A computational investigation of sources of variability in sentence comprehension difficulty in aphasia

Paul Mätzig (pmaetzig@uni-potsdam.de)

University of Potsdam, Germany, Human Sciences Faculty, Department of Linguistics

Shravan Vasishth (vasishth@uni-potsdam.de)

University of Potsdam, Germany, Human Sciences Faculty, Department of Linguistics

Felix Engelmann (felix.engelmann@manchester.ac.uk)

The University of Manchester, UK, Faculty of Biology, Medicine and Health, Child Study Centre

David Caplan (dcaplan@partners.org)

Massachusetts General Hospital, USA, Neurology

Frank Burchert (burchert@uni-potsdam.de)

University of Potsdam, Germany, Human Sciences Faculty, Department of Linguistics

Abstract

We present a computational evaluation of three hypotheses about sources of deficit in sentence comprehension in aphasia: slowed processing, intermittent deficiency, and resource reduction. The ACT-R based Lewis & Vasishth (2005) model is used to implement these three proposals. Slowed processing is implemented as slowed execution time of parse steps; intermittent deficiency as increased random noise in activation of elements in memory; and resource reduction as reduced spreading activation. As data, we considered subject vs. object relative sentences, presented in a self-paced listening modality to 56 individuals with aphasia (IWA) and 46 matched controls. The participants heard the sentences and carried out a picture verification task to decide on an interpretation of the sentence. These response accuracies are used to identify the best parameters (for each participant) that correspond to the three hypotheses mentioned above. We show that controls have more tightly clustered (less variable) parameter values than IWA; specifically, compared to controls, among IWA there are more individuals with slow parsing times, high noise, and low spreading activation. We find that (i) individual IWA show differential amounts of deficit along the three dimensions of slowed processing, intermittent deficiency, and resource reduction, (ii) overall, there is evidence for all three sources of deficit playing a role, and (iii) IWA have a more variable range of parameter values than controls. An important implication is that it may be meaningless to talk about sources of deficit with respect to an abstract average IWA; the focus should be on the individual's differential degrees of deficit along different dimensions, and on understanding the causes of variability in deficit between participants.

Keywords: Sentence Comprehension; Aphasia; Computational Modeling; Cue-based Retrieval

A computational investigation of sources of variability in sentence comprehension difficulty in aphasia

Introduction

In healthy adults, sentence comprehension has long been argued to be influenced by individual differences; a commonly assumed source is differences in working memory capacity (Daneman & Carpenter, 1980; Just & Carpenter, 1992). Other factors such as age (Caplan & Waters, 2005) and cognitive control (Novick, Trueswell, & Thompson-Schill, 2005) have also been implicated.

An important question that has not received much attention in the computational psycholinguistics literature is: what are sources of individual differences in healthy adults versus impaired populations, such as individuals with aphasia (IWA)?

It is well known that individuals with aphasia often experience difficulties in comprehending sentences. These difficulties are mainly observable as lower accuracy scores in comprehension tasks such as sentence-picture matching, in which a picture must be selected in accordance with the meaning of a sentence, or in object-manipulation task, in which the meaning of a sentence must be reenacted with figurines (see literature review in Patil, Hanne, Burchert, De Bleser, & Vasishth, 2016). Furthermore, eye-tracking during comprehension studies have revealed that IWA exhibit slower overall processing times (Hanne, Sekerina, Vasishth, Burchert, & Bleser, 2011).

Two factors that are known to affect performance in sentence comprehension tasks (such as sentence-picture-matching, see Hanne et al., 2011) are canonicity (i.e., word order), and reversibility of thematic roles of animate nouns. Comprehension difficulties in IWA are selective in nature and particularly pronounced in sentences that are semantically reversible and have non-canonical word order, for example passives or object relative clauses. For these sentence structures, response accuracy is often indistinguishable from guessing (50% accuracy). Such a pattern is referred to as chance performance. On the other hand, performance for canonical structures (e.g., actives or subject relative clauses) and irreversible sentences is often within normal range (Hanne et al., 2011). While chance performance is a typical trait of Broca's aphasia, it can be observed in other aphasia syndromes as well.

