Reactivation of antecedents by overt versus null pronouns: Evidence from Persian

Niloofar Keshtiari[1] *and Shravan Vasishth*[2]
[1] Cluster of Languages of Emotion, Freie Universität Berlin, Germany
[2] Department of Linguistics, University of Potsdam, Germany
and School of Mathematics and Statistics, University of Sheffield, UK

ABSTRACT

In Persian, a construction exists in which a gap can optionally be replaced by an overt pronoun. A self-paced reading study (110 participants) suggests that the overt pronoun results in deeper encoding (higher activation) of the antecedent noun, presumably because of richer retrieval cue specifications during antecedent retrieval at the pronoun; this higher activation has the consequence that the antecedent is easier to retrieve at a subsequent stage. This provides new evidence for reactivation effects of the type assumed in the cue-based retrieval model of parsing (Lewis and Vasishth 2005), and shows that dependency resolution is not simply a matter of connecting two co-dependents; the retrieval cue specification has a differential impact on processing.

Keywords: sentence comprehension, pronoun processing, reactivation, expectation-based processing

1 INTRODUCTION

It is well known that both overt and null pronouns render their antecedents more active (more salient) in memory (MacDonald 1989; Emmorey and Lillo-Martin 1995). One way to characterize the underlying processes in antecedent–pronoun/gap resolution is in terms of the ACT-R based (Anderson *et al.* 2004) architecture of sentence processing discussed in (Lewis and Vasishth 2005; Lewis *et al.* 2006). The computational model developed in these papers has been widely applied in the study of various phenomena in psycholinguistics (Vasishth *et al.* 2008; Vasishth and Lewis 2006; Patil *et al.* 2012; Reitter *et al.* 2011; Dillon 2011; Boston *et al.* 2011; Patil *et al.* 2013; Dillon *et al.* 2011; Engelmann *et al.* 2013).

A central assumption in the ACT-R architecture is that, in any information-processing task, memory representations must be associated with each other in order to build a mental representation that allows the task to be carried out. In the context of sentence comprehension, the primary events that are modeled are structure building and dependency resolution. All other things being equal, the speed with which a memory representation can be accessed depends on its activation level (this is an abstract, unitless quantity) and on the retrieval cues (these are essentially feature–value matrices) that guide access. Generally speaking, the higher the activation, the faster the retrieval. For example, when the parser encounters a reflexive like *himself*, an antecedent noun may be searched using the fact that the antecedent must c-command the reflexive (one retrieval cue; see Dillon *et al.* 2011) or using an additional cue, here gender (Patil *et al.* 2012). Activation of memory representations is assumed to be undergoing constant decay; this models forgetting over time. An assumption in ACT-R is that decay can be counteracted by a process of reactivation: every retrieval event is assumed to increase the activation of the item retrieved. Such an increase in activation has the obvious consequence of facilitating subsequent retrieval (unless enough time goes by such that decay levels out the activation). Previous work has addressed some of the empirical consequences of this theoretical assumption. For example, in Hindi, processing a verb in a relative clause has been argued to be easier when the relative clause is long vs. short; under the assumption that a long relative clause repeatedly accesses and modifies the head noun and does so more often than a shorter relative clause, we expect a faster reading time when the head noun must be accessed, for example, while processing the verb of the relative clause. This was one of the arguments presented by Vasishth and Lewis (2006) in order to explain faster reading times observed at the verb of the relative clause in long vs. short relative clauses (cf. Levy 2008; Husain *et al.* 2013).

Given such an architecture, it is reasonable to assume that completing a dependency between an antecedent and a pronoun, or between an antecedent and a gap, will increase activation of the antecedent, making subsequent retrieval easier; this assumes that the decay component has not had enough time to counteract the effect of such an activation increase. Indeed, Bever and McElree (1988) have shown experimentally that such reactivation occurs with gaps and pronouns, and that "gaps access their antecedents during comprehension in the same way as pronouns."

Persian presents an interesting construction in this context. Sentences such as in Example (1) have the property that the first gap can optionally be an overt or null pronoun (Taghvaipour 2004). For example, consider (1a); here, two gaps are present. As shown in (1b), the first one can be replaced with the pronoun *un ro*, 'it DOM').[1]

(1) a. Nazanin [in ketabcha ro]$_i$ [$_{CP}$ ghablaz inke *gap*$_i$
Nazanin this booklet DOM before that gap
be-khun-eh] *gap*$_i$ be man dad, dorost hamoon
prefix-read-3SG gap to me gave.3SG, just that
moghe ke kelas ha tamoom shod.
moment that class PL finish became.3SG.

(Lit.) 'Nazanin gave me this little book before reading (it), when the classes finished.'

b. Nazanin [in ketabcha ro]$_i$ [$_{CP}$ ghablaz inke **unro**
Nazanin this booklet DOM before that it-DOM
be-khun-eh] *gap*$_i$ be man dad, dorost hamoon
prefix-read-3SG to me gave.3SG, just that moment
moghe ke kelas ha tamoom shod.
that class PL finish became.3SG.

'Nazanin gave me this little book before reading it, when the classes finished.'

This construction is interesting in the context of reactivation effects in parsing because it allows us to investigate whether there is a difference in activation increase due to antecedent–pronoun vs. antecedent–gap dependency resolution. Our study was motivated by the speculation that there might be a difference in the way a anteced-

[1] Abbreviations used are as follows. DOM: direct object marking; 3SG: third singular; PL: plural.

ent–gap and antecedent–pronoun dependency is completed: the antecedent might be activated to a greater extent in the antecedent–pronoun case vs. the antecedent–gap case. This could happen because the pronoun uses a richer set of cues; for example, pronouns provide number information, whereas gaps do not. Another possibility is that the pronoun may focus the antecedent (thereby encoding it more richly) in a way that the gap does not. These two explanations may be related: a richer set of cues would lead to better encoding due to the greater extent of activation increase.

