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

Paul Mätzig (pmaetzig@uni-potsdam.de)
University of Potsdam, Human Sciences Faculty, Department Linguistics,
24–25 Karl-Liebknecht-Str., Potsdam 14476, Germany

Shravan Vasishth, (vasishth@uni-potsdam.de)
University of Potsdam, Human Sciences Faculty, Department Linguistics,
24–25 Karl-Liebknecht-Str., Potsdam 14476, Germany

Felix Engelmann (felix.engelmann@manchester.ac.uk)
The University of Manchester, School of Health Sciences
Child Study Centre, Coupland 1, Oxford Road, Manchester M13 9PL

David Caplan (dcaplan@partners.org)
Massachusetts General Hospital
175 Cambridge St, #340, Boston, Massachusetts 02114

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 ? (?) model is used to implement these three proposals. Slowed processing is implemented as slowed default production-rule firing time; intermittent deficiency as increased random noise in activation of chunks in memory; and resource reduction as reduced goal activation. As data, we considered subject vs. object relatives 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 low goal activations, high noise, and slow default action times. This suggests that (i) individual patients show differential amounts of deficit along the three dimensions of slowed processing, intermittent deficient, 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. In sum, this study contributes a proof of concept of a quantitative implementation of, and evidence for, these three accounts of comprehension deficits in aphasia.

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

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 (?, ?, ?). Other factors such as age (?, ?) and cognitive control (?, ?) 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 sentence processing performance in IWA is characterised by a performance deficit that expresses itself as slower overall processing times, and lower accuracy in question-response tasks (see literature review in ?, ?). These performance deficits are especially pronounced when IWA have to engage with sentences that have non-canonical word order and that are semantically reversible, e.g. Object-Verb-Subject versus Subject-Verb-Object sentences (?, ?).

Regarding the underlying nature of this deficit in IWA, there is a consensus that some kind of disruption is occurring in the syntactic comprehension system. The exact nature of this disruption, however, is not clear. Although a broad range of proposals exist (see ?, ?), we focus on three influential proposals here:

1. *Intermittent deficiencies*: ? (?) suggest that occasional temporal breakdowns of parsing mechanisms capture the observed behaviour.
2. *Resource reduction*: A third hypothesis, due to ? (?), is that the deficit is caused by a reduction in resources related to sentence comprehension.
3. *Slowed processing*: ? (?) argue that a slowdown in parsing mechanisms can best explain the processing deficit.

Computational modelling can help evaluate these different proposals quantitatively. Specifically, the cue-based retrieval account of ? (?), which was developed within the ACT-R framework (?, ?), 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 (?, ?, ?, ?, ?) and the interaction between oculomotor control and sentence comprehension (?, ?).[1]

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

The ? (?) model is particularly attractive for studying sentence comprehension because it 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, which does not entail that there are no language specific constraints on sentence comprehension. A further advantage of the ? (?) 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.

In ? (?), the ? (?) architecture was used to model aphasic sentence processing on a small scale, using data from seven patients. They modelled fixations proportions in a visual world task, response accuracies and response times for empirical data of a sentence-picture matching experiment by ? (?). Their goal was to test two of the three hypotheses of sentence comprehension deficits mentioned above, slowed processing and intermittent deficiency.

In the present work, we provide a proof of concept study that goes beyond ? (?) by evaluating the evidence for the three hypotheses—slowed processing, intermittent deficiencies, and resource reduction—using a larger data-set from ? (?) with 56 IWA and 46 matched controls.

Before we describe the modelling carried out in the present paper and the data used for the evaluation, we first introduce the cognitive constraints assumed in the ? (?) 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 (?, ?) from controls and IWA. When fitting individual participants, we vary three parameters that map to the three theoretical proposals mentioned above. The goal was to determine whether the distributions of parameter values 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 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 ? (?) 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.

The ACT-R architecture assumes a distinction between long-term declarative memory and procedural knowledge. The latter is implemented as a set of rules, consisting of condition-action pairs known as production rules. These production rules operate on units of information known as chunks, which are elements in declarative memory that are defined in terms of feature-value specifications. For example, 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. For example, if the noun *book* is to be retrieved, cues such as {part-of-speech nominal, number singular, and animate no} could be used to retrieve it. Production rules are written to trigger such a retrieval event. Retrieval only succeeds if the activation of a to-be-retrieved chunk is above a minimum threshold, which is a parameter in ACT-R.

The activation of a chunk is determined by several constraints. Let C be the set of all chunks in declarative memory. The total activation of a chunk $i \in C$ equals

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

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; the third summand is a penalty for mismatches between a cue value j and the value in the corresponding slot of chunk i; and finally, ε is noise that is logistically distributed, approximating a normal distribution, with location 0 and scale ANS which is related to the variance of the distribution. It is generated at each new retrieval request. The retrieval time T_i of a chunk 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 scale parameter ANS of the logistic distribution from which ε is generated can be interpreted as implementing the *intermittent deficiency* hypothesis because higher values of the scale will tend to lead to more fluctuations in activation of a chunk and therefore higher rates of retrieval failure.[2]

The second summand in (1), representing the process of *spreading activation* within the ACT-R framework, can be made more explicit for the goal buffer and for retrieval cues $j \in \{1, \ldots, J\}$ as

[2] As an aside, note that ? (?) implemented intermittent deficiency using another source of noise in the model (utility noise). In future work, we will compare the relative change in quality of fit when intermittent deficiency is implemented in this way.