Regarding the underlying nature of this deficit in IWA, two primary approaches have been proposed in the literature: *representational* vs. *processing accounts*. Whereas representational accounts (Grodzinsky, 1995) argue that chance performance is caused by impaired syntactic representations on which the parser operates, processing accounts assume an underlying deficit in parsing procedures proper. The exact nature of the impairment in mechanisms that are employed in parsing operations, however, are still not clear, and several proposals have been made. In this paper, we focus on processing accounts of sentence comprehension deficits in IWA and, specifically, evaluate three influential proposals (for a detailed evaluation of representational accounts, see Patil et al., 2016):

1. *Slowed processing*: Burkhardt, Piñango, and Wong (2003) argue that a slowdown in parsing mechanisms can best explain the processing deficit.
2. *Intermittent deficiencies*: Caplan, Michaud, and Hufford (2015) suggest that occasional temporal breakdowns of parsing mechanisms capture the observed behavior.
3. *Resource reduction*: A further hypothesis, due to Caplan (2012), is that the deficit is caused by a reduction in resources related to sentence comprehension.

Computational modeling can help evaluate these different proposals quantitatively. Specifically, the cue-based retrieval account of Lewis and Vasishth (2005), which was developed within the ACT-R framework (Anderson et al., 2004), is a computationally implemented model of unimpaired sentence comprehension that has been used to model a broad array of empirical phenomena in sentence processing relating to similarity-based interference effects (Engelmann, Jäger, & Vasishth, 2016; Jäger, Engelmann, & Vasishth, 2017; Nicenboim & Vasishth, 2018) and the interaction between oculomotor control and sentence comprehension (Engelmann, Vasishth, Engbert, & Kliegl, 2013).[1]

[1]The model can be downloaded in its current form from https://github.com/felixengelmann/act-r-sentence-parser-em.

The Lewis and Vasishth model (and its more recent version by Engelmann, 2015) integrates top-down predictions into bottom-up, word-by-word parsing steps by using a left-corner parsing algorithm. The model assumes that words are being read on a (simulated) screen. Simplifying somewhat, the processing cycle consists of the following stages, which are repeated until there are no more words to process: (1) lexical access and integration with the current parse, (2) predictive structure building; and (3) dependency completion (e.g., connecting a grammatical subject and a verb) through a so-called retrieval process. The Lewis and Vasishth (2005) model is particularly attractive for studying sentence comprehension because the retrieval process relies on the general constraints on cognitive processes that have been laid out in the ACT-R framework. This makes it possible to investigate whether sentence processing could be seen as being subject to the same general cognitive constraints as any other information processing task (of course, this does not entail that there are no language-specific constraints on sentence comprehension). A further advantage of the Lewis and Vasishth (2005) model in the context of theories of processing deficits in aphasia is that several of its numerical parameters (which are part of the general ACT-R framework) can be interpreted as implementing the three proposals mentioned above.

The objective of the present paper is to demonstrate that it is possible to map the three types of deficits mentioned above to three distinct parameters (described below) of the ACT-R architecture, and to use, for each individual IWA, the best-fitting values of these parameters as a means to distinguish between impaired and unimpaired individuals.

In Patil et al. (2016), the Lewis and Vasishth (2005) architecture was used to model aphasic sentence processing on a small scale, using data from seven IWA. They modeled proportions of fixations in a visual world task, response accuracies and response times for empirical data of a sentence-picture matching experiment by Hanne et al. (2011). Their goal was to test two of the three hypotheses of sentence comprehension deficits mentioned above, slowed processing and intermittent deficiency. Their results revealed the best fit for the model that implemented both of the accounts, compared to models that only implemented one. Further, the results lead to the conclusion that IWA exhibit deficits to differing amounts.

One major limitation of the Patil et al. study was the limited data it was based on: 7 IWA. In the present work, we provide a proof of concept study that goes beyond Patil et al. (2016) in two respects: first, we use a much larger data-set from Caplan et al. (2015) with 56 IWA and 46 matched controls; and second, we evaluate the evidence for all the three hypotheses mentioned above. In future work, we intend to extend the data that will be used to evaluate the model.

Before we describe the modeling carried out in the present paper and the data used for the evaluation, we first introduce the cognitive constraints assumed in the Lewis and Vasishth (2005) model that are relevant for this work, and show how the theoretical approaches to the aphasic processing deficit can be implemented using specific model parameters. Having introduced the essential elements of the model architecture, we simulate comprehension question-response accuracies for unimpaired controls and IWA, and then fit the simulated accuracy data to published data (Caplan et al., 2015) from controls and IWA.