We investigated whether we could find any evidence for differential amounts of activation increase in the above construction. We employed the self-paced reading methodology (Just *et al.* 1982) described below. In (1a), the word *ketabche*,[2] 'booklet', is co-indexed with a gap; this gap presumably activates the antecedent once the dependency is completed. Due to reactivation effects, the activation increase of *ketabche* should increase the speed or rate at which its retrieval is completed subsequently at the verb *dad*, 'gave.' Example (1b) is identical except that instead of the gap we have an overt pronoun *unro*. Our speculation was that this might boost activation of *ketabche* to a greater extent than the gap, facilitating retrieval at the verb.[3]

In order to understand the role of the overt pronoun, we compared sentences with null and overt pronouns, as shown in (2a,b) (the front-slashes in the examples represent the partitioning of the segments in the self-paced reading task; this is described below). We were also interested in exploring, in the same experiment, a related kind of reactivation effect: modification of the noun *ketabche*, 'booklet', by a relative clause. As mentioned above, it has been argued (Vasishth and Lewis 2006; Hofmeister 2011; Vasishth *et al.* 2012) that modification of a noun increases its activation, making subsequent retrieval easier (cf. Levy 2008 for an alternative explanation in terms of expectations).

[2] In (1a,b), the direct object marker *ro* induces a sound change on the word it modifies, changing *ketabche* to *ketabcha*.

[3] Note that the exact location of the gap before the verb *dad*, 'gave' is not important; even assuming that there is a gap there is not necessary. All that we need to assume is that a dependency must be completed between the noun *ketabche* and the verb in order to determine who did what to whom. This assumption of a dependency resolution requirement is well-motivated by previous work; see, e.g., Gibson (2000); Bartek *et al.* (2011); Vasishth *et al.* (2008).

It follows that relative clause modification should also increase activation of the noun, resulting in faster retrieval of the noun at the verb. We were interested in determining whether we find facilitation at the verb due to relative clause interposition (2c,d); if yes, this would provide new evidence for the proposal in the literature that modification increases activation of the modified element.

An alternative possibility is greater processing difficulty at the verb in the relative clause conditions; this effect has been found by Grodner and Gibson (2005) for English (also see Bartek *et al.* 2011), and could be a consequence of increased distance between the verb and its argument(s). Such locality effects can be explained in terms of distance as defined in the Dependency Locality Theory (Gibson 2000) or in terms of decay and interference (Lewis and Vasishth 2005; Van Dyke and McElree 2006).

(2) a. Nazanin / [in ketabcha ro]$_i$ / [$_{CP}$ ghablaz inke /
Nazanin this booklet DOM before that
gap_i be-khun-eh] / gap_i be man dad / ,
gap prefix-read-3SG gap to me gave.3SG ,
dorost hamoon moghe ke / kelas ha / tamoom
just that moment that class PL finish
shod / .
became.3SG .

‘Nazanin gave me this little book before reading (it), when the classes finished.’

b. Nazanin / [in ketabcha ro]$_i$ / [$_{CP}$ ghablaz inke /
Nazanin this booklet DOM before that
unro be-khun-eh] / gap_i be man dad / ,
it-DOM prefix-read-3SG gap to me gave.3SG ,
dorost hamoon moghe ke / kelas ha / tamoom
just that moment that class PL finish
shod / .
became.3SG .

’Nazanin gave me this little book before reading it, when the classes finished.’

c. Nazanin / [in ketabcha ro /]$_i$ [**ke hafte pish** /
Nazanin this booklet DOM that week last

kharid-eh **bood** /] ghablaz inke / *gap*$_i$
bought-3SG.PC was before this gap
be-khun-eh / *gap*$_i$ be man dad / , dorost
prefix-read-3SG gap to me gave.3SG , just
hamoon moghe ke / kelas ha / tamoom
that moment that class PL finish
shod / .
became.3SG .

'Nazanin gave me this little book which she has bought last week, before reading (it), when the classes finished.'

d. Nazanin / [in ketabcha ro]$_i$ / [**ke hafte pish** /
Nazanin this booklet DOM that week last
kharid-eh **bood**] / ghablaz inke / **unro**
bought-3SG.PC was before that it-DOM
be-khun-eh / *gap*$_i$ be man dad / , dorost
prefix-read-3SG gap to me gave.3SG , just
hamoon moghe ke / kelas ha / tamoom
that moment that class PL finish
shod / .
became.3SG .

'Nazanin gave me this little book which she has bought last week, before reading it, when the classes finished.'

Thus, our predictions are: the presence of overt pronoun interposition should result in a facilitation in processing at the main verb (compared to the conditions where a gap is present); modifying the antecedent with a relative clause could show a facilitation due to reactivation, or increased difficulty due to locality effects. We had no predictions about whether there would be an interaction between the pronoun/gap and relative clause factors. The results of the study are presented next.

2 EXPERIMENT

2.1 *Method: Participants, stimuli and fillers, procedure*

One hundred and ten native speakers of Persian, all living in Tehran, participated in the experiment in August and September 2009. Since we had no access to a laboratory in Tehran, the first author visited the participants at their homes and carried out the experiment there. Participants were asked to complete a questionnaire on their educational background, language background, and average reading time per day. The questionnaire and items are available from the second author. Participants' ages ranged from 18 to 75 years, with mean age 34.6 years. Each participant was paid the Iranian-Rial equivalent of five Euros.

