In press, Language and Cognitive Processes September 3, 2009

Short-term forgetting in sentence comprehension: Crosslinguistic evidence from verb-final structures

Shravan Vasishth
University of Potsdam

Katja Suckow
University of Dundee

Richard L. Lewis
University of Michigan

Sabine Kern
University of Potsdam

Submitted for publication September 3, 2009

Abstract

Seven experiments using self-paced reading and eyetracking suggest that omitting the middle verb in a double center embedding leads to easier processing in English but leads to greater difficulty in German. One commonly accepted explanation for the English pattern—based on data from offline acceptability ratings and due to Gibson & Thomas (1999)—is that working-memory overload leads the comprehender to forget the prediction of the upcoming verb phrase (VP), which reduces working memory load. We show that this VP-forgetting hypothesis does an excellent job of explaining the English data, but cannot account for the German results. We argue that the English and German results can be explained by the parser's adaptation to the grammatical properties of the languages: in contrast to English, German subordinate clauses always have the verb in clause-final position, and this property of German may lead the German parser to maintain predictions of upcoming verb phrases more robustly compared to English. The evidence thus argues against language-independent forgetting effects in online sentence processing: working memory constraints can be conditioned by countervailing influences deriving from grammatical properties of the language under study.

The act of comprehending a sentence necessarily involves a working memory system. Linguistic units, such as lexical items, must be incrementally accessed from a long-term declarative memory, encoded as a memory trace, and maintained in memory until they are integrated with other linguistic units, such as phrase structure. The end-product of these processes is a sufficiently complete representation of the syntax, semantics and discourse structure of a sentence.[1]

A perfect memory system would maintain in pristine form every linguistic unit still active in the parse, with no deterioration in its memorial representation as the sentence unfolds. In reality, however, memory representations degrade over time, resulting in comprehension difficulty and outright comprehension failures.

An interesting example of degradation of memory representations can be illustrated by the contrast in (1), discussed first by Frazier (1985) (the original observation is attributed by Frazier to Janet Fodor). Sentence (1a) is a grammatical sentence: the rules of English allow such embedded structures. Such complex but grammatical structures are perceived by native English speakers to be less acceptable than their ungrammatical counterparts (1b), which has the middle verb phrase *was cleaning every week* missing.

(1) a. The apartment that the maid who the service had sent over was cleaning every week was well decorated.

b. *The apartment that the maid who the service had sent over was well decorated.

The first published study involving this contrast was an offline questionnaire-based experiment by Gibson and Thomas (1999). They found (inter alia) that ungrammatical sentences such as (1b) were rated no worse than grammatical ones such as (1a). In related work, Christiansen and Macdonald (2009) (this work was mentioned in Christiansen & Chater, 2001 and in Gibson & Thomas, 1999), found that ungrammatical sentences were rated significantly better than the grammatical ones. Christiansen and MacDonald used an online self-paced reading task where participants were asked to decide after each word whether the sentence seen so far was grammatical ("stops-making sense" task). After the presentation of the whole sentence, participants rated the overall grammaticality of the sentence (Gibson & Thomas, 1999, 239). Thus, Christiansen and MacDonald's grammaticality judgements were gathered during an online, incremental sentence processing task which required making grammaticality decisions from one word to the next. It is worth noting that Christiansen and MacDonald controlled for sentence length in the two sentence types; by doing so, they also demonstrated that the explanation for the illusion cannot lie in the length difference between grammatical and ungrammatical sentences.

The greater acceptability of ungrammatical sentences compared to the grammatical ones can be explained in terms of the degradation of memory representations as syntactic structure is built incrementally. Gibson and Thomas (1999) argue that the prediction for the middle verb is forgotten if memory cost exceeds a certain threshold. Specifically, when the third NP, *service*, is read (example 1a), five heads are predicted, each with a cost, quantified in Memory Units (MUs). These predicted heads and their associated costs are:

[1] We do not mean that the sentence representation is held in complete form in short-term memory; see Lewis and Vasishth (2005) for a detailed proposal for a parser that meets our criteria but has a severely limited short-term memory.

We are grateful to Heiner Drenhaus, Lyn Frazier, Reinhold Kliegl, Umesh Patil, Titus von der Malsburg, Ruben Van de Vijver and Julie Van Dyke for comments and useful discussions. The reviewers (among them Janet Fodor) also played an important role in improving this work; our thanks to them.

1. the matrix verb (which by assumption involves no cost),

2. a verb to head the first relative clause (RC); this has cost 2 MUs because two discourse referents (*maid* and *service*) have been processed since the prediction for that verb was first made,

3. an NP empty-category to be coindexed with the first RC pronoun; this also involves a cost of 2 MUs because of the two discourse referents,

4. a verb to head the second RC; this involves a cost of 1 MU corresponding to one new discourse referent (*service*) that was processed since the prediction was made,

5. an NP empty-category to be coindexed with the second RC pronoun; this involves a cost of 1 MU.

This high overall memory cost (6 MUs) has the effect that the prediction with the highest cost is forgotten; since the prediction associated the second verb phrase carries the highest cost of 4 MUs (see 2 and 3 above), this prediction is forgotten.

Thus, according to Gibson and Thomas (1999), the prediction for the second VP is forgotten at the third NP because that prediction is associated with the greatest memory cost. Crucially, the cost of the prediction is computed by adding up the number of discourse referents that have been processed *since the prediction was first made*. This calculation is intended to model increasing difficulty in storing and maintaining the prediction in memory.

It is also possible to explain the grammaticality illusion in terms of the Dependency Locality Theory or DLT (Gibson, 2000), although in this case the explanation is quite different. Under the DLT view, memory cost can be quantified as the sum of two components: (a) integration cost: the number of new discourse referents between the verb and the dependent to be integrated with it (crucially, this cost is computed when the verb is processed); and (b) storage cost: the number of predicted (as-yet unseen) heads.[2]

Unlike the Gibson and Thomas account, in the DLT the cost of a stored prediction per se does not increase as more discourse referents are processed: any one predicted head adds an invariant storage cost of 1. Thus, under DLT's assumptions, at the third NP in example (1)a the storage cost should be 2 (1 memory unit for each verb predicted; the prediction of the matrix verb involves no cost by assumption); there is no motivation for forgetting the prediction for the second VP as opposed to the prediction of the first VP.

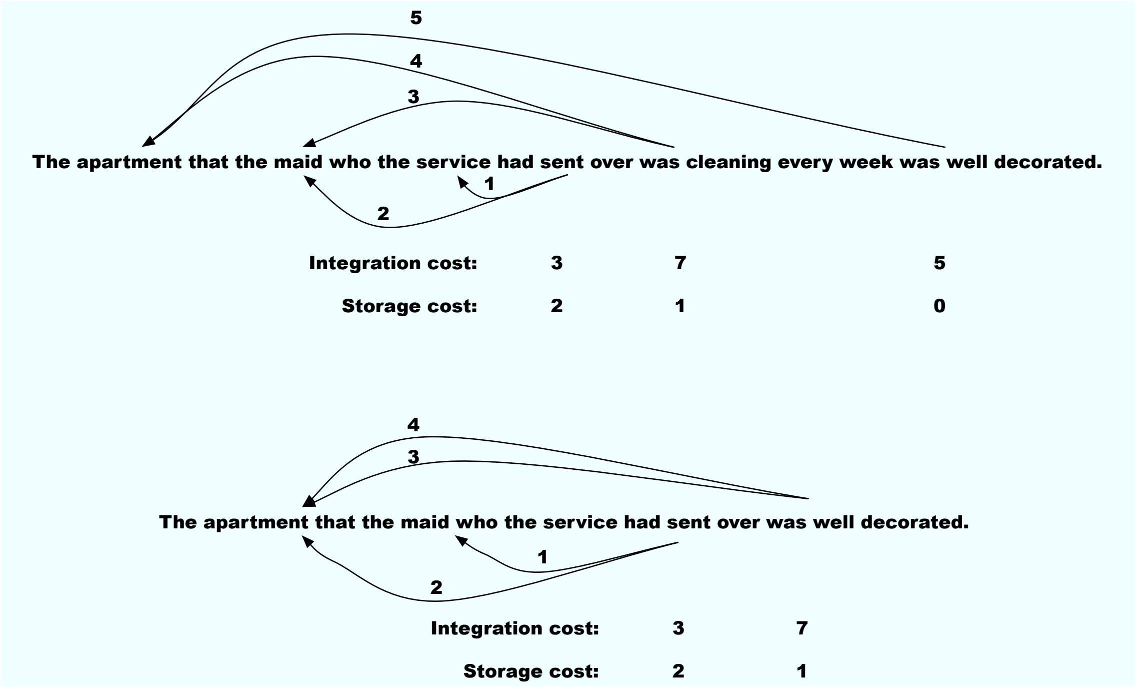
In other words, in the DLT the explanation for the grammaticality illusion cannot be ascribed to forgetting the second VP's prediction; rather, the explanation lies in the greater overall integration cost at the verbs, as we discuss below with reference to Figure 1. The point of greatest difficulty is at the second (embedded) verb *admitted*: the subject *nurse* is separated by three new discourse referents, *clinic*, *hired*, and *admitted*;[3] and the object *patient* by four discourse referents.