When fitting individual participants, we vary three parameters that map to the three theoretical proposals mentioned above. The goal is to determine whether the distributions of optimal parameter values computed for individual participants furnish any support for any of the three sources of deficits in processing. We expect that if there is a tendency in one parameter to show non-default values in individual model fits for IWA, for example slowed processing, then there is support for the claim that slowed processing is an underlying source of processing difficulty in IWA. Similar predictions hold for the other two constructs, intermittent deficiency and resource reduction; and for combinations of the three proposals.

Constraints on sentence comprehension in the model

In this section, we describe some of the constraints assumed in the Lewis and Vasishth (2005) sentence processing model. Then, we discuss the model parameters that can be mapped to the three theoretical proposals for the underlying processing deficit in IWA.

ACT-R is a computational framework for modeling cognitive processes. Within this framework, domain-specific models are usually developed that are intended to explain some specific aspect of cognition; for example, a word-list memorization task. The Lewis and Vasishth (2005) model is an example of such an implementation, and aims to spell out the link between general memory constraints and sentence processing. It is useful for computational evaluations of empirical data because it is possible to calculate quantitative predictions for sentence comprehension phenomena, in the form of reading times and response accuracies.

The Lewis and Vasishth model adopts the ACT-R distinction between long-term declarative memory and procedural knowledge. In the sentence processing context, the latter is a set of parsing rules. These rules operate on units of information known as chunks to build syntactic structure. Chunks are elements in declarative memory that are defined in terms of feature-value specifications, much as in the syntactic formalism head-driven phrase structure grammar (Sag, Wasow, Bender, & Sag, 1999). In the context of sentence comprehension, both single words and syntactic subtrees, such as verb phrases, are chunks in memory. As an example of a chunk, a noun like *book* could be stored as a feature-value matrix that states that the part-of-speech is nominal, number is singular, and animacy status is inanimate:

$$\begin{pmatrix} \text{pos} & \textit{nominal} \\ \text{number} & \textit{sing} \\ \text{animate} & \textit{no} \end{pmatrix}$$

Each chunk is associated an *activation*, a numeric value that determines the probability and latency of access from declarative memory. Accessing chunks in declarative memory happens via a cue-based retrieval mechanism that is specified in ACT-R. For example, when we read a sentence like *The book is a good read*, at the auxiliary verb *is*, a dependency must be built between *book* and *is* in order to arrive at the meaning of the sentence. In the Lewis and Vasishth model, this dependency is built by a parsing rule triggering an access at the auxiliary for a noun that has certain features. The lexical specification of the auxiliary verb demands that the subject have the following feature specification: {part-of-speech nominal, number singular, and animate no}. These retrieval cues are used to search for and access the relevant chunk in memory. Retrieval only succeeds if the activation of a to-be-retrieved chunk is above a minimum threshold, which is a parameter in ACT-R. If retrieval fails, we say that there was a retrieval failure.

The retrieval of a chunk described above depends on its activation level, which is determined by several constraints; we discuss a subset of these next. Let C be the set of all chunks in declarative memory. The total activation of a chunk $i \in C$ equals

$$A_i = B_i + S_i + \epsilon, \tag{1}$$

where B_i is the base-level or resting-state activation of the chunk i; the second summand S_i represents the spreading activation that a chunk i receives during a particular retrieval event; and ϵ is noise that is logistically distributed, approximating a normal distribution, with mean 0 and standard deviation ANS; the noise term is generated at each new retrieval request. The time it takes for a chunk i to be retrieved T_i depends on its activation A_i via $T_i = F\exp(-A_i)$, where F is a scaling constant which we kept constant at 0.2 here.

The parameter ANS of the logistic distribution from which ϵ is generated can be interpreted as implementing the *intermittent deficiency* hypothesis, because higher values of ANS will tend to lead to more fluctuations in activation of a chunk and therefore higher rates of unsuccessful retrievals.[2] Increasing ANS leads to a larger influence of random fluctuation in activation on a chunk's activation, which represents the core idea of *intermittent deficiency*: that there is not a constantly present damage to the processing system, but rather that the deficit occasionally interferes with parsing,

[2]Note that Patil et al. (2016) implemented intermittent deficiency using another source of noise in the model which affects not activation levels but rather the odds of a processing operation being chosen. In future work, we will compare the relative change in quality of fit when intermittent deficiency is implemented in this way.

leading to more errors.