A total of 161 Persian sentences (5 examples, 60 fillers, 96 stimulus sentences) were prepared by the first author. The 96 stimulus sentences were designed as follows: Following standard experimental methodology for repeated measures (within-subjects) designs, twenty-four stimuli sentences were prepared, and four versions of each sentence were constructed; these correspond to the four conditions in the experiment (see Example 2). Each version of the twenty-four sentences was assigned to one of four lists; that is, each list contains only one of the four versions of a stimulus sentence. Because each participant is shown items from only one list, they read (apart from the fillers and examples) a total of twenty-four target sentences, each representing one condition in the experiment design. This has the consequence that, for example, a subject exposed to List 1 would see Sentences 1a, 2b, 3c, 4d, 5a,...; and a subject exposed to List 2 would see Sentences 1b, 2c, 3d, 4a, 5b,...This partitioning into lists is commonly referred to as counterbalancing and serves to minimize bias introduced by any one stimulus sentence.

Thus, each participant saw $5 + 60 + 24 = 89$ sentences. Because reading time generally increases at the end of a sentence (so-called sentence final wrap-up effects, thought to reflect higher-level integration processes that are triggered after a sentence is read), we added an adverbial phrase to the end of each of the stimuli sentences. In addition, at the end of each sentence a period was presented after pressing the space bar as a separate final segment. This extra material at the end of the sentence makes it less likely that our critical region (the verb of the main clause) is contaminated by higher-level sentence-final processing effects.

The experiment began by the first author explaining the task to each participant verbally; then, the five practice sentences were presented, and these were followed by the actual experiment (fillers and stimulus sentences, pseudo-randomly ordered). Participants pressed the spacebar (marked with a star) to reveal each successive segment; every time the space bar was pressed, the previous segment would disappear and the next segment would appear in the center of the screen. The time the participants spent reading each segment was recorded as the time between key presses. The segmentation is shown in Example 2 and in the items file provided as supplementary material with this paper.

The experiment was run using Linger version 2.88 by Douglas Rohde on a laptop.[4] Participants were asked to read at a pace that was normal for them. A true/false statement was presented after each sentence; this was meant to ensure that subjects were attending to the sentences and not just pressing the space bar without reading. In order to prevent subjects from developing a strategy for answering the true/false statements without completely parsing the sentence, the statements were directed at every part of the previous sentence, including the noun and the verb of both the main clause and the relative clause. These true/false statements were balanced in their yes–no responses. No feedback was given for correct/incorrect responses. For examples, see the items file provided as supplementary information with this paper.

Participants took approximately 30 minutes to complete the experiment. Reading time at the verb of the main clause (in milliseconds) was taken as a measure of relative momentary processing difficulty. In the following section, the results of the study are reported and discussed.

2.2 *Results*

We fit linear mixed models using the package `lme4` (Bates and Sarkar 2007) in the R programming environment (R Development Core Team 2006); see the appendix for some background on the statistical models used here. The reader unfamiliar with psycholinguistic data analysis methods would benefit from reading the appendix before proceeding with the present section. For the critical analyses at the verb and the region following it, we used JAGS (Plummer 2010) to also fit a hierarchical Bayesian model (a linear mixed model) using non-informative

[4] See `http://tedlab.mit.edu/~dr/Linger/`.

priors. For brevity, the Bayesian analysis is omitted from the main paper, but is included in the supplementary material.

The response variables were response accuracy and response time, and reading times at the verb of the main clause (hereafter, critical region) and the spillover region (hereafter, post-critical region).

Due to an error in the design, nine items (labeled 7, 12, 14, 15, 17, 20, 22, 23, 24 in the supplementary material) were removed from the analysis. In these items, the argument that was co-indexed with the pronoun/gap was either not modified by the relative clause, or the pronoun was not co-indexed with the correct antecedent. This reduced the original data by about 40%.

The response time and reading time data (both in milliseconds) were transformed to a negative reciprocal ($-\frac{1000}{rt}$) in order to stabilize variance; the choice of transform was determined using the Box-Cox procedure (Box and Cox 1964; Venables and Ripley 2002). The reciprocal transform converts speed to rate; see (Kliegl *et al.* 2010) for further discussion on the use of this transform for reading time data.

In the reading times, the transform revealed some extremely fast values (0.7% of the data) that dramatically affected the residuals; these were a few values that were 200–250 ms long. Although such reading times cannot in general be categorized as "too fast", in the context of the present experiment they are not representative of the reading time distributions (based on our experience in our own lab and elsewhere, in languages like Hindi, German, and Japanese, we also see remarkably slow reading times compared to English). These extreme values were removed in the final analysis.

All data and R code associated with the analyses presented here are provided as supplementary material with this paper.

2.2.1 Response accuracy and response time

Response accuracy and response time and their analyses are summarized in Tables 1 and 2. For accuracy, a generalized linear mixed model was fit using the binomial link function to evaluate the effect of pronoun (pron), the effect of relative clause insertion (RC), and the interaction of these two factors (see the appendix for more detail on generalized linear models). A standard ANOVA contrast coding was used: the factor pron was coded −0.5 for the gapped conditions (Examples 2a,c) and 0.5 for the pronoun conditions (Examples 2b,d); the factor RC was coded −0.5 for the −RC conditions (Examples 2a,b), and 0.5 for the +RC conditions (Examples 2c,d). Items and participants were included as crossed random factors (crossed varying intercepts; see Appendix for discussion). Consistent with the predictions of the locality-based accounts discussed earlier, we found significantly lower accuracies in the conditions where the relative clause was present; these conditions also had longer response times. No other effect reached statistical significance.