Although the second verb has the highest integration cost among the three verbs, this fact alone cannot explain the grammaticality illusion: even in the easier-to-process ungrammatical sentence, which has the missing second verb (1)b, the reader would be forced to assume that the final verb is the second verb and experience the same integration cost as in the grammatical sentence; see Figure 1. In order for DLT to explain the grammaticality il-

[2] Gibson's metric relies on the notions Energy Units and Memory Units; however, the DLT's predictions can be derived by simply counting the number of intervening discourse referents (integration cost) and the number of predicted heads at each point (storage cost).

[3] Finite verbs are assumed to introduce new discourse referents (Gibson, 2000).

lusion, an additional assumption must be made: the *total* integration cost at the verbs (i.e., the sum of the integration costs incurred at the verbs) must be assumed to determine overall processing cost. The observed preference for the grammaticality illusion would follow because the total integration cost is lower in the ungrammatical sentence. An important point is that this explanation does not predict that there is anything special about omitting the second verb; the total integration cost at the verbs would be lower regardless of which verb is omitted. One could nevertheless argue that omitting the second verb yields the largest reduction in processing load compared to the other verbs because it involves the greatest reduction in total integration cost. Since the total integration cost would be registered at the final verb and possibly at the region following the final verb, the DLT would predict increased difficulty in the grammatical condition at either the final verb and/or the region following the final verb.



The apartment that the maid who the service had sent over was cleaning every week was well decorated.

Integration cost: 3 7 5

Storage cost: 2 1 0

The apartment that the maid who the service had sent over was well decorated.

Integration cost: 3 7

Storage cost: 2 1

Figure 1. A schematic illustration of DLT's predictions for the grammatical and ungrammatical structures. Integration costs are labeled along the arcs that define the argument-head dependencies, and storage costs are presented below each verb phrase. The storage costs are the number of heads predicted at each point; only the storage costs at the verb are shown because these are the only critical ones for this paper.

In summary, the explanation for the examples in (1) is closely linked to the idea of memory overload induced by distance, but the explanations offered by Gibson and Thomas (1999) and the more recent DLT (Gibson, 2000) differ. Under the Gibson and Thomas view, the VP-prediction for the second verb is forgotten because of high memory cost, which arises due to the number of intervening new discourse referents since the head was first predicted. Under the DLT view, the difficulty arises due to the higher total integration cost at the verbs in the grammatical sentence. Although the explanations differ, both predict that the difficulty in online processing should be observed at the final verb or the region following the final verb.

Our first goal in this paper was to establish whether any evidence is available from online methods for the grammaticality illusion (as opposed to offline acceptability ratings as in (Gibson & Thomas, 1999) and (Gimenes, Rigalleau, & Gaonac'h, 2009)). The use of online methods is necessary because although the explanations for the grammaticality illusion refer to online processing difficulty, the published evidence for forgetting comes from acceptability judgements, an offline task. The only exception is the online processing study reported by Christiansen and Chater (2001), but here too the dependent measure was acceptability rating. Given that DLT and its precursor, the Syntactic Prediction Locality Theory (Gibson, 1998), are not primarily theories of acceptability judgements but of online, incremental processing difficulty, it is reasonable to examine the claims using online methods.[4]

A second goal was to investigate whether the forgetting hypothesis (driven by VP-forgetting) applies to languages other than English. In this context, it is interesting that French speakers also rate the ungrammatical versions better than their grammatical counterparts (Gimenes et al., 2009). Given the consistency between English and French acceptability ratings it seems reasonable to assume that VP-forgetting is not a language-specific matter but rather the consequence of a general working memory system. It is therefore critical to establish whether this is true. If language-specific factors can modulate, arrest, or even reverse the forgetting effect, this is a strong argument against an unconditionally language-independent explanation.

In order to address the cross-linguistic and online issues, we used English and German and two online methods to determine whether we could uncover evidence for the forgetting effect.

Regarding the choice of languages, English and German were chosen because the two have different word orders: the former has the default surface order subject-verb-object (SVO) and the latter SOV (at least in subordinate clauses). This difference in word order is interesting because previous work by Konieczny (2000) has shown that the predictions of Dependency Locality Theory (which are also based on integration costs) are not upheld in the case of German. German is therefore an important test case for locality-based explanations of the grammaticality illusion. Regarding methods, we chose two well-known techniques, self-paced reading (SPR) and eyetracking. SPR was chosen because it is has been used quite often in work relating to locality effects; and eyetracking was chosen because the eye movement record could provide additional evidence (such as re-reading times) relevant for evaluating the theoretical claims.

To sum up the discussion so far, we reasoned as follows: (a) If the grammaticality illusion is found in online processing as well, this would provide stronger support for the explanations provided by Gibson and Thomas and the DLT. (b) If the grammaticality illusion occurs irrespective of the headedness of the language, this provides further evidence for the cross-linguistic generality of existing accounts; but if headedness determines whether the

[4] This does not mean that acceptability ratings are inappropriate for investigating the claims of such theories; we mean merely that offline ratings, by definition, cannot provide us with a detailed picture of events unfolding in real time.

illusion occurs at all, the explanatory power of cross-linguistically applicable explanations would be considerably weakened.

Experiment 1: English self-paced reading

Method

Participants. Forty nine English native-speakers from the University of Michigan participated in the experiment. They received course credit as compensation. All had normal or corrected-to-normal vision.

Procedure and Design.

The present experiment and all other experiments presented here had two conditions: grammatical and ungrammatical sentences.[5] The grammatical sentences had three verbs, while the ungrammatical sentences had the second verb missing.

(2) a. Grammatical
The carpenter who the craftsman that the peasant carried hurt supervised the apprentice.

b. Ungrammatical
The carpenter who the craftsman that the peasant carried - supervised the apprentice.

A self-paced reading comprehension experiment (Just, Carpenter, & Woolley, 1982) was carried out in English at the University of Michigan, USA. Sixteen target sentences were presented in a counterbalanced manner, with 56 filler sentences pseudo-randomly interspersed between the target sentences. The counterbalancing meant that each participant saw each sentence only once. The experiment was run on a Macintosh computer using the Linger software developed by Doug Rohde (http://tedlab.mit.edu/∼dr/Linger/).

Participants read the introduction to the experiment on the computer screen. In order to read each word of a sentence successively in a moving window display, they had to press the space bar. The word seen previously was masked. Each word was shown separately. At the end of each sentence subjects had to answer yes-no comprehension questions in order to ensure that they would try to comprehend the sentences; no feedback was given as to whether the response was correct or not. One concern was that participants might develop a question-answering strategy without paying attention to the entire sentence; therefore, questions were designed to probe different argument-verb relations in the sentences. More specifically, in all the experiments the questions relating to the target sentences targeted all verbs in the grammatical condition, and the two verbs in the ungrammatical one; the questions were designed to have "yes" answers in 50% cases. All stimulus items and accompanying questions, along with the expected correct answers, are shown in the Appendix.

[5] In experiments 1-4 presented here, another factor was included, but this was orthogonal to the present issue. This other factor was interference: we manipulated the similarity of the second NP with respect to the first and third NPs. Since the results of that manipulation do not concern this study, we omit discussion of this factor in the paper. The items presented in the appendix show all four conditions, and the experimental data, which will be made available online, will contain a full discussion of the interference results and their interpretation.

An important point to note is that the questions were not designed to probe the forgetting account but rather to ensure that participants were reading for comprehension. As a consequence, although we report comprehension accuracies for completeness, we do not interpret them theoretically.

Statistical analysis.

A linear mixed-effects model (LME) was fitted to the data, with crossed random intercepts for participants and items, and with grammaticality as the predictor (fixed factor). LMEs have several advantages over repeated measures ANOVA, one of them being that they allow us to take by-item and by-participant variance into account simultaneously, which is an improvement over separate analyses or the calculation of min-F (Clark, 1973), (Raaijmakers, Schrijnemakers, & Gremmen, 1999); see (Baayen, 2008) for further discussion of this issue. Throughout this paper we present coefficient estimates, their standard errors, t- or z-scores (depending on the dependent measure), and highest posterior density (HPD) intervals derived from 50,000 Monte Carlo Markov Chain runs. An absolute t-score of 2 or greater indicates significance at the α level 0.05. Note also that the t-score is not accompanied by degrees of freedom or p-values. This is because degrees of freedom can only be approximated in LMEs (Baayen, 2008).

In all experiments presented here, the statistical analyses on reading times were carried out on log-transformed values. We report the log-transformed analyses because linear models based on untransformed reading times do not meet the assumptions of additivity and linearity (Gelman & Hill, 2007, 59-65). Another reason is that untransformed reading times often lead to non-normal residuals in the fitted linear models, which violates a central assumption of linear regression. We also removed reading times greater than 2000 ms in all the experiments on the grounds that such long reading times are unlikely to reflect online processing events; however, the results of the experiments presented in this paper remain unchanged regardless of whether these reading times are removed or retained. We also repeated all the analyses using untransformed values, and except in Experiment 4 (see footnote 8), the results were similar regardless of whether we transformed the data or not.

Predictions

The predictions of the original VP-forgetting hypothesis of Gibson and Thomas and those of Gibson's more recent theory, Dependency Locality Theory, are discussed next with reference to examples 2. We discuss the two explanations separately.