The second summand in (1) represents the process of *spreading activation* within the ACT-R framework. For a given chunk i to be retrieved, and given retrieval cues $j \in \{1, \ldots, J\}$, the amount of activation spread to the chunk i as a function of the retrieval cues is quantified by computing:

$$S_i = \sum_{j=1}^{J} W_j S_{ji}. \tag{2}$$

The weighting term W_j is assumed by default to be $\frac{\mathrm{GA}}{J}$, where GA is an ACT-R parameter[3] and S_{ji} is a value that reflects the strength of association between the retrieval cue j and the chunk i. To give a simplified example of what spreading activation does, assume that two chunks i_1 and i_2 are in memory, and there are two retrieval cues c_1 and c_2 that are being used to search for a chunk. Chunk i_1 matches fully with the two retrieval cues, and chunk i_2 matches with only one of the retrieval cues, say c_2. Suppose also that the strength of association S_{ji} between a retrieval cue j and a chunk i is 1 if the retrieval cue matches a feature on the chunk, and 0 otherwise. Assume that GA is 1. Then, the activation spread to item i_1 as a function of the two retrieval cues is:

$$S_{i_1} = W_1 S_{11} + W_2 S_{21} = 1/2 \times 1 + 1/2 \times 1 = 1 \tag{3}$$

and the activation spread to item i_2 is:

$$S_{i_2} = W_1 S_{12} + W_2 S_{22} = 1/2 \times 1 + 1/2 \times 0 = 1/2 \tag{4}$$

More activation is spread to chunk i_1 than to chunk i_2 due to a full match with the retrieval cues. Now consider the case where $GA = 2$. Now, $S_{i_1} = 2/2 \times 1 + 2/2 \times 1 = 2$ and $S_{i_2} = 2/2 \times 1 + 2/2 \times 0 = 1$. Thus, with a higher GA value, more activation is spread to a chunk in memory when it matches the retrieval cues.

The parameter GA therefore determines the total amount of activation increase in a chunk as a function of the number of retrieval cues it matches. It is a free parameter

[3] GA stands for goal activation in ACT-R, for reasons that are not relevant here.

in ACT-R. This parameter has already been used to model individual differences in working memory capacity (see, for example, Daily, Lovett, & Reder, 2001 and Engelmann, 2015). The lower the GA value, the lower the increase in activation of the chunk i due to a match with the retrieval cues. Low GA values thus model low working memory capacity: when an item needs to be retrieved from memory using certain retrieval cues, the probability of successful retrieval will go down if GA is lower. Similarly, higher GA values model higher working memory capacity. Thus, it can be seen as one way (although by no means the only way) to implement the resource reduction hypothesis.

Finally, the hypothesis of *slowed processing* can be mapped to the *default action time* (DAT) parameter in ACT-R. This defines the constant amount of time it takes a selected parsing rule to “fire”, i.e., to start the actions specified in the action part of the rule. Higher values would lead to a greater delay in firing of parsing rules. Due to the longer decay in this case, retrieval may be slower and more failed retrieval attempts may occur.

In the following section, we evaluate the model’s performance on the empirical data for IWA and unimpaired individuals, implementing the three theoretical claims by varying the three parameters described above.

Simulations

In this section we describe our modeling method and the procedure we use for fitting the model results to the empirical data from Caplan et al. (2015).

Materials

We used the data from 56 IWA and 46 matched controls published in Caplan et al. (2015). In this data-set, participants listened to recordings of sentences presented word-by-word; they paced themselves through the sentence, providing self-paced listening data. Participants processed 20 examples of 11 spoken sentence types and indicated which of two pictures corresponded to the meaning of each sentence. This yielded accuracy data for each sentence type.

Out of the 11 sentence types, we chose the subject/object relative clause contrast for the current simulation: subject relatives (*The woman who hugged the girl washed the boy*) represent the arguments of the sentence (woman, girl) in canonical order, whereas in object relatives (*The woman who the girl hugged washed the boy*), they occur in non-canonical order. We chose relative clauses for two reasons. First, relative clauses have been very well-studied in psycholinguistics and serve as a typical example where processing difficulty is (arguably) experienced due to deviations in canonical word ordering (Just & Carpenter, 1992). Second, the Lewis and Vasishth model already has productions defined for these constructions, so the relative clause data serve as a good test of the model as it currently stands.

