitatory) shrinks toward 0. Individual participants may weight the subject cue differently, based on fluency or varying degrees of attention, leading to differences in the magnitude of the effect across individuals (Figure 5).

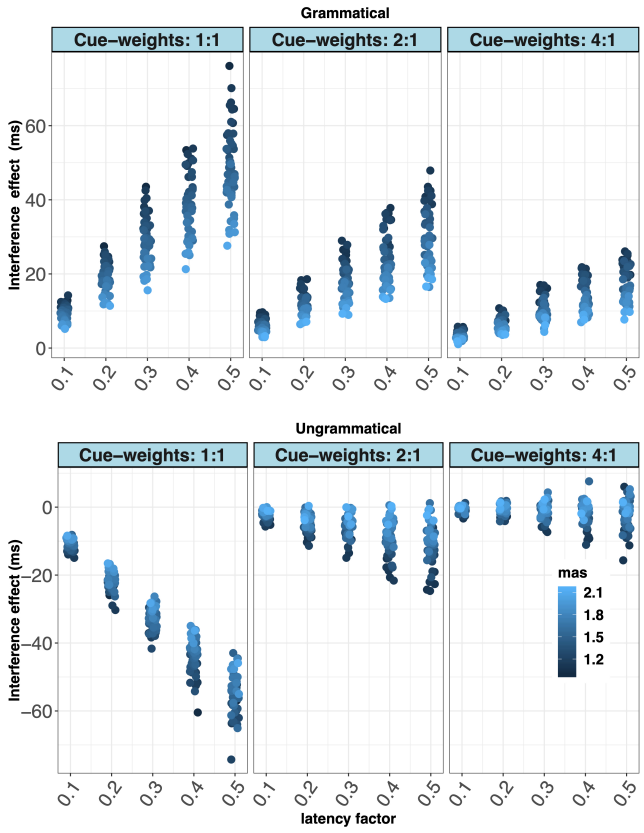

Fig. I. **The effect of cue-weighting on interference effect magnitudes**. Shown are the quantitative predictions of inhibitory and facilitatory interference effects in the activation model [17] as a function of several theoretically important parameters in the activation model that can be used to account for individual-level differences in the magnitude of interference effects.

Outstanding questions

- **How would facilitatory interference effects be explained in the direct-access model?** It would be very valuable to have a computational implementation of the direct-access model that makes explicit its predictions for facilitatory interference effects.
- **What benchmark data should a computational model be able to explain?** Appropriately powered, large-sample benchmark data from different languages are needed from a broad spectrum of syntactic phenomena and languages in order to evaluate quantitative model predictions.
- **What is the relative performance of competing computational models of sentence processing against benchmark empirical data?** Several competing computational models of sentence processing are now available, but their relative predictive performance needs to be systematically tested against published benchmark data.
- **What predictions do models of sentence processing make for sentence comprehension in different populations?** A potentially rich and until now underdeveloped area for research is using models of unimpaired processing to attempt to explain sources of variability in comprehension difficulty in impaired populations such as individuals with aphasia, and in non-native speakers. For example, bilingual non-native speakers may display processing characteristics, such as slowed processing, or cue-weighting different from fluent native speakers, that could arise from a lack of fluency.
- **Can computational models explain both aggregate and individual-level behavior?** Quantitative models should aim to explain both group-level effects as well as the (often systematic) variability seen in individual participants.