For ease of exposition, we will refer the verb sequence in (2a) with indices V3, V2, V1. The verb V3 is the most embedded verb, V2 the middle verb and V1 the matrix verb. This convention simply reflects the nesting of argument-verb dependencies (as opposed to Dutch (Bach, Brown, & Marslen-Wilson, 1986), where they are crossed; also see Joshi, Becker, & Rambow, 2000; Rambow & Joshi, 1994). We will refer to the verbs in the conditions as Vn where n is the verb index. In addition, we will use the phrase "second verb" to refer to whichever verb sequentially follows the embedded verb V3; in the grammatical condition it is V2, and in the ungrammatical condition V1.

First consider the predictions of the Gibson and Thomas account. The parsing events of interest begin immediately after V3 is processed. When V3 is read, the prediction for the embedded verb, a verb phrase (VP3), is retrieved, the verb is integrated with the predicted

VP, and processing continues to the next word. No reading time difference is predicted at the embedded verb for the grammatical and ungrammatical conditions. We separate the theoretically interesting events into two stages:

- Process second verb: The next word is V2 (in the grammatical condition) or V1 (in the ungrammatical condition). Since the prediction for the middle verb phrase node is by assumption forgotten, the only predicted node available for retrieval in both the ungrammatical and grammatical cases is the matrix VP node. That is, in the grammatical condition, the *middle* verb V2 triggers a retrieval of the matrix-VP prediction and attaches to it; by contrast, in the ungrammatical condition the matrix verb V1 retrieves and attaches to the predicted matrix VP node. Here too, if the VP-forgetting hypothesis is correct, no difference in reading time should occur at grammatical (V2) versus ungrammatical (V1) conditions, other than due to any differences between the semantic fit of previously processed arguments with V2 versus V1.

- Process region following second verb:

- Process V1 (grammatical condition): In the grammatical condition, the matrix verb V1 is processed but no predicted VP-node is available for attachment—the middle verb's predicted VP node has been forgotten and all the other predicted VP nodes have been consumed. Thus, an error should be registered by the parser. It is possible that the effects of this error appear in the post-matrix verb region, as a spillover effect (Mitchell, 1984).

- Process post-matrix verb region (ungrammatical condition): In the ungrammatical condition, after V1 is processed, the word following the matrix verb should be processed successfully since all the (unforgotten) predicted VP nodes have been filled. Thus, the reading time at the post-matrix-verb region in the ungrammatical condition should be shorter compared to the post-matrix verb region in the grammatical condition.

These processing steps together predict longer reading times at the matrix verb V1 in the grammatical condition compared to the matrix verb V1 in the ungrammatical condition. In self-paced reading, the behavioral correlate of this error—longer reading time—should also occur in the post-matrix verb region. Thus the prediction for self-paced reading is that reading time in the post-matrix verb region of the grammatical condition could be longer than in the post-matrix verb region of the ungrammatical condition. We discuss the predictions for eyetracking when the relevant experiment is discussed.

The predictions of DLT are identical to those of the Gibson and Thomas account, although the explanation is based on integration cost. As discussed earlier with reference to Figure 1, if we assume that the total or cumulative integration cost affects processing difficulty, then at V1 the ungrammatical condition should be easier to process. It is possible that the effect of cumulative integration cost persists beyond the final verb and shorter reading times are seen at the region following the final verb.

Results

Question-response accuracies.

As mentioned earlier, in the experiments presented in this paper, question-response accuracies cannot provide definitive evidence relating to the theoretical issues, which relate to online processing events. In addition, the questions were not designed to probe the grammaticality illusion but were rather designed to ensure that participants read for comprehension. However, for completeness we present statistical analyses of response accuracy as a function of grammaticality. In addition to presenting the mean accuracies by factor, the overall accuracy (including all distractors) would also tell us whether readers were attending to the sentences. Therefore, we present mean overall accuracy as well.

In the present experiment, the mean question-response accuracy over all items (including distractors) was 79.6%, suggesting that participants were attending to the meaning of the sentences. A contrast was defined for fitting a generalized linear mixed effects model (Gelman & Hill, 2007): the grammatical conditions was coded as 1 and the ungrammatical condition as -1.[6]

The grammatical target sentences were more difficult to answer correctly (57.75% correct responses) compared to the ungrammatical ones (65.75%), z=-2.34, p<0.05; coefficient -0.17189, SE 0.0734.

Reading time.

The mean reading times and 95% confidence intervals are summarized in Figure 2 and the statistical analyses in Table 1. The VP-forgetting hypothesis makes predictions about the V1 and post-V1 region We therefore present analyses for these regions in Table 1. Here and in all subsequent analyses, the grammaticality factor was coded as 1 for grammatical sentences, and -1 for ungrammatical sentences. This coding aids in the interpretation of the estimated coefficients: a positive sign indicates slower reading time for the grammatical condition, whereas a negative sign indicates faster reading time.

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	0.007273	0.017104	0.43	-0.02693120	0.04134948
Post-V1	0.141575	0.014559	9.72 *	0.11161532	0.16940903

Table 1: Coefficients, standard errors and t-scores (the asterisk indicates statistical significance at α=0.05) for the grammaticality effect in the English self-paced reading experiment 1. The grammatical condition was coded as 1 and the ungrammatical condition as -1.

As predicted by Gibson and Thomas' VP-forgetting hypothesis and the DLT, in the Post-V1 region the ungrammatical condition was significantly easier to process than the grammatical; in the V1 region the difference between the conditions had the predicted pattern but did not reach significance.

Thus, Experiment 1 yielded clear evidence in favor of Gibson and Thomas' VP-forgetting hypothesis and the DLT. Regarding the evidence for the VP-forgetting account, we were not overly concerned that a statistically significant effect was not found at V1 because in self-paced reading, processing difficulty often manifest itself in a subsequent region (spillover, as mentioned earlier), and because there was a numerical tendency at V1 that was consistent with the VP-forgetting hypothesis. Moreover, subsequent experiments show the predicted effects at V1, as discussed later.

[6] Sentence 11 was not considered in any of the analyses because it contained a typographic error. The sentence in question was *The clerk who the bureaucrat that the visitor forgotten about...*, where it should have been *...forgot about...*.

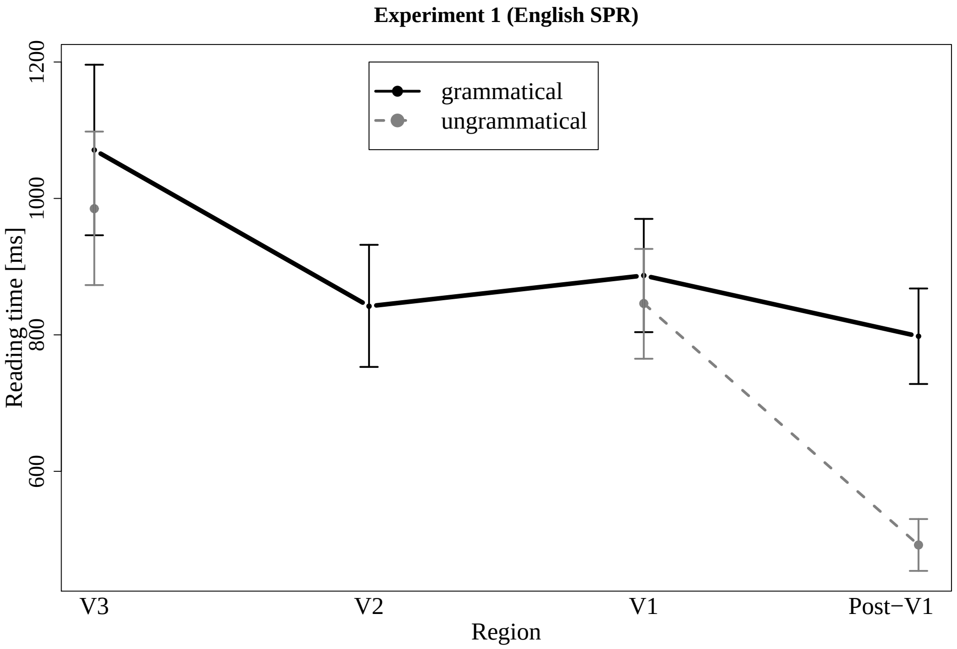


Figure 2. Mean reading times and 95% confidence intervals for the three verb regions and the post-V1 word in the English self-paced reading study (experiment 1). The figure shows reading times with respect to the grammaticality manipulation.

Although the results of Experiment 1 were consistent with the VP-forgetting hypothesis, we were concerned that the presence of three-verb sequences appearing in succession may have rendered processing unusually difficult (Frazier, 1985). We therefore repeated the study (we refer to this replication as Experiment 1a below) with sentences that had more material between the verbs; examples are shown in (3) and the full list of stimuli is available from the authors (these are not included in the paper in order to conserve space). The three verb regions are shown in square brackets. This experiment had 62 fillers interspersed between the target items (the same method was used as described above).

(3) a. The carpenter who the craftsman that the peasant [$_{V_3}$ had carried] to the bus-stop [$_{V_2}$ had hurt] yesterday [$_{V_1}$ supervised] the apprentice.

b. The carpenter who the craftsman that the peasant [$_{V_3}$ had carried] to the bus-stop [$_{V_1}$ supervised] the apprentice.