Table 1:
Mean question accuracy (percentages) and negative reciprocal response time (abbreviated as response rate)

	−RC gap	−RC pronoun	+RC gap	+RC pronoun
accuracy	87	86	78	81
response rate	−0.22	−0.22	−0.21	−0.21

Table 2:
Summary of the effects of pronoun (pron), relative clause insertion (RC), and the pron×RC interaction on response accuracy

contrast	coef	se	z	p
pron	0.04	0.14	0.29	n.s.
RC	−0.48	0.15	−3.3*	<0.01
pron×RC	0.12	0.15	0.85	n.s.

Table 3:
Summary of the effects of pronoun (pron), relative clause insertion (RC), and the pron×RC interaction on negative reciprocal response time

contrast	coef	se	t
pron	−0.00	0.004	−0.9
RC	0.02	0.004	3.7*
pron×RC	−0.00	0.004	−0.6

2.2.2 Analyses of reading times

The negative mean reciprocal reading times ($-s^{-1}$) with 95% confidence intervals are summarized in Figure 1. The results of the statistical analyses are shown in Table 4. The linear mixed models had varying intercepts and slopes for subject and item, and varying slopes by subject for pron, RC, and the pron×RC interaction.

Analyses at the critical region (the main verb) showed a marginally significant main effect of pron: the overt pronoun resulted in faster

region	contrast	coef	se	t
critical	pron	−0.05	0.03	−1.89
	RC	−0.04	0.03	−1.58
	pron×RC	−0.03	0.03	−1.02
post-critical	pron	−0.07	0.03	−2.62*
	RC	0.008	0.03	0.30
	pron×RC	−0.03	0.03	−1.10

Table 4:
Summary of planned comparisons in the linear mixed models analysis

reading times at the verb, as predicted by the reactivation hypothesis. In addition, there was only a marginal effect of relative clause interposition: reading time (rather, reading rate) was marginally faster at the verb in the RC conditions. The interaction between the factors pron and RC did not reach statistical significance either. Figure 1 suggests that the marginally significant facilitation at the critical region due to the overt pronoun is driven by the RC conditions (2c,d). This was confirmed in a post-hoc analysis where the effect of pronoun was investigated within the −RC and +RC conditions. Table 5 summarizes these analyses (the third contrast in Table 5, the effect of RC, is redundant since this was already investigated in our planned comparisons shown in Table 4, but is included because we wanted to use all three degrees of freedom available for parameter estimation of fixed effects).

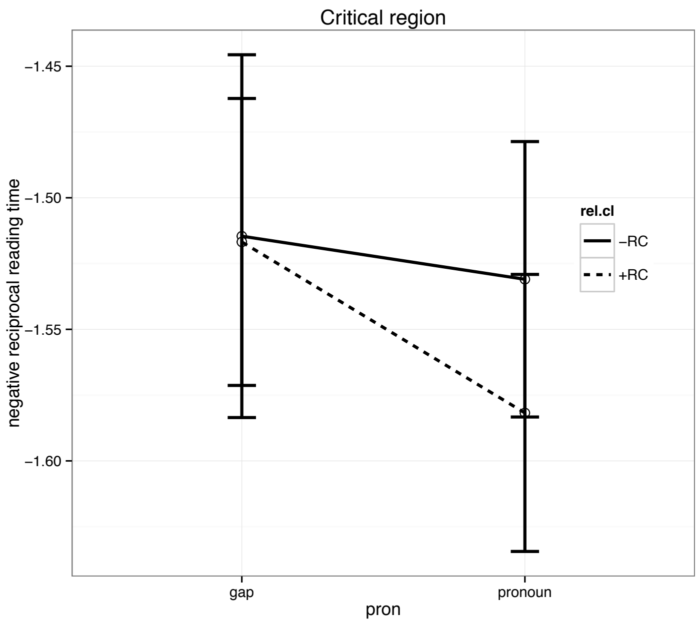
The post-critical region showed an effect of pron: reading rate was faster when the pronoun was present. The post-hoc analysis showed that the effect of pron was present in both the −RC and the +RC conditions with approximately the same magnitude.[5]

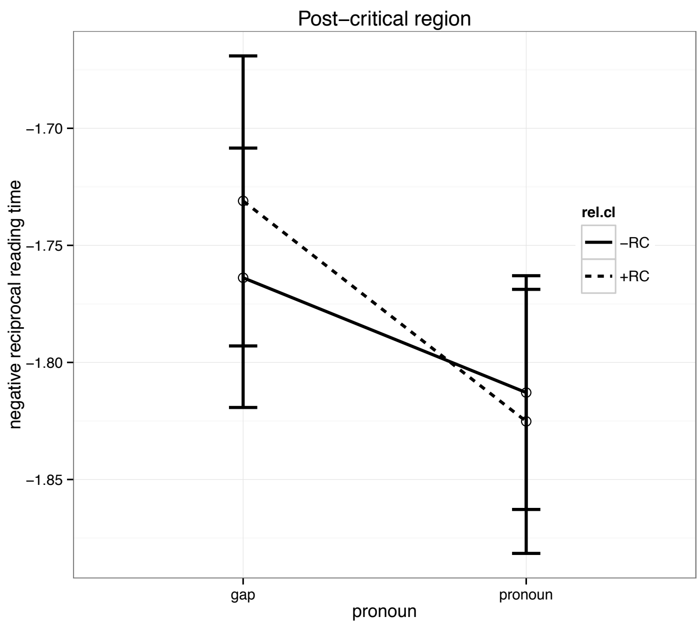
2.2.3 Discussion

To summarize the results for response accuracy and response time, we find lower accuracies and longer response times when the relative

[5] As an aside, note that in psycholinguistics analyses are conventionally carried out on raw reading times. Had we done this conventional analysis for the present data, we would have reported null results. However, in the model based on such untransformed data, the normality assumption for residuals and the homoscedasticity assumption are not met; this would make the model output based on raw reading times meaningless for purposes of statistical inference. The source code provided with this paper gives more detail.