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

Here, $W_j = \frac{\mathrm{GA}}{J}$, where GA is the *goal activation* parameter and S_{ji} is a value that increases for each matching retrieval cue. S_{ji} reflects the association between the content of the goal buffer (that can be interpreted as part of the current focus of attention) and the chunk i. The parameter GA determines the total amount of activation that can be allocated for all cues j of the chunk in the goal buffer. It is a free parameter in ACT-R. This parameter, sometimes labelled the "*W* parameter", has already been used to model individual differences in working memory capacity (?, ?). Thus, it can be seen as one way (although by no means the only way) to implement the resource reduction hypothesis. The lower the GA value, the lower the difference in activation between the retrieval target and other chunks. This leads to more retrieval failures and lower differences in retrieval latency.

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

Next, we evaluate whether there is evidence consistent with the claims regarding slowed processing, intermittent deficiency, and resource reduction, when implemented using the parameters described above.

Simulations

In this section we describe our modelling method and the procedure we use for fitting the model results to the empirical data from ? (?).

Materials

We used the data from 56 IWA and 46 matched controls published in ? (?). 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.

Of the 11 sentence types, for the current simulation we chose subject relatives (*The girl who hugged the boy washed the woman*) and object relatives (*The boy who the girl hugged washed the woman*). 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 (?, ?). Second, the ? 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. Since the production rules in the model were designed for modelling unimpaired processing, using them for IWA amounts to assuming that there is no damage to the parsing system per se, but rather that the processing problems in IWA are due to some subset of the cognitive constraints discussed earlier.

Method

For the simulations, we refer to as the parameter space Π the set of all vectors $(\mathrm{GA}, \mathrm{DAT}, \mathrm{ANS})$ with GA, DAT, 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 $\mathrm{ANS} \in \{0.15, 0.2, \ldots, 0.45\}$.[3] 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. This is motivated by the results of ? (?): there, the combined model varying both parameters (default action time (DAT) and utility noise, in their case) achieved the best fit to the data. Including all models allows us to do a similar investigation.

For all participants in the ? (?) data-set, we calculated comprehension question response accuracies, averaged over all items of the subject / object relative clause conditions. 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 made the correct attachment. We counted parsing failures, where the model did not create the target dependency, as incorrect.

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

Cluster analysis In order to investigate the clustering of parameter estimates that we observed in Figure **??**, we also performed a cluster analysis on the data. We used a cluster-

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

ing method on the combined data (i.e., one data-set for subject/object relatives each) to see to which degree controls and IWA could be discriminated. This functions as a test for the impression from Figure **??** that clustering is generally tighter in controls vs. IWA. If this interpretation of the plots is correct, we would 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.

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 1. 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 ? (?) 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 1: Discrimination ability of hierarchical clustering on the combined data for **simple subject/object relative clauses**. Numbers in bold show the number of correctly clustered data points. The bottom row shows the percentage accuracy.

	Subject relatives		Object relatives	
predicted group	controls	IWA	controls	IWA
control	**31**	17	**27**	45
IWA	15	**39**	19	**11**
accuracy	67%	70%	59%	20%

Table 2: Discrimination ability of hierarchical clustering on the combined data for **subject/object relative clauses with reflexives**. The numbers in bold are the correct classifications of controls/IWA. The bottom row shows the percentage accuracy.

Distribution of normal parameter values Table 3 shows the number of participants for which a parameter value outside the default values was predicted. By default values we mean the values GA = 1, DAT = 0.05 (or 50 ms), and ANS = 0.15. 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 surprisingly often. In controls, the main difference between subject and object relatives is a clear increase in elevated noise values in object relatives. For IWA, the single-parameter models in subject relatives are very similar, whereas in object relatives, most IWA (95%) exhibit elevated noise values, while a far smaller proportion (71%) showed reduced goal activation values.

Discussion

The simulations and cluster analysis above demonstrate overall tighter clustering in parameter estimates for controls, and more variance in IWA. This is visible in the scatterplots in Figure **??** and from the cluster analysis results in Table 1. These findings are consistent with the predictions of the small-scale study in ? (?). However, there is considerable variability even in the parameter estimates for controls, more than expected based on the results of ? (?).

The marginal distributions of parameter estimates (Figure **??**) 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 appear to be more pronounced in object relatives vs. subject relatives. This means that all the three hypotheses can be considered viable candidate explanations.

This pattern is also evident in Table 3, where generally more IWA than controls have 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 subject relatives than in object 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 modelling 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 ? (?) 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) We plan to investigate whether an approach as in ? (?), using lognormal race models and mixture models, can be applied to our research question.

Second, the use of accuracies as modelling 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 will implement an approach modelling both accuracies and listening times. Also, counting each parsing failure as 'wrong' might yield overly conservative accuracy values for the model; this will be addressed by

		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 3: Number of participants (out of 46 controls and 56 IWA) 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.

assigning a random component into the calculation. This reflects more closely a participant who guesses if he/she did not fully comprehend the sentence.

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. For example, 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. Also, as both the GA and ANS parameters effect higher failure rates on the surface, it will be worth investigating in future work whether a more focussed source of noise, such as utility noise, may be a better way to model intermittent deficiency.

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 ? (?) data-set.

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.