Here, the two non-matrix verb regions were the auxiliary plus main verb (in the examples above, V3: *had carried*, and V2: *had hurt*). The other regions of interest were as in Experiment 1 (NP3: *the peasant*, V1: *supervised*, and post-V1: *the*). Thus, V3 was separated from V2 (V1 in the ungrammatical condition) by a PP *to the bus-stop*, and V2 from V1 by an adverb *yesterday*.

Results: Experiment 1a

Question-response accuracies.

The mean response accuracy over all items (including distractors) was 83.7%. The grammatical target sentences were more difficult to answer correctly (54% correct) compared to the ungrammatical ones (63.30%), z=-2.604, p<0.05, coefficient -0.195, SE 0.078. As in experiment 1, this analysis also relied on a generalized linear mixed-effects model with a logit link function (participants and items were treated as crossed random factors).

Reading time.

The results are summarized in Figure 3 and Table 2. Consistent with the VP-forgetting hypothesis, the reading times at the V1 and post-V1 region were significantly shorter in the ungrammatical condition than the grammatical one. This is as predicted by the VP-forgetting hypothesis and the DLT. Table 2 presents the results of a linear mixed-effects model fitted with participants and items as crossed random factors; reading times were log-transformed, as discussed earlier.

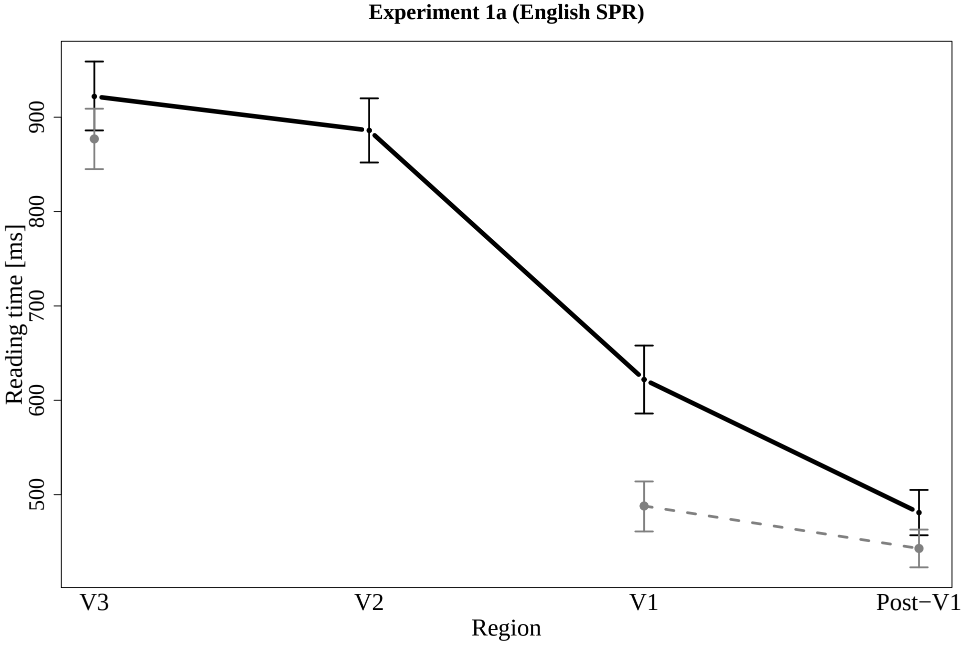


Figure 3. Mean reading times and 95% confidence intervals for the three verb regions and the post-V1 word in the second English self-paced reading study (experiment 1a). The figure shows reading times with respect to the grammaticality manipulation.

In conclusion, the findings in Experiment 1a are also consistent with the VP-forgetting hypothesis and the DLT: reading time at the V1 and Post-V1 regions is longer in the grammatical conditions compared to the ungrammatical conditions. The next question we address is whether eyetracking during reading also provides corroborating evidence for these theories. This experiment is described next.

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	0.106307	0.013856	7.67 *	0.07834222	0.13396855
Post-V1	0.033516	0.011542	2.9 *	0.01093247	0.055340952

Table 2: Coefficients, standard errors and t-scores (the asterisk indicates statistical significance) for the grammaticality effect in the English self-paced reading experiment 1a. Also shown are 95% Highest Posterior Density intervals (based on 50,000 MCMC runs).

Experiment 2: English Eyetracking study

Method

Participants.

Forty seven native English speakers (undergraduates at the University of Michigan) took part in this study, each receiving either course credit or 10 US dollars for completing the task. Participants were tested in individual sessions, and took approximately 30 minutes to complete the experiment. All had normal or corrected-to-normal vision.

Procedure and Design.

Participants were seated approximately 50 cm from a 19-inch color monitor with 1024 × 768 pixel resolution; twenty-three pixels equaled about one degree of visual angle. Participants wore a SensoMotoric Instruments Eyelink I head-mounted eye-tracker running at 250 Hz sampling rate. Although viewing was binocular, only data from the right eye was used in analyses. Participants were instructed to avoid strong head movements or large shifts in position throughout the experiment. A standard PC keyboard was used to record responses. The presentation of the materials and the recording of the responses was controlled by a PC running software developed (using the Eyelink API) in the University of Michigan eyetracking laboratory

Each participant was randomly assigned one of four stimuli lists which comprised different item-condition combinations according to a Latin Square. The trials per session were randomized individually per participant, subject to the constraints that experimental trials were always separated by at least one filler, and that each session started with at least five fillers. At the start of the experiment, the experimenter performed the standard Eyelink calibration procedure, which involves participants looking at a grid of nine fixation targets in random succession. Then a validation phase followed to test the accuracy of the calibration against the same targets. If there was a discrepancy between calibration and validation of more than 1.2 degrees of visual angle for any target, or if the average discrepancy across all nine targets was greater than 0.7 degrees of visual angle, then calibration and validation were repeated until the discrepancy was acceptable. Calibration and validation were repeated during the session if the experimenter noticed that measurement accuracy was poor (e.g., after strong head movements or a change in the participant's posture).

Each trial consisted of the following steps. First, a quick single target recalibration ("drift correction") was performed to correct for any degradation of measurement accuracy due to subject movement or slippage of the eyetracker on the head. Then, a fixation target appeared 12 pixels to the left of where the left edge of the text would appear. The stimulus was presented only after the subject fixated on this target for 900 consecutive milliseconds. The participant was instructed to press the spacebar on the keyboard when he/she had finished reading. This triggered the presentation of a simple comprehension question which the participant had to answer either with the 'F' key ('yes') or the 'J' key ('no'). The text stimuli were presented using a Courier New font, printed in white on a black background. The characters (including spaces) were all the same width, approximately 9.1 pixels or 0.39 degrees of visual angle.

The presentation software automatically recorded the coordinates of rectangular interest areas around each word; fixations were then associated with words according to whether their coordinates fell within a word's interest area. The left and right boundaries of each interest area were the midpoints of the spaces between the words (or in the case of the leftmost and rightmost words, the midpoint of where a space would have been had there been another word to the left or right). The upper and lower boundaries were 24 pixels above and below the top and bottom of the line of text. The line of text (i.e., the space taken by a capital letter) was 10 pixels high.

Results and Discussion

Question response accuracies.

The mean response accuracy over all items (including distractors) was 79.3%, indicating that participants were attending to the comprehension task. Accuracy in the grammatical condition was 54.93%, which was significantly lower than in the ungrammatical condition, 70.05% (coefficient -0.34578, SE 0.07793, z=-4.437, p<0.01).

Reading time.

In order to map the predictions of the forgetting hypothesis to eyetracking dependent measures, it is necessary to first briefly review our assumptions regarding the mapping between eyetracking dependent measures and human parsing processes. The most common dependent measures and their interpretation in terms of reading processes are as follows. *First fixation duration* (FFD) is the time elapsed during the first fixation during first pass (the first encounter with a region of interest as the eye traverses the screen from left to right), and has been argued to reflect lexical access costs (Inhoff, 1984). *Gaze duration* or *first pass reading time* (FPRT) is the summed duration of all the contiguous fixations in a region before it is exited to a preceding or subsequent word; Inhoff (1984) has suggested that FPRT reflects text integration processes, although Rayner and Pollatsek (1987) argue that both FFD and FPRT may reflect similar processes and could depend on the speed of the cognitive process. ***First-pass regression probability*, the probability of the eye making a leftward saccade to a previous during first pass, is another measure that is sometimes used as an index of processing difficulty.** *Right-bounded reading time* (RBRT) is the summed duration of all the fixations that fall within a region of interest before it is exited to a word downstream; it includes fixations occurring after regressive eye movements from the region, but does not include any regressive fixations on regions outside the region of interest. RBRT may reflect a mix of late and early processes, since it subsumes first-fixation durations. *Re-reading time* (RRT) is the sum of all fixations at a word that occurred after first pass; RRT has been assumed to possibly reflect the costs of late-stage processes (Clifton, Staub, & Rayner, in press, 349), and recent work suggests that it may be informative about retrieval costs (Gordon, Hendrick, Johnson, & Lee, 2006, 1308), (Vasishth, Bruessow, Lewis, & Drenhaus, 2008). Another measure that has been invoked in connection with late-stage processes is *regression path duration*, which is the sum of all fixations from the first fixation on the region of interest up to, but excluding, the first fixation downstream from the region of interest. Finally, *total reading time* (TRT) is the sum of all fixations on a word.