Figure 1: Negative reciprocal reading times at the critical and post-critical regions





region	contrast	coef	se	t
critical	pron (−RC)	−0.02	0.04	−0.63
	pron (+RC)	−0.08	0.04	−2.03*
	RC	−0.04	0.03	−1.58
post-critical	pron (−RC)	−0.04	0.04	−0.94
	pron (+RC)	−0.10	0.04	−2.43*
	RC	0.008	0.03	0.3

Table 5:
Summary of post-hoc nested contrasts at the critical and post-critical regions

clause is present. Response accuracy and response time showed no effect of pronoun and no interaction between the pronoun and relative clause conditions. In the reading time data, at the post-critical region (the word following the verb), reading rate is faster if the pronoun (*unro*) is present. A post-hoc analysis revealed that the facilitation due to the pronoun was driven by the relative clause conditions. At the verb, there is a slight facilitation due to relative clause interposition, but this does not reach statistical significance. No interactions were found.

Although response accuracy and response time are of secondary interest to the research question, they do show that interposing a relative clause renders the sentence more difficult to process in later stages of processing. This finding is partly consistent with locality accounts such as Dependency Locality Theory (Gibson 2000) and the cue-based retrieval architecture (Lewis and Vasishth 2005), both of which predict increased integration cost at the verb when the distance between the subject of the sentence (*Nazanin*) and the main verb (*dad*) is increased. This increased distance (or increased syntactic complexity) could plausibly make it more difficult to retain an accurate representation of the sentence meaning in memory in order to respond to the question.

The Dependency Locality Theory and the cue-based retrieval account, however, also predict a slowdown at the verb in reading times; this prediction turns out to be incorrect because relative clause interposition at the verb results in a marginal facilitation. This tendency towards a facilitation makes sense given prior findings; it can be explained in terms of reactivation due to relative clause modification (Vasishth and Lewis 2006; Hofmeister 2011; Vasishth *et al.* 2012), as discussed earlier. An alternative explanation for the relative-clause facilitation effect lies in expectation-based processing (Levy 2008). Assuming that treebank-based distributions in Persian turn out to be similar to the distributions in languages that Levy examined, the proposal would be that the expectation for a verb gets stronger and stronger if the appearance of the verb is delayed – this is the situation when the relative clause is interposed. In the relative clause conditions, by the time the verb is encountered, it is highly expected compared to the non-relative clause conditions. This expectation-based account has been proposed as an alternative to the reactivation account. Both explanations are plausible, but the expectation-based account's key prediction has been falsified by Levy *et al.* (2013): they showed in a series of experiments that in Russian relative clauses, if the verb's appearance is delayed inside a relative clause, there is a slow-down at the verb, not a speed-up. The evidence from Russian relative clauses is therefore strongly in favor of locality based explanations. In any case, in our study, the reactivation-based explanation seems more plausible, as discussed in connection with the pronoun results below.

Next, we turn to the main research question in this paper, the effect of the pronoun/gap manipulation. At the verb and the region following it, the pronoun conditions have a faster reading rate than the gap conditions. This suggests that completing the antecedent–pronoun dependency results in higher activation of the antecedent compared to the antecedent–gap case; as a consequence, the antecedent of the pronoun is retrieved faster at the verb. This facilitation probably spills over to the word following the verb. In sum, the data are consistent with our original speculation: replacing a gap with a pronoun appears to increase the activation of the antecedent, making it easier to retrieve at a subsequent stage.

Why was the facilitation effect due to the pronoun driven by the relative clause conditions? In the non-relative clause conditions, even though the pronoun may be activating the antecedent more than the gap, decay of the antecedent noun might be setting in by the time the verb is encountered. By contrast, in the relative clause conditions, the reactivation of the antecedent noun by the relative clause (which modifies this noun) could be providing a counteracting activation boost that reverses the effect of decay. If this is correct, the reactivation account may be the correct explanation for the relative clause facilitation effect discussed above. Naturally, this conclusion does not challenge the expectation-based explanation per se, which probably also plays a role in sentence processing; there is considerable evidence for expectation-based effects (see, e.g., Husain *et al.* (2013) for new evidence from Hindi), and these effects cannot be explained in terms of reactivation.

Returning to our main finding, that the pronoun increases activation of the antecedent, our results raise the question: what is it about antecedent–pronoun vs. antecedent–gap dependencies that results in a more robust encoding or higher activation of the antecedent? One explanation may be that the pronoun may be focusing the antecedent; another may be that a richer set of retrieval cues is used to complete the antecedent–pronoun dependency. We do not have a clear answer for the underlying reason; but given the present data, it seems clear that pronouns activate the antecedent to a greater degree than gaps do.

Why is it that accuracy scores are lower for the relative clause conditions, but in reading times the relative clause conditions result in a (marginal) facilitation at the verb? The former supports the locality account but the latter directly contradicts it. One possible explanation lies in the relative timing of retrieval and expectation effects: locality costs, which reflect retrieval difficulty, may be appearing at a later stage during parsing, whereas expectation-based effects appear earlier. Vasishth and Drenhaus (2011) have proposed this in the context of German. One problem with this account is that locality effects have been found in reading times in English (Grodner and Gibson 2005; Bartek *et al.* 2011) and most recently Hindi (Husain *et al.* 2013) and Hungarian (Kovács and Vasishth 2013). Perhaps a more plausible explanation is that locality effects are longer lasting than expectation effects: the former show effects in online as well as offline measures, but expectation only shows effects in online measures. Under this view, locality effects could have been masked by or are much weaker than reactivation and/or expectation effects while processing the verb; in the offline question–response stage, only locality effects remain visible. All the above explanations are speculative and need to be investigated in new studies pitting locality against expectation.