Parameter estimation

We used grid search to find the best fitting parameters. We refer to the parameter space Π_i as the set of all vectors $(\mathrm{GA}, \mathrm{DAT}, \mathrm{ANS})$ with $\mathrm{GA},\ \mathrm{DAT},\ \mathrm{ANS} \in \mathbb{R}$. For computational convenience, we chose a discretisation of Π by defining a step-width and lower and upper boundaries for each parameter. In this discretised space Π', we chose $\mathrm{GA} \in \{0.2, 0.3, \ldots, 1.1\}$, $\mathrm{DAT} \in \{0.05, 0.06, \ldots, 0.1\}$, and $\mathrm{ANS} \in \{0.15, 0.2, \ldots, 0.45\}$.[4] Π' could be visualised as a three-dimensional grid of 420 dots, which are the elements $p' \in \Pi'$.

The default parameter values were included in Π'. This means that models that vary only one or two of the three parameters were included in the simulations.

For all participants in the Caplan et al. data-set, we calculated comprehension question response accuracies, averaged over all items of the subject / object relative clause condition. For each $p' \in \Pi'$, we ran the model for 1000 iterations for the subject and object relative tasks. From the model output, we determined whether the model made the correct attachment in each iteration, i.e., whether the correct noun was selected as subject of the embedded verb, and we calculated the accuracy in a simulation for a given parameter $p' \in \Pi'$ as the proportion of iterations where the model

[4]The standard settings in the Lewis and Vasishth (2005) model are $\mathrm{GA} = 1$, $\mathrm{DAT} = 0.05$ (or 50 ms), and $\mathrm{ANS} = 0.15$.

made the correct attachment. We counted a parsing failure, where the model did not create the target dependency, as an incorrect response.

The problem of finding the best fit for each subject can be phrased as follows: for all subjects, find the parameter vector that minimises the absolute distance between the model accuracy for that parameter vector and each subject's accuracy. Because there might not always be a unique p' that solves this problem, the solution can be a set of parameter vectors. If for any one participant multiple optimal parameters were calculated, we averaged each parameter value to obtain a unique parameter vector. This transforms the parameter estimates from the discretised space Π' to the original parameter space Π.

Results

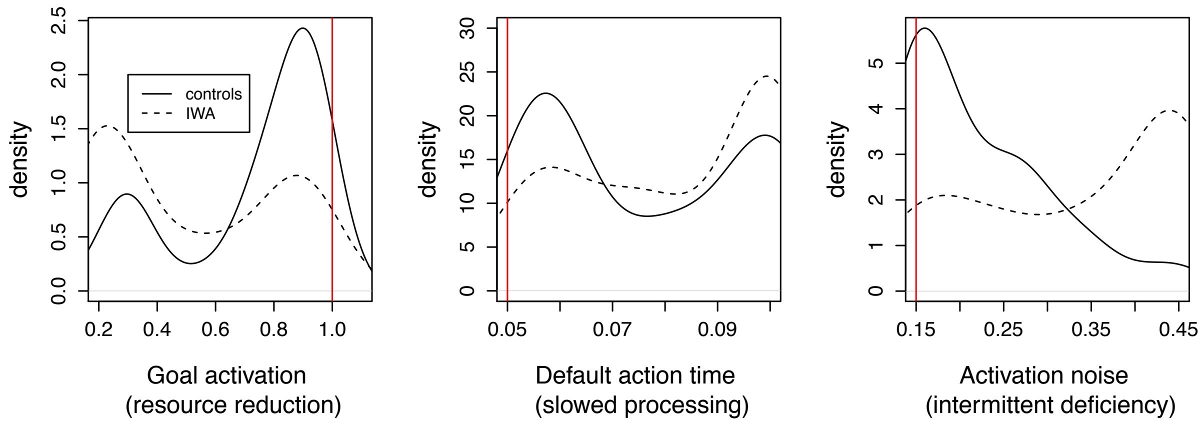

Figure 1. Marginal distributions of each of the three parameters for subject relatives in controls (solid lines) vs. IWA (dotted lines). The vertical line shows the default setting for the respective parameter.