The early dependent measures (e.g., FFD, FPRT) did not show any effect in any of the eyetracking experiments reported here; the effect showed up in the late measures, re-reading time and total reading time (the latter subsumes re-reading time and first-pass reading time). We present results based on re-reading times (rather than total reading time) in this paper because of the recent findings that retrieval difficulty may be reflected in this dependent measure. As mentioned above, Gordon et al. (2006) and Vasishth et al. (2008) have found evidence from re-reading time for predicted retrieval/integration difficulty and retrieval failures. Since the forgetting hypothesis and DLT both predict integration difficulty (in the case of the Gibson and Thomas account, retrieval failure of the predicted verb phrase; and in the case of DLT, increased cumulative integration cost), it is reasonable to focus on these measures.

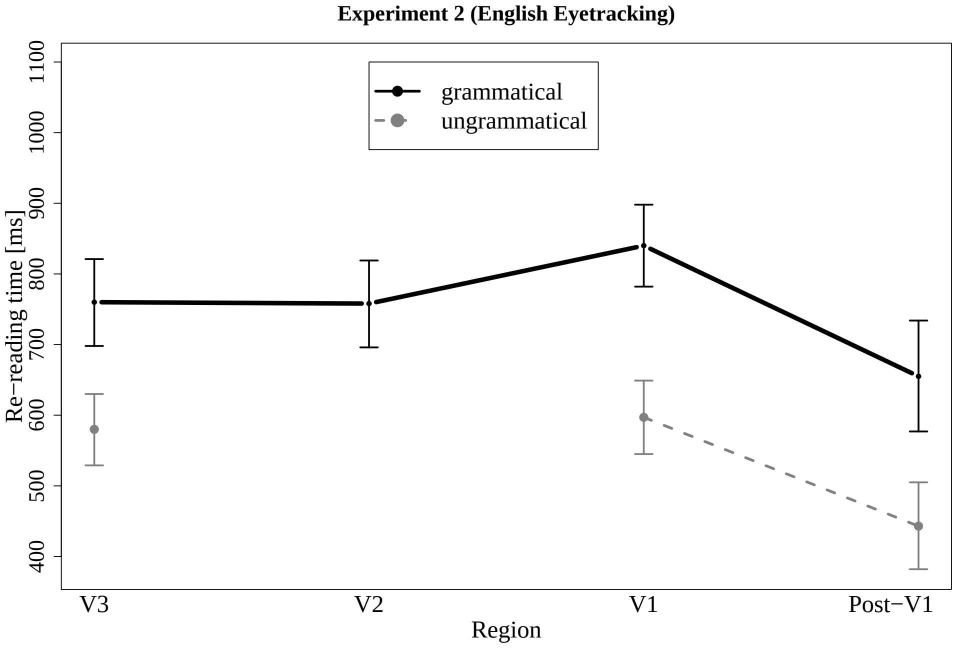


Figure 4. Mean re-reading times and 95% confidence intervals for the verb and post-verbal regions in the English eyetracking study (experiment 2). The figure shows the effect of the grammaticality manipulation.

The mean re-reading times and 95% confidence intervals are summarized in Figure 4 and Table 3. As predicted by the Gibson and Thomas account and DLT, re-reading times at

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	0.19247	0.02933	6.56 *	0.132494098	0.24943728
Post-V1	0.15313	0.03859	3.97 *	0.07935625	0.23827624

Table 3: Summary of results of the comparisons for experiment 2. The asterisk (*) indicates statistical significance at $\alpha = 0.05$.

V1 and the Post-V1 regions were significantly shorter in the ungrammatical condition compared to the grammatical one. In addition, although V3 was not a theoretically interesting region for the grammaticality manipulation, re-reading time in this region was also shorter in the ungrammatical condition compared to the grammatical one (coefficient -0.32631, SE 0.06021, t=-5.42).

To sum up the English experiments, in both the SPR and eyetracking studies we found evidence consistent with the predictions. One surprising result in the eyetracking data was the shorter re-reading time at V3 in the ungrammatical condition. However, this effect can be explained. Re-reading time is a function of revisits to regions that have already been viewed during first pass. Given that regressions are generally more frequent in complex sentences (where complexity is defined as increased ambiguity (Clifton et al., in press) or any other kind of integration difficulty), and given that the ungrammatical sentences are predicted to be less complex overall, it may not be surprising after all that re-reading time is shorter at V3 in the ungrammatical condition.

We describe next the German experiments.

Experiment 3: German self-paced reading study

Method

Participants.

Thirty nine native German speakers (undergraduates at the University of Potsdam) took part in this study, each receiving 7 Euros for participating. Participants were tested in individual sessions, and took approximately 30 minutes to complete the experiment. All had normal or corrected-to-normal vision.

Procedure and Design.

The procedure of the German self-paced reading experiment was identical to the English experiment 1. This experiment had 16 critical items and 60 distractors. Example items are shown below; see the Appendix for the full list of experiment items (as in the previous experiments, there were two other conditions in the experiment but we do not discuss these here).

(4) a. Grammatical
Der Anwalt, den der Zeuge, den der Spion betrachtete, schnitt, überzeugte den Richter.

b. Ungrammatical
Der Anwalt, den der Zeuge, den der Spion betrachtete, überzeugte den Richter.

As mentioned earlier, German and English differ in several respects, two of which are particularly important in the present context. First, German has head-final (SOV) subordinate clauses whereas English has SVO order. Second, in German relative clauses, commas obligatorily demarcate the beginning and end of a relative clause; thus, the presence of a comma with a verb inside a relative clause is a clear cue to the reader that the relative clause has come to an end. Note that the matrix verb does not require a comma.

The predictions of DLT are analogous to the English case and need no discussion. For the VP-forgetting hypothesis of Gibson and Thomas, it is critical that commas are obligatory for embedded verbs but do not occur with matrix verbs. If—as the VP-forgetting hypothesis predicts—participants forget the middle verb's prediction, the absence of the comma on V1 should not alert the reader that an embedded verb was expected: in the grammatical condition, they should experience the same difficulty at V1 that we found in the English experiments. Thus, the VP-forgetting hypothesis predicts that German should show the same reading time patterns that English does. The alternative possibility is that the reader does not forget the prediction for the middle verb; this is equivalent to the parsing mechanism retaining the information that it is processing an embedded clause. If the parsing mechanism is expecting an embedded clause context (has not forgotten the middle verb's prediction), an error should be triggered upon encountering the comma-less matrix verb in the grammatical condition. This predicts—contrary to the VP-forgetting hypothesis—that processing should be more difficult at V1 and the Post-V1 word in the *un*grammatical condition, compared to the grammatical one.

We tested these predictions for German using self-paced reading.

Results

Question-response accuracy.

Question-response accuracy (for all items, including distractors) was 79.51%. The mean accuracy for the grammatical condition was 65%, which was significantly lower than in the ungrammatical condition, 71.5%, coefficient -0.19423, SE 0.09953, z-score=-1.952, p=0.05.

Reading time.

The mean reading times and 95% confidence intervals at the verb regions are summarized in Figure 5 and the results of the mixed-effects model analysis are shown in Table 4.

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	-0.067315	0.015935	-4.22	-0.097603787	-0.03427237
Post-V1	-0.070175	0.016191	-4.33	-0.10162431	-0.03635828

Table 4: Coefficients, standard errors and t-scores (the asterisk indicates statistical significance at α=0.05) for the grammaticality effect in the German self-paced reading experiment 3.

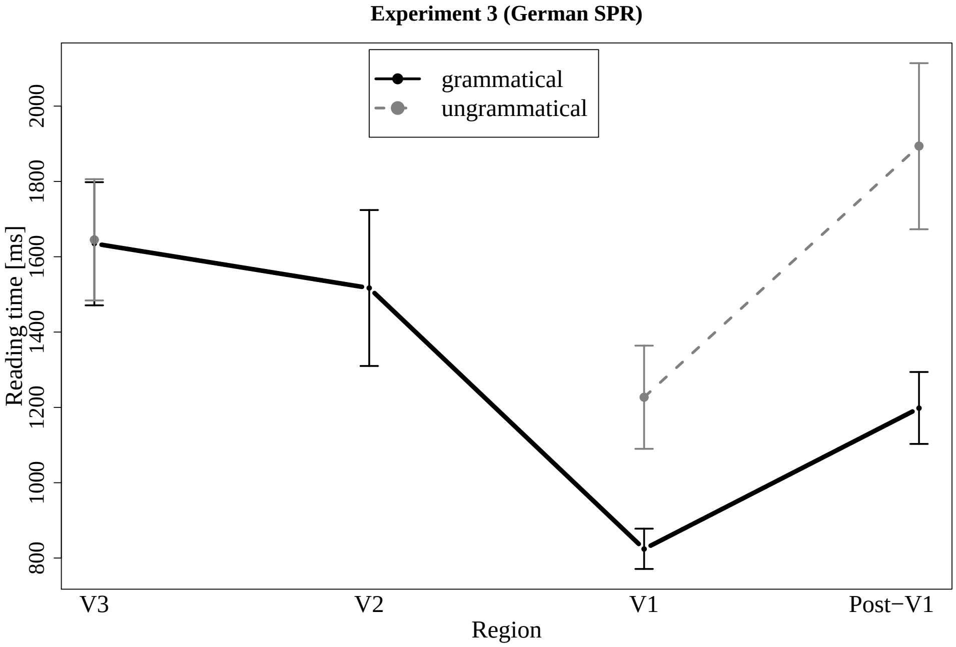


Figure 5. Mean reading times and 95% confidence intervals for the critical regions in the German self-paced reading study (experiment 3). The figure shows reading times with respect to the grammaticality manipulation.