A broader consequence of the pronoun-driven facilitation we report is that the notion of dependency resolution in parsing needs to be made more precise. It is widely assumed, implicitly or explicitly, that dependency resolution is simply a matter of connection to elements subject to certain constraints, such as locality (Gibson 2000). But completing an antecedent–gap dependency and an antecedent–pronoun dependency cannot be only a function of locality; it matters which retrieval cues are deployed in retrieval. This has important implications for theories of parsing: an architecture driven by retrieval cues seems to be better motivated than one that ignores the nature of the cue.

3 AUTHOR NOTE

We are grateful to Felix Engelmann for assisting with setting up the software for running the experiments, and to Philip Hofmeister for comments on a draft. Colin Philips and Brian Dillon provided very detailed and useful comments on the paper and have contributed to improving it considerably. We are also grateful to the anonymous reviewers for their careful, high quality reviews. Dr. Jeremy Oakley of the School of Mathematics and Statistics, University of Sheffield, provided very useful guidance on fitting linear mixed models and Bayesian models, but we are of course responsible for any errors in this paper. This paper is based on the master's thesis of Niloofar Keshtiari in the MSc program European Masters in Clinical Linguistics, University of Potsdam. The experiment design is due to Shravan Vasishth; the experiment items were prepared by Niloofar Keshtiari and the experiment was conducted by her in Tehran. The data analysis and the majority of the text in the paper (except the materials and methods section) were written by Shravan Vasishth. The corresponding author for this paper is Shravan Vasishth (Email: vasishth@uni-potsdam.de).

REFERENCES

John R. ANDERSON, Dan BOTHELL, Michael D. BYRNE, Scott DOUGLASS, Christian LEBIERE, and Yulin QIN (2004), An integrated theory of the mind, *Psychological Review*, 111(4):1036–1060.

Brian BARTEK, Richard L. LEWIS, Shravan VASISHTH, and Mason SMITH (2011), In search of on-line locality effects in sentence comprehension, *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(5):1178–1198.

Douglas BATES and Deepayan SARKAR (2007), *lme4: Linear mixed-effects models using S4 classes*, R package version 0.9975-11.

Thomas G. BEVER and Brian MCELREE (1988), Empty categories access their antecedents during comprehension, *Linguistic Inquiry*, 19(1):35–43.

Marisa F. BOSTON, John T. HALE, Shravan VASISHTH, and Reinhold KLIEGL (2011), Parallel processing and sentence comprehension difficulty, *Language and Cognitive Processes*, 26(3):301–349.

George E.P. BOX and David R. COX (1964), An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)*, 26(2):211–252.

Brian DILLON, Alan MISHLER, Shayne SLOGGETT, and Colin PHILLIPS (2011), A computational cognitive model of syntactic priming, *Journal of Memory and Language*, 69(4):85–103.

Brian W. DILLON (2011), *Structured access in sentence comprehension*, Ph.D. thesis, University of Maryland.

Karen EMMOREY and Diane LILLO-MARTIN (1995), Processing spatial anaphora: Referent reactivation with overt and null pronouns in American Sign Language, *Language and Cognitive Processes*, 10(6):631–653.

Felix ENGELMANN, Shravan VASISHTH, Ralf ENGBERT, and Reinhold KLIEGL (2013), A framework for modeling the interaction of syntactic processing and eye movement control, *Topics in Cognitive Science*, 5(3):452–474.

Edward GIBSON (2000), Dependency Locality Theory: A distance-based theory of linguistic complexity, in Alec MARANTZ, Yasushi MIYASHITA, and Wayne O'NEIL, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, MIT Press, Cambridge, MA.

Daniel GRODNER and Edward GIBSON (2005), Consequences of the serial nature of linguistic input, *Cognitive Science*, 29:261–290.

Philip HOFMEISTER (2011), Representational complexity and memory retrieval in language comprehension, *Language and Cognitive Processes*, 26(3):376–405.

Samar HUSAIN, Shravan VASISHTH, and Narayanan SRINIVASAN (2013), Locality effects depend on processing load and expectation strength, in *Proceedings of the Conference on Architectures and Mechanisms for Language Processing*, p. 96, Aix-Marseille Université, Marseille, France.

Marcel A. JUST, Patricia A. CARPENTER, and Jacqueline D. WOOLLEY (1982), Paradigms and processes in reading comprehension, *Journal of Experimental Psychology: General*, 111(2):228–238.

Reinhold KLIEGL, Michael E.J. MASSON, and Eike M. RICHTER (2010), A linear mixed model analysis of masked repetition priming, *Visual Cognition*, 18(5):655–681.

Nóra KOVÁCS and Shravan VASISHTH (2013), The processing of relative clauses in Hungarian, in Cheryl FRENCK-MESTRE, F-Xavier ALARIO, Noël NGUYEN, Philippe BLACHE, and Christine MEUNIER, editors, *Proceedings of the Conference on Architectures and Mechanisms for Language Processing*, p. 13, Aix-Marseille Université, Marseille.

Roger LEVY (2008), Expectation-based syntactic comprehension, *Cognition*, 106:1126–1177.

Roger LEVY, Evelina FEDORENKO, and Edward GIBSON (2013), The syntactic complexity of Russian relative clauses, *Journal of Memory and Language*, 69(4):461–495.