In this section we presents the results of the simulations and the fit to the data. First, we describe the general pattern of results reflected by the distribution of non-default parameter estimates per subject. Following that, we test whether tighter clustering occurs in controls.

Distribution of parameter value estimates. Table 1 shows the number of participants for which a non-default parameter value was predicted. We refer to the

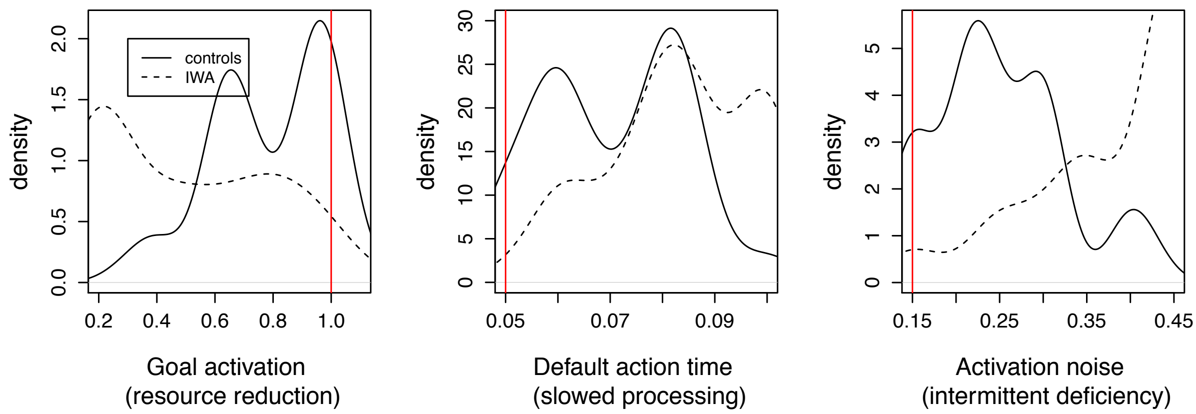

Figure 2. Marginal distributions of each of the three parameters for object relatives in controls (solid lines) vs. IWA (dotted lines). The vertical line shows the default setting for the respective parameter.

values GA = 1, DAT = 0.05 (or 50 ms), and ANS = 0.15 as the default values, as set in the ACT-R architecture. It is clear that, as expected, the number of subjects with non-default parameter values is always larger for IWA vs. controls, but controls show non-default values unexpectedly often. In controls, the main difference between subject and object relatives is a clear increase in elevated noise values in object relatives.

For IWA in subject relatives, the single-parameter models are very similar, whereas in simple object relatives, most IWA (95%) exhibit elevated noise values, while a far smaller proportion (71%) showed reduced goal activation values.

Figures 1 and 2 illustrate the smoothed marginal distributions of parameter value estimates, for subject and object relative clauses, respectively. Most importantly, it is visible in both subject and object relatives that the distributions of controls' estimates have their point of highest density around the default value of the respective parameter. Deviations from this observation are mainly visible in the distributions for object relatives, where a second peak further away from the default is visible for each parameter. Distributions for IWA, on the other hand, are much flatter, and most density is concentrated relatively far away from the default parameter setting. This situation is exacerbated in object relatives compared to subject relatives.

Overall, most IWA exhibit non-default parameter settings ANS and DAT, and to

		GA	DAT	ANS	GA & DAT	GA & ANS	DAT & ANS	GA & DAT & ANS
SR	control	19	24	18	18	11	16	10
	IWA	38	41	42	32	33	36	27
OR	control	21	26	36	21	20	25	20
	IWA	40	48	53	38	40	48	38

Table 1

The number of participants in ***subject / object relatives*** *(SR/OR) for which non-default parameter values were predicted, in the subject vs. object relative tasks, respectively; for goal activation (GA), default action time (DAT) and noise (ANS) parameters.*

a lesser extent in GA. Table 1 shows further that the only combined model (i.e., the model that varied two or more parameters instead of keeping the other two at their default value) that matches the single variation model for DAT or ANS is the one combining DAT and ANS. We suspect that the lower number of IWA for which non-default GA values were estimated are due to GA and ANS eliciting similar model behavior. We address this point in the discussion below.