The reading-time analyses show that, contrary to the predictions of DLT and the VP-forgetting account, at the matrix verb V1 as well as the Post-V1 word, the ungrammatical condition was read *slower* than the grammatical one. This is the opposite of the result found for English in experiments 1, 1a and 2.

Experiment 4: German eyetracking study

Method

Participants.

Fifty one native German speakers (undergraduates at the University of Potsdam) took part in this study, each receiving 7 Euros for participating. Participants were tested in individual sessions, and took approximately 30 minutes to complete the experiment. All had normal or corrected-to-normal vision.

Procedure and Design.

The procedure for this experiment is similar to that of the English experiment 2. There were some minor differences in procedure and apparatus, as discussed next. Participants were seated 55 cm from a 17” color monitor with 1024 × 768 pixel resolution. The eyetracker used was an IView-X eye-tracker (SensoMotoric Instruments) running at 240 Hz sampling rate, 0.025 degree tracking resolution, < 0.5 degree gaze position accuracy. Participants were asked to place their head in a frame and to position their chin on a chin-rest for stability. The angle per character was 0.26 degrees (3.84 characters per degree of visual angle).

Participants were asked to avoid large head movements throughout the experiment. A standard three-button mouse was used to record button responses. The presentation of the materials and the recording of responses was controlled by two PCs running proprietary software (the software used was Presentation, and SensoMotoric Instruments' own software for eyetracker control).

At the start of the experiment the experimenter performed a standard calibration procedure, which involves participants looking at a grid of thirteen fixation targets in random succession in order to validate their gazes. Calibration and validation were repeated after every 10-15 trials throughout the experiment, or if the experimenter noticed that measurement accuracy was poor (e.g., after large head movements or a change in the participant's posture).

As in the self-paced reading study, there were 60 distractor sentences and 16 stimulus sentences in each list (see appendix for the full list of stimuli), and each list was pseudorandomly reordered. The trials per session were randomized once for each file, subject to the constraint that each session started with at least one filler.

Results and Discussion

Question-response accuracy.

The mean response accuracy for all items in the experiment (distractors and targets) was 77.25%. The mean accuracy for grammatical conditions was 68.63% and for ungrammatical conditions was 70.59%; this difference did not reach statistical significance.

Reading time.

The mean re-reading times and 95% confidence intervals are summarized in Figure 6 and the statistical analyses in Table 5.

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	-0.07993	0.03755	-2.13	-0.15309065	-0.00092967
Post-V1	-0.07227	0.05347	-1.35	-0.18210067	0.03440298

Table 5: Summary of results of the two regions of interest in the German eyetracking experiment 4. The asterisk (*) indicates statistical significance at $\alpha = 0.05$.

As in the German SPR study, the matrix verb V1 in the ungrammatical condition had longer reading time than V1 in the grammatical condition. The re-reading time at the word following V1 showed no significant difference.[7] This result is inconsistent with the DLT and the VP-forgetting hypothesis, and is consistent with the hypothesis that the middle verb's prediction is not forgotten.

[7] As mentioned earlier, we also analyzed the data with untransformed reading times. At the post-V1 region, a significant difference was seen in the untransformed data: coefficient -38.28, SE 18.52, t=-2.067.

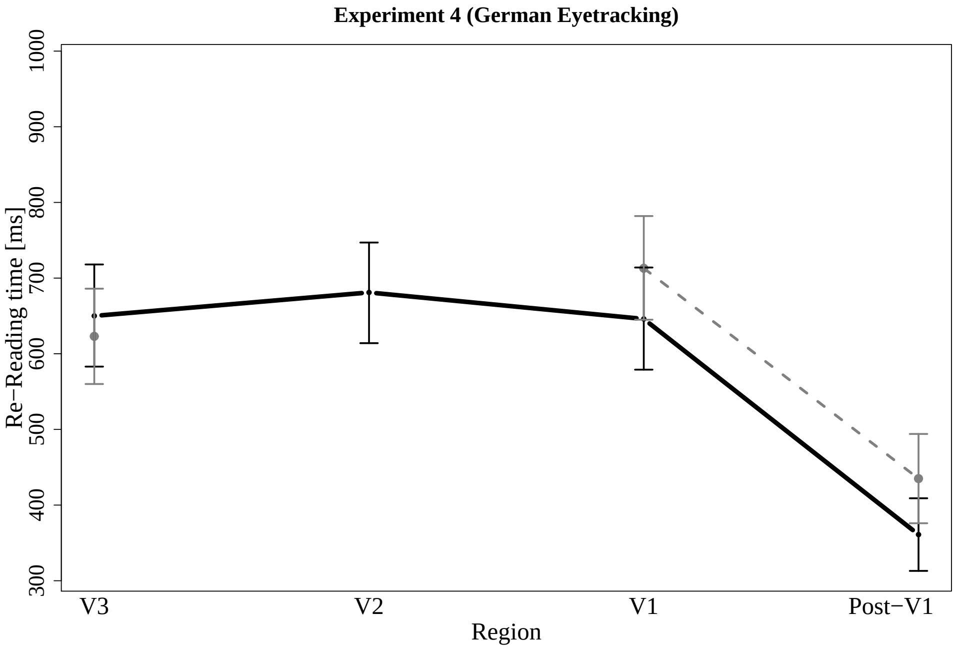


Figure 6. Mean re-reading times and 95% confidence intervals for the critical regions in the German eyetracking study (experiment 4). The figure shows the effect of grammaticality.

One concern with the present experiments is the differences between the English and German experiments could be attributed merely to the obligatory presence of commas in German relative clauses. It is possible that German speakers simply rely on counting commas to establish the dependencies between arguments and heads. If the commas *alone* provide an extra cue for processing, i.e., independent of the fact that verbs occur head-finally in German relative clauses, and if the presence of commas alone could reverse the forgetting effect, then English should also show the same pattern as German if commas are used in the critical English structures.

In order to test this comma-hypothesis, we carried out two further studies involving English in which the comma was present. In the first study (experiment 5), we merely added commas to the stimuli used in the previous English experiments. In experiment 6, we modified the stimuli such that the relative clauses were uniformly non-restrictive relative clauses. The same items (with the modifications mentioned above) were used in both studies as in experiments 1, 1a and 2.

Experiment 5: English self-paced reading study

Method

Participants.

Fifty English native-speakers from the University of Michigan participated in the experiment. They received course credit for taking part. All had normal or corrected-to-normal vision.

Procedure and Design.

The procedure and design were as described in earlier SPR studies, with the difference that commas were added to the relative clauses; see 5.

(5) a. The carpenter, who the craftsman, that the peasant carried, hurt, supervised the apprentice.

b. The carpenter, who the craftsman, that the peasant carried, supervised the apprentice.

Results and Discussion

The overall mean response-accuracy (including all items) was 79%. The accuracy for the grammatical sentences was 54% and for the ungrammatical sentences the accuracy was 63.5%. The ungrammatical sentences had a significantly higher accuracy than the grammatical ones (coefficient -0.20966, SE 0.07314, z=-2.86, p<0.01).

Regarding the reading times, as summarized in Figure 7 and Table 6, the English reading pattern remained unaffected by the presence of commas. As in the previous experiments involving English, the ungrammatical sentences were read faster at the final verb as well as in the region following the final verb (the word *the*).

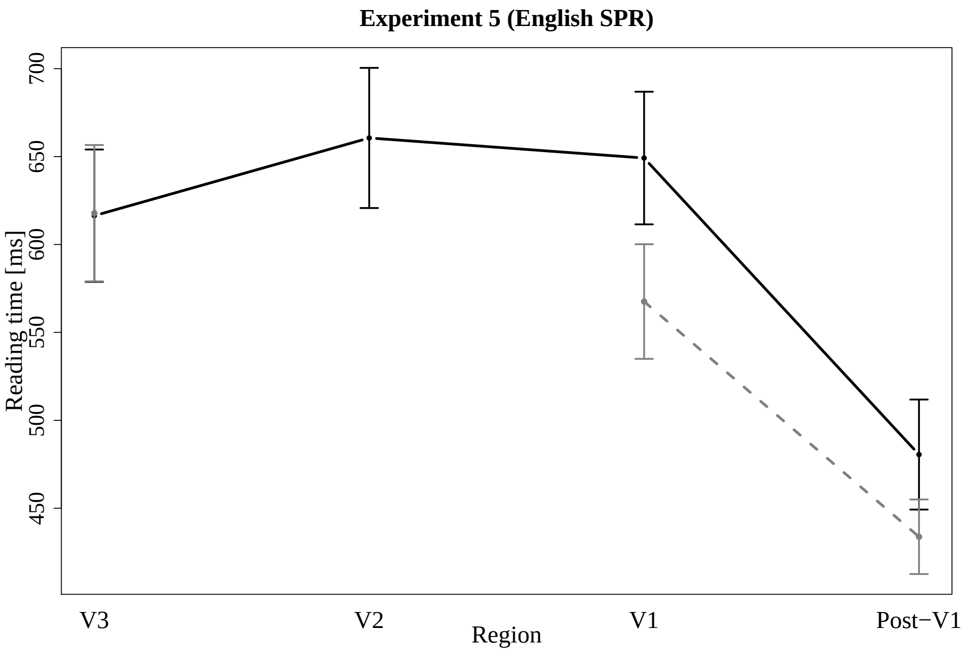


Figure 7. Mean reading times and 95% confidence intervals for the critical regions in the English SPR study with commas (experiment 5).