Richard L. LEWIS and Shravan VASISHTH (2005), An activation-based model of sentence processing as skilled memory retrieval, *Cognitive Science*, 29:1–45.

Richard L. LEWIS, Shravan VASISHTH, and Julie VAN DYKE (2006), Computational principles of working memory in sentence comprehension, *Trends in Cognitive Sciences*, 10(10):447–454.

Maryellen C. MACDONALD (1989), Priming effects from gaps to antecedents, *Language and Cognitive Processes*, 4(1):35–56.

Umesh PATIL, Sandra HANNE, Frank BURCHERT, Ria De BLESER, and Shravan VASISHTH (2013), Sentence comprehension in aphasia: A computational evaluation of representational and processing accounts, manuscript, accepted pending revision, Cognitive Science.

Umesh PATIL, Shravan VASISHTH, and Richard L. LEWIS (2012), Retrieval interference in syntactic processing: The case of reflexive binding in English, manuscript.

Martin PLUMMER (2010), JAGS Version 2.2.0 user manual.

R DEVELOPMENT CORE TEAM (2006), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org, ISBN 3-900051-07-0.

David REITTER, Frank KELLER, and Johanna D. MOORE (2011), A computational cognitive model of syntactic priming, *Cognitive Science*, 35(4):587–637.

Mehran TAGHVAIPOUR (2004), An HPSG analysis of Persian relative clauses, in Stefan MÜLLER, editor, *Proceedings of the HPSG-2004 Conference, Center for Computational Linguistics, Katholieke Universiteit Leuven*, pp. 274–293, CSLI Publications, Stanford.

Julie VAN DYKE and Brian MCELREE (2006), Retrieval interference in sentence comprehension, *Journal of Memory and Language*, 55:157–166.

Shravan VASISHTH, Sven BRUESSOW, Richard L. LEWIS, and Heiner DRENHAUS (2008), Processing polarity: How the ungrammatical intrudes on the grammatical, *Cognitive Science*, 32(4).

Shravan Vasishth and Heiner Drenhaus (2011), Locality in German, *Dialogue and Discourse*, 1:59–82, http://elanguage.net/journals/index.php/dad/article/view/615.

Shravan Vasishth and Richard L. Lewis (2006), Argument–head distance and processing complexity: Explaining both locality and antilocality effects, *Language*, 82(4):767–794.

Shravan Vasishth, Rukshin Shaher, and Narayanan Srinivasan (2012), The role of clefting, word order and given–new ordering in sentence comprehension: Evidence from Hindi, *Journal of South Asian Linguistics*, 5:35–56.

William N. Venables and Brian D. Ripley (2002), *Modern Applied Statistics with S-PLUS*, Springer, New York.

APPENDIX: A NOTE ON THE STATISTICAL METHODS USED

Here, we summarize the statistical methods used in this paper.

Experimental designs such as the present one are generally referred to as repeated measures or within-subjects designs; this refers to the fact that each participant is exposed to each level of each factor in the experiment design (in our design, we have four factors). The logic of the experiment in general is that our dependent variable or DV (this could be accuracy, measured for example as a percentage, or reading times in milliseconds; or a transformation of these values – see below) is expected to be a linear function of the predictors, which are the factors of our experiment:

$$DV \propto predictors \tag{1}$$

The central idea is that the observed data (the dependent variable, DV) is generated by an underlying statistical model with unknown parameters θ. Formally, DV is a random variable with a particular probability density/distribution function (PDF) associated with it; the PDF is specified in terms of the parameters θ. The goal of statistical analysis is to obtain estimates of these parameters and to draw inferences from these estimates (as discussed below). For example, our pronoun vs. gap manipulation tests the prediction that the presence of the pronoun will result in shorter reading time at the verb. In the linear model setting, this amounts to the claim that there exists some point value (an unknown parameter) with a particular sign (positive or negative) that represents the mean speed-up due to pronoun insertion. Generalizing this, we define a statistical model as shown below:

$$DV_i = \beta_0 + \beta_1 Pronoun_i + \beta_2 RC_i + \epsilon_i \tag{2}$$

Arranging the data in an arbitrary but fixed ordering, let DV_i represent the i-th dependent variable, for example, reading time at the verb in one of the four conditions from each of the participants. Note that each participant would have seen 24 sentences in our experiment, with six instances each of the four levels of the experiment (Pronouns present vs. absent $\times$ Relative Clause (RC) present vs. absent). This is where the term repeated measures comes from: we have multiple measurements from each group (here, subject).

According to the statistical model above, the i-th DV_i is assumed to be generated by a random component, ϵ, and a systematic component (the rest of the right hand side in the above equation). Here, ϵ_i is assumed to be generated by a normal distribution with mean zero and standard deviation σ (yet another unknown parameter which is estimated from the data). We write this compactly as $\epsilon \sim N(0, \sigma)$.[6] The variables Pronoun and RC are indicator variables; for example, when the pronoun is present, the variable Pronoun could be coded as 1, otherwise 0. Similarly, when the relative clause is present, the indicator variable RC could be coded as 1, otherwise 0. Thus, in the statistical model, β_0 is the mean reading time of the gap no-relative clause condition. Setting up indicator variables in this manner is called contrast coding, and the example of contrast coding given here is called treatment contrasts. Different contrast codings are possible; each reflects the theoretical question to be studied (in this paper, we use an anova-style contrast coding and a nested contrast coding; see main text). In the above example, we expect the parameter β_1 to be negative; this reflects our prediction that pronoun insertion will facilitate processing. We can state this as a hypothesis test; we could write that our null hypothesis H_0 is:

$$H_0 : \beta_1 = 0 \tag{3}$$

Standard statistical theory then attempts to obtain, given the data, the best estimate, for example, a maximum likelihood estimate for β_1, call it $\hat{\beta}_1$, along with an uncertainty measure for the estimate, the standard error. Note that β_1 is the true (unknown) difference between the gap and pronoun conditions; and $\hat{\beta}_1$ is the difference in the mean reading times for the gap vs. pronoun conditions. The goal of the frequentist hypothesis testing procedure is to determine, given the null hypothesis above, the probability of obtaining an absolute value of $\hat{\beta}_1$ or something more extreme. This probability, called a p-value, is the conditional probability of the data given a particular hypothesized parameter value (in the example above, $\beta_1 = 0$): we can write it as P(data | parameter).