Cluster analysis. In order to investigate the predicted clustering of parameter estimates, we performed a cluster analysis on the data too see to which degree controls and IWA could be discriminated. If our prediction is correct that, compared to IWA, clustering is tighter in controls, we expect that a higher proportion of the data should be correctly assigned to one of two clusters, one corresponding to controls, the other one corresponding to IWA. We chose hierarchical clustering to test this prediction (Friedman, Hastie, & Tibshirani, 2001).

We combined the data for subject and object relatives into one respective data set. We calculated the dendrogram and cut the tree at 2, because we are only looking for the discrimination between controls and IWA. The results of this are shown in Table 2. The clustering is able to identify controls better than IWA, but the identification of IWA is better than chance (50%). Discriminative ability might improve if all 11 constructions in Caplan et al. (2015) were to be used; this will be investigated in future work.

	Subject relatives		Object relatives	
predicted group	controls	IWA	controls	IWA
control	**34**	21	**42**	24
IWA	12	**35**	4	**32**
accuracy	74%	63%	91%	57%

Table 2

Discrimination ability of hierarchical clustering on the combined data for ***subject / object relatives****. Numbers in bold show the number of correctly clustered data points. The bottom row shows the percentage accuracy.*

Discussion

The simulations and cluster analysis above demonstrate overall tighter clustering in parameter estimates for controls, and more variance in IWA. This is evident from the clustering results in Table 2. These findings are consistent with the predictions of the small-scale study in Patil et al. (2016). However, there is considerable variability even in the parameter estimates for controls, more than expected based on the results of Patil et al..

The distribution of non-default parameter estimates (see Figures 1, 2 and Table 1) suggest that all three hypotheses are possible explanations for the patterns in our simulation results: compared to controls, estimates for IWA tend to include higher default action times and activation noise scales, and lower goal activation. These effects generally appear to be more pronounced in object relatives vs. subject relatives. This means that all the three hypotheses can be considered viable candidate explanations. Overall, more IWA than controls display non-default parameter settings. Although there is evidence that many IWA are affected by all three impairments in our implementation, there are also many patients that show only one or two non-default parameter values. Again, this is more the case in object relatives than in subject relatives.

In general, there is evidence that all three deficits are plausible to some degree. However, IWA differ in the degree of the deficits, and they have a broader range of parameter values than controls. Nevertheless, even the controls show a broad range of differences in parameter values, and even though these are not as variable as IWA, this suggests that some of the unimpaired controls can be seen as showing slowed processing, intermittent deficiencies, and resource reduction to some degree.

There are several problems with the current modeling method. First, using the ACT-R framework with its multiple free parameters has the risk of overfitting. We plan to address this problem in three ways in future research: (1) Testing more constructions from the Caplan et al. (2015) data-set might show whether the current estimates are unique to this kind of construction, or if they are generalisable. (2) We plan to create a new data-set analogous to Caplan's, using German as the test language. Once the English data-set has been analysed and the conclusions about the different candidate hypotheses have been tested on English, a crucial test of the conclusions will be cross-linguistic generalisability. (3) A systematic model comparison method such as k-fold cross-validation might serve as a means to formally compare the implementations of the theoretical claims of aphasic sentence processing.

Second, the use of accuracies as modeling measure has some drawbacks. Informally, in an accuracy value there is less information encoded than in, for example, reading or listening times. In future work, we aim to implement an approach modeling both accuracies and listening times (Nicenboim & Vasishth, 2018). Also, counting each parsing failure as 'wrong' might yield overly conservative accuracy values for the model; this can be addressed by assigning a random component into the calculation. This reflects more closely a participant who guesses if he/she did not fully comprehend the sentence.

Third, related to the overfitting problem addressed above, at least two of the varied parameters – goal activation and activation noise – lead to similar effects when manipulated in the way described here. More specifically, the decision to use the ANS parameter makes the assumption that the high noise levels for IWA influence all declarative memory retrieval processes, and thus the whole memory, not only the production system. Similarly, assuming lower GA values for IWA amounts to assuming generally lower working memory capacity in those participants, not specifically lower verbal working memory. Both parameters lead to a higher rate of retrieval failures. Because of this, it will be worth investigating in future work whether other sources of noise in the ACT-R framework may be a better way to model intermittent deficiencies (see Patil et al., 2016 for an example).