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	0.05609	0.01547	3.62 *	0.02578169	0.08650596
Post-V1	0.04302	0.01473	2.92 *	0.01313406	0.07198286

Table 6: Summary of results of the two comparisons for the (English SPR) experiment 5. The asterisk (*) indicates statistical significance at $\alpha = 0.05$.

Thus, in English, commas alone do not appear to help in maintaining predictions of the upcoming verbs. However, this does not entail that the German pattern is completely unrelated to the presence of the commas. Because commas are obligatory in German relative clauses but not in English, German speakers could be highly practised in using the comma as a cue for building predictions of upcoming verbs, whereas the English speakers may not be.

One possible objection to the above experiment is that the presence of the comma in the relative clause beginning with *that* is not felicitous—in English, commas are used only with non-restrictive relative clauses. In order to address this objection, we carried out a second study where we replaced the *that* with a *who*; see 6. This experiment is described next.

(6) a. The carpenter, who the craftsman, who the peasant carried, hurt, supervised the apprentice.

b. The carpenter, who the craftsman, who the peasant carried, supervised the apprentice.

Experiment 6

Method

Participants.

Forty one English native-speakers from the University of Dundee participated in the experiment. They received course credit or monetary compensation for taking part. All had normal or corrected-to-normal vision.

Procedure and Design.

The procedure was as described in earlier SPR studies. The design was identical to the previous experiments with the difference that the relative pronoun was always *who*.[8]

Results and Discussion

Response accuracies.

The overall mean response-accuracy (including all items) was 79%. The accuracy for the grammatical sentences was 55% and for the ungrammatical sentences the accuracy was 61%; this difference did not reach significance.

[8] Please see the appendix for a discussion of a further manipulation carried out on question type in response to a reviewer's comments.

Reading time.

Regarding the reading times, as summarized in Figure 8 and Table 7, the English reading pattern did not change substantially due to the presence of commas. As in the previous experiments involving English, the ungrammatical sentences were read faster in the region following the final verb. A statistically significant effect was not found at V1 but the sign of the coefficient was consistent with the previous English studies. The reading time at the region following V1 showed a significantly longer RT in the grammatical condition, consistent with the preceding studies on English.

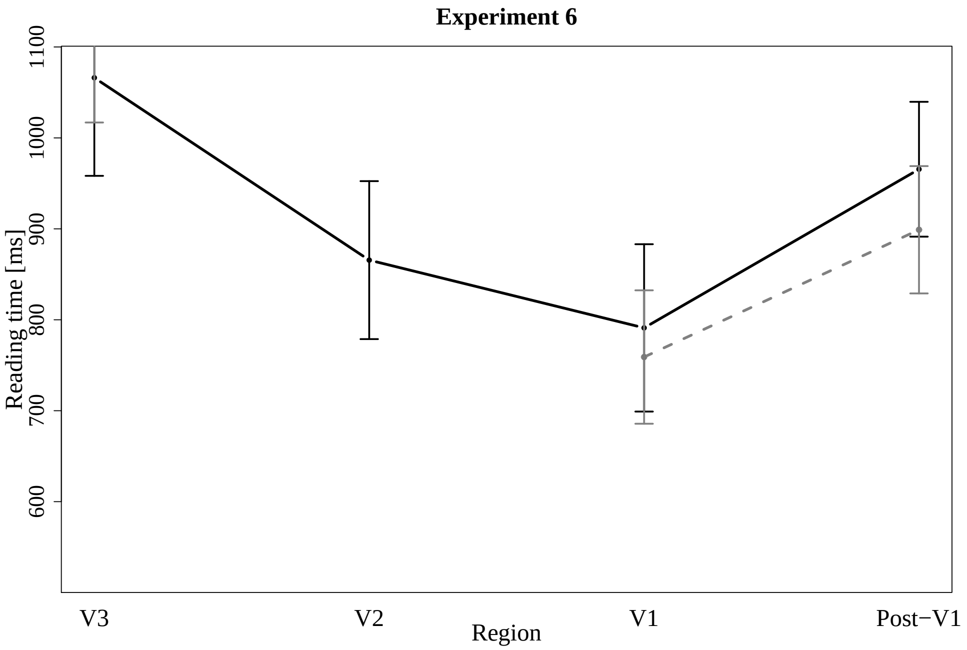


Figure 8. Mean reading times and 95% confidence intervals for the critical regions in the English SPR study with commas (experiment 6).

region	coefficient	se	t-score	hpd.lower	hpd.upper
V1	-0.00675	0.01408	-0.48	-0.0378045	0.01832894
Post-V1	-0.04304	0.01525	-2.82 *	-0.0736196	-0.01280071

Table 7: Summary of results of the two comparisons for the (English SPR) experiment 6. The asterisk (*) indicates statistical significance at $\alpha = 0.05$.

Thus, in English, the presence of commas alone does not appear to help in processing the grammatical structures. However, this does not entail that the German pattern is completely unrelated to the presence of the commas. Because commas are obligatory in German relative clauses but not in English, German speakers could be highly practised in using the comma as a cue for building predictions of upcoming verbs, whereas the English speakers may not be.

In other words, the reading pattern found in German could be due to its head-final nature as well as to the obligatory presence of commas in relative clauses. In either case, the evidence suggests that the parsing mechanism is not invariant across languages: the activation of a predicted node in memory is conditioned by the statistical properties of the language being parsed. Specifically, we have demonstrated that the ability to maintain predictions in memory for VP nodes corresponding to as-yet-unseen verbs is dependent on the language-specific factors such as head-finality and possibly also the use of punctuation to mark clause breaks.

General Discussion

We have presented seven experiments involving English and German center embeddings using two methods—self-paced reading and eyetracking—and shown that (i) English readers experience the grammaticality illusion in online processing; (ii) German speakers do not in general experience the grammaticality illusion; and (iii) the presence of commas in English does not appear to save the English readers from the illusion.

Regarding the VP-forgetting hypothesis of Gibson and Thomas and DLT, the English self-paced reading studies (experiments 1 and 1a) and the eyetracking study (experiment 2) were remarkably consistent with the explanation provided by Gibson and Thomas. Reading times were longer at the matrix verb V1 and the Post-V1 region in the grammatical condition compared to the ungrammatical one. Only in experiment 1 did the difference fail to reach statistical significance for the V1 region, but even there the pattern seen was consistent with the other English experiment results.

The German data, however, pose a significant problem for DLT and the VP-forgetting hypothesis as stated in (Gibson & Thomas, 1999): in both SPR and eyetracking, at the matrix verb V1 and the Post-V1 word the grammatical condition had *faster* reading times than the ungrammatical one. Only in experiment 4 did this difference fail to reach statistical significance for the Post-V1 region. German speakers apparently tend to not forget the predicted middle verb-phrase node.

As discussed earlier, one plausible explanation for this difference in behavior is the difference in word order between German and English. One consequence of German head-finality is that—due to the relatively frequent occurrence of head-final structures—predictions of upcoming verbs may have more robust memory representations in German than in English. This could result in reduced susceptibility to forgetting the upcoming verb's prediction, even in the face of increased memory load.

What do these results mean for the VP-forgetting explanation and decay-based theories in general? Clearly, whether the prediction is forgotten or not depends on the *a priori* grammatical constraints in the language (in this case head-finality and perhaps also the obligatory presence of commas in German relative clauses), which in turn affect structural frequency patterns of the language. Predictions of verb phrases could be maintained more robustly in head-final languages than non-head-final languages, and this could determine whether or not a prediction is forgotten. Importantly, memory overload as defined in different versions of locality-based theories does not appear to be an invariant, language-independent constraint operating on online sentence comprehension; it can be modulated and as we have shown, even reversed, by experience. In related work, Engelmann (2009); Engelmann and Vasishth (2009) show how a simple recurrent network based model can account for both the English and German findings in this paper; also see Christiansen & Macdonald, 2009 for a model of their earlier English results reported originally in (Christiansen & Chater, 2001).