[6] One typically defines the normal distribution in terms of variance, σ^2, but we can simply talk about standard deviation here.

A convention widely accepted in psycholinguistics is that a p-value of less than 0.05 gives grounds for rejecting the null hypothesis and accepting the alternative hypothesis. The value 0.05 is related to the fact that we conventionally fix the probability of incorrectly rejecting the null hypothesis (i.e., rejecting it when it is actually true) to 0.05. This quantity is called Type I error probability. Note here that if we fit k separate statistical models, the Type I error probability increases, and a correction is needed in order to retain an overall Type I error probability of 0.05. One popular one is the Bonferroni correction: the corrected Type I error is $0.05/k$.

The above statistical model includes an important assumption, namely that the DV_i are independent and identically distributed. The independence requirement means that each data point i is assumed to be independent from all others – this assumption is implausible when the same participant is delivering 24 data points. The other requirement, identical distribution, is that all DVs are assumed to come from a normal distribution with the same variance. Linear mixed models have been developed to address the fact that the DVs are not independent; these models estimate between-subject variance (or equivalently, within-subject covariance) in addition to the σ mentioned above, thereby taking the dependency within each subject's responses into account. The above linear model can be expanded quite easily to a linear mixed model:

$$DV_i = \beta_0 + b_i + \beta_1 Pronoun_i + \beta_2 RC_i + \epsilon_i \tag{4}$$

Here we have an extra term, b_i, called a varying intercept term, which represents each subject's adjustment to the baseline reading time β_0: subjects who are faster than average would have negative b_i, and those slower than average, positive b_i. These b_i are not estimates; they are best linear unbiased predictors (BLUPs) generated after between-subject variance has been estimated. Thus, the above model assumes that $\epsilon \sim N(0, \sigma)$ but also that $b_i \sim N(0, \sigma_{b1})$, where σ_{b1} is the between-subject variance (or, equivalently, within-subject covariance). For our purposes, subject-level variance is a nuisance variable, and we only need to take it into account in the model; our principal interest remains focused on the estimates of the coefficients β_1 and β_2, and in hypothesis tests associated with these coefficients.

Note that in linear mixed models p-values are difficult to compute for various technical reasons, but an absolute t-value greater than 2 can reasonably be assumed to be statistically significant at Type I error probability 0.05.

Two extensions of the above model are as follows. Apart from between-subject variance, we also want to take into account between-item variance. For this reason, we can introduce another additive term in the model analogous b_i, call it c_i, which comes from a distribution $N(0, \sigma_{b2})$. Thus, it is easy to add a varying intercept for items as well, and this requires the model to estimate a further variance component, that due to between-item variability. A second extension involves an adjustment for subject (and item) in the coefficients β_1 and β_2. This is called a varying slopes model. Such adjustments to the coefficients β_1 and β_2 simply take into account possible variability between subjects (and items) in their response to the pronoun and relative clause manipulation; for example, some subjects may speed up much more than others due to the pronoun vs. gap manipulation, and some items may have a greater impact on the pronoun vs. gap manipulation. Models with varying slopes take this variability into account. This is by no means the whole story regarding linear mixed models, but it does provide the reader with some guidance on how we analyzed the data.

One further twist is the issue of non-normality of residuals. Statistical inference based on the models discussed above crucially depends on normal distribution theory; the residuals ϵ are assumed to be normally distributed, and the BLUPs are too. When residuals are not normally distributed (this is usually the case in analyses of reading studies using raw reading times as dependent measures) the above models may no longer be applicable. One solution to this issue is to use generalized linear mixed models (we will not discuss this solution here); another is to find a transformation to the data such that the variance is stabilized and the normality assumptions are approximately satisfied. Box and Cox (1964) developed a procedure for stabilizing variance, which is now known as the Box-Cox procedure and is available through the `MASS` package (Venables and Ripley 2002) in R as the function `boxcox`. Briefly, the function discovers (using maximum likelihood estimation) the transformation needed in order to stabilize variance.

In our reading time data, the transform suggested by the Box-Cox procedure is the reciprocal. We changed this to a negative reciprocal in order to make it easier to interpret the sign of the estimated coefficients (the reciprocal converts the reading time to rate of processing; so a smaller value on the transformed scale is a slower rate, and a larger value corresponds to a faster rate).

Next, we briefly explain the model fitted for response accuracy, which is a proportion. To analyze these, we used the generalized linear modeling (GLM) framework (more specifically, generalized linear mixed models). The basic idea in GLMs is that we assume that responses are generated by a binomial distribution. Instead of assuming that the dependent variable is a proportion μ, we transform it to log-odds and specify a linear relationship between the predictor x and the log-odds: $\log[\frac{\mu}{1-\mu}] = \beta_0 + \beta_1 x$. The generalized linear mixed model extends this framework to deal with grouped data, as discussed above, with varying intercepts, etc. The essential point here is that again we are testing null hypotheses of the form $H_0 : \beta_1 = 0$, where the coefficient is in the log-odds scale.