Lastly, simulating the subject vs. object relative tasks separately yields the undesirable interpretation of participants' parameters varying across sentence types. While this is not totally implausible, estimating only one set of parameters for all sentence types would reduce the necessity of making additional theoretical assumptions on the underlying mechanisms, and allows for easier comparisons between different syntactic constructions. We plan to do this in future work.

Although our method, as a proof of concept, showed that all three hypotheses are supported to some degree, it is worth investigating more thoroughly how different ACT-R mechanisms are influenced by changes in the three varied parameters in the present work. Implementing more of the constructions from Caplan et al. (2015) will, for example, enable us to explore how the different hypotheses interact with each other in our implementation.

One possible way to delve deeper into identifying the sources of individual variability in IWA could be to investigate whether sub-clusters show up within the IWA parameter estimates. For example, different IWA being grouped together by high noise values could be interpreted as these patients sharing a common source of their sentence processing deficit (in this hypothetical case, our implementation of intermittent deficiencies). We will address this question once we have simulated data for more constructions of the Caplan et al. (2015) data-set.

Concluding remarks

We evaluated three well-known verbally stated hypotheses about causes of deficits in sentence comprehension in aphasia: slowed processing, intermittent deficiency, and resource reduction. We implemented these hypotheses within a computational model of sentence processing, the Lewis and Vasishth (2005) model of cue-based retrieval. The three hypotheses can be implemented by changing the default values of three different parameter values within the Lewis and Vasishth model. Using a large data-set from IWA and unimpaired controls, we estimated the optimal values for each of these parameters for each individual separately. We found that, compared to controls, IWA have more variable optimal parameter values than controls, and that IWA show differential degrees of deficit, where a deficit is considered to exist if an optimal parameter value for an individual deviates from the default value of the parameter. Thus, all three hypotheses about deficits in sentence comprehension may be viable explanations of processing difficulty; however, for each individual, the degree of impairment along each of these dimensions of slowed processing, intermittent deficiency, and resource reduction is likely to differ. An important implication is that it is not meaningful to state hypotheses about deficits in IWA in terms of average behavior: multiple causes of deficit may exist in any one individual, and the degree of deficit along each dimension in each individual may differ. Understanding deficits in IWA requires shifting the focus toward understanding the nature of the variability between individuals.

Acknowledgements

Paul Mätzig was funded by the Studienstiftung des deutschen Volkes. This research was partly funded by the Volkswagen Foundation grant 89 953 to Shravan Vasishth, and grant B2 in the Collaborative Research Center (Sonderforschungsbereich 1287) entitled *Limits of Variability in Language*, funded by the Deutsche Forschungsgemeinschaft and granted to Shravan Vasishth, Frank Burchert, and Nicole Stadie.

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.

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.

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.

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. (2005). The relationship between age, processing speed, working memory capacity, and language comprehension. *Memory*, *13*(3-4), 403-413.

Daily, L. Z., Lovett, M. C., & Reder, L. M. (2001). Modeling individual differences in working memory performance: A source activation account. *Cognitive Science*, *25*(3), 315–353.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Engelmann, F. (2015). *Toward an integrated model of sentence processing in reading.* Unpublished doctoral dissertation, University of Potsdam.

Engelmann, F., Jäger, L. A., & Vasishth, S. (2016). *The effect of prominence and cue association in retrieval processes: A computational account.* Retrieved from `https://osf.io/b56qv/`

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.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning.*

Berlin: Springer.

Grodzinsky, Y. (1995). Trace deletion, theta-roles, and cognitive strategies. *Brain and Language*, *53*(1), 469–497.

Hanne, S., Sekerina, I., Vasishth, S., Burchert, F., & Bleser, R. D. (2011). Chance in agrammatic sentence comprehension: What does it really mean? Evidence from Eye Movements of German Agrammatic Aphasics. *Aphasiology*, *25*, 221-244.

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.

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

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

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. doi: 10.1016/j.jml.2017.08.004

Novick, J. M., Trueswell, J. C., & Thompson-Schill, S. L. (2005). Cognitive control and parsing: Reexamining the role of Broca's area in sentence comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(3), 263–281.

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.

Sag, I. A., Wasow, T., Bender, E. M., & Sag, I. A. (1999). *Syntactic theory: A formal introduction* (Vol. 92). Center for the Study of Language and Information Stanford, CA.