This role of experience is also relevant for ongoing debates regarding the role of expectation in incremental parsing (Levy, 2008). If the ability to maintain a prediction for an upcoming VP is modulated by the frequency with which a verb appears clause-finally, one important implication is that the head-final structures in English and German should not be predicted to exhibit the same pattern. Interestingly, recent work by Jaeger, Fedorenko, Hofmeister, and Gibson (2008) has argued precisely that head-final structures in English and German should behave similarly with regard to expectations of upcoming heads. In a series of self-paced reading experiments, Jaeger and colleagues showed quite conclusively that increasing the distance between an argument and a verb in English by interposing material (adjuncts such as prepositional phrases) between them results in faster reading time at the verb (rather than slower reading time, as predicted by locality theories). Their explanation for the speedup is the heightened expectation for the verb as distance is increased. Jaeger and colleagues explain in a similar way speedups observed in German (Konieczny, 2000) and Hindi (Vasishth & Lewis, 2006). Now, if—as the present work suggests—predicted VP nodes are maintained with differing degrees of robustness in English and German, integrating the highly-expected verb in English with the predicted VP node should be costlier than integrating the highly expected verb in German with the predicted VP: German and English should not both show similar antilocality effects. It is an interesting open question how the present results can be reconciled with expectation-based accounts such as Levy's and Jaeger et al's.

The findings presented here have rather broad implications for theories of sentence comprehension. Specifically, they appear to be inconsistent with the view that the human parsing mechanism can predominantly be characterized as a universal set of mechanisms and constraints that can be applied to languages; the role of purely grammatical constraints (which express themselves as experience) appears to be under-represented in such theories. Rather, the constraints on parsing—for example, the ability to maintain a verb-phrase prediction in memory—seem to be greatly dependent on independently determined grammatical properties of the language in question. In the present case, the relevant grammatical property is head-finality and possibly also punctuation rules (which in this case can be considered to be part of the grammar of written communication), but in principle it could be anything relevant for online, incremental parsing decisions: morphological cues, case-marking, agreement properties, and so on (MacWhinney, Bates, & Kliegl, 1984). It is uncontroversial, for example, that an accusative case-marked sentence initial noun phrase in German would be construed as an object and not a subject, with the result that a subject is predicted to occur later on. This predictive ability is a function of the grammatical properties of German and emerges from exposure to the language.

Of course, our results do not imply that there is no language-independent core in the parsing mechanism. It seems very reasonable to assume that the parsing mechanism is an emergent system constrained by a combination of task demands and independent properties of the human brain. Under this view, the parser would be seen as an adaptive system that changes depending on the kinds of actions it has to perform. For example, if predicting upcoming verbs is a frequent requirement for a language, the parsing mechanism would become better at it. There is ample independent evidence for such an adaptive parsing system. For example, the kinds of cues used in parsing depend on the language: parsing German would involve the use of case-marking and agreement cues to a much greater extent than English (MacWhinney et al., 1984).

How do such differential abilities come about, and are such adaptations an overlay on a core parsing architecture? These are important and complicated questions which are beyond the scope of this paper. Answering these questions would require developing articulated models specifying how the parsing mechanism emerges, and identifying the precise role of grammatical knowledge and experience in shaping the emergent parsing mechanism. Connectionist models of parsing (e.g., MacDonald & Christiansen, 2002; Christiansen & Macdonald, 2009; Engelmann & Vasishth, 2009) provide one framework for addressing this question, but clearly much more work is necessary before we can begin to understand how underlying grammatical properties of a language shape the parser's actions and constraints.

References

Baayen, H. (2008). *Practical data analysis for the language sciences.* Cambridge University Press.

Bach, E., Brown, C., & Marslen-Wilson, W. (1986). Crossed and nested dependencies in German and Dutch: A psycholinguistic study. *Language and Cognitive Processes*, *1(4)*, 249–262.

Christiansen, M. H., & Chater, N. (2001). Finite models of infinite language: A connectionist approach to recursion. In M. H. Christiansen & N. Chater (Eds.), *Connectionist psycholinguistics* (p. 138-176). Westport, Connecticut: Ablex Publishing.

Christiansen, M. H., & Macdonald, M. (2009). *A usage-based approach to recursion in sentence processing.* (Submitted)

Clark, H. (1973). The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–59.

Clifton, C., Staub, A., & Rayner, K. (in press). Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, & R. L. Hill (Eds.), *Eye movements: A window on mind and brain* (chap. 15). Elsevier.

Engelmann, F. (2009). *Connectionist modeling of experience-based effects in sentence comprehension.* Unpublished master's thesis, University of Potsdam, Potsdam, Germany. (Downloadable from http://www.ling.uni-potsdam.de/∼/vasishth/Papers/engelmanndiploma09.pdf)

Engelmann, F., & Vasishth, S. (2009). Processing grammatical and ungrammatical center embeddings in English and German: A computational model. In A. Howes, D. Peebles, & R. Cooper (Eds.), *Proceedings of 9th International Conference on Cognitive Modeling.* Manchester, UK.

Frazier, L. (1985). Syntactic complexity. In D. R. Dowty, L. Kartunnen, & A. M. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 129–189). Cambridge University Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, UK: Cambridge University Press.

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

Gibson, E. (2000). Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, & W. O'Neil (Eds.), *Image, language, brain: Papers from the first mind articulation project symposium.* Cambridge, MA: MIT Press.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14(3)*, 225–248.

Gimenes, M., Rigalleau, F., & Gaonac'h, D. (2009). When a missing verb makes a French sentence more acceptable. *Language and Cognitive Processes*, *24*, 440–449.

Gordon, P. C., Hendrick, R., Johnson, M., & Lee, Y. (2006). Similarity-Based Interference During Language Comprehension: Evidence from Eye Tracking During Reading. *Journal of Experimental Psychology: Learning Memory and Cognition*, *32*(6), 1304–1321.

Inhoff, A. W. (1984). Two stages of word processing during eye fixations in the reading of prose. *Journal of verbal learning and verbal behavior*, *23*(5), 612–624.

Jaeger, F. T., Fedorenko, E., Hofmeister, P., & Gibson, E. (2008). Expectation-based syntactic processing: Antilocality outside of head-final languages. In *Cuny sentence processing conference.* North Carolina.

Joshi, A. K., Becker, T., & Rambow, O. (2000). Complexity of scrambling: A new twist to the competence-performance distinction. In A. Abeillé & O. Rambow (Eds.), *Tree adjoining grammars* (pp. 167–181). CSLI.

Just, M., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111(2)*, 228–238.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29(6)*, 627–645.

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

Lewis, R. L., & Vasishth, S. (2005, May). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1-45.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: A reply to Just and Carpenter and Waters and Caplan. *Psychological Review*, *109*(1), 35–54.

MacWhinney, B., Bates, E., & Kliegl, R. (1984). Cue validity and sentence interpretation in English, German. *and Italian. Journal of Verbal Learning and Verbal Behavior*, *23*, 127–150.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods of investigating immediate processes in reading. In D. E. Kieras & M. Just (Eds.), *New methods in reading comprehension research.* Hillsdale, N.J.: Erlbaum.

Raaijmakers, J., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with the "language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*(3), 416–426.

Rambow, O., & Joshi, A. K. (1994). A processing model for free word order languages. In C. C. Jr., L. Frazier, & K. Rayner (Eds.), *Perspectives on sentence processing.* Hillsdale, NJ: L. Erlbaum.

Rayner, K., & Pollatsek, A. (1987). Eye movements in reading: A tutorial review. *Attention and performance XII: The psychology of reading*, 327–362.

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

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*(4), 767-794.

Appendix A
Experiments 1 and 2 stimuli

(1) a. The carpenter who the craftsman that the peasant carried hurt supervised the apprentice.
b. Did the peasant carry the craftsman? Y

(2) a. The carpenter who the pillar that the peasant carried hurt supervised the apprentice.
b. Did the pillar hurt the carpenter? Y

(3) a. The carpenter who the craftsman that the peasant carried supervised the apprentice.
b. Did the carpenter supervise the apprentice? Y

(4) a. The carpenter who the pillar that the peasant carried supervised the apprentice.
b. Did the peasant carry the pillar? Y

(5) a. The mother who the daughter that the sister found frightened greeted the grandmother.
b. Did the sister find the daughter? Y

(6) a. The mother who the gun that the sister found frightened greeted the grandmother.
b. Did the gun frighten the mother? Y

(7) a. The mother who the daughter that the sister found greeted the grandmother.
b. Did the mother greet the grandmother? Y

(8) a. The mother who the gun that the sister found greeted the grandmother.
b. Did the sister find the gun? Y

(9) a. The worker who the tenant that the foreman looked for injured questioned the shepherd.
b. Did the foreman look for the tenant? Y

(10) a. The worker who the bucket that the foreman looked for injured questioned the shepherd.
b. Did the bucket injure the worker? Y

(11) a. The worker who the tenant that the foreman looked for questioned the shepherd.
b. Did the worker question the shepherd? Y

(12) a. The worker who the bucket that the foreman looked for questioned the shepherd.
b. Did the foreman look for the bucket? Y

(13) a. The trader who the businessman that the professor hired confused annoyed the investor.

b. Did the professor hire the businessman? Y

(14) a. The trader who the computer that the professor hired confused annoyed the investor.

b. Did the computer confuse the trader? Y

(15) a. The trader who the businessman that the professor hired annoyed the investor.

b. Did the trader annoy the investor? Y

(16) a. The trader who the computer that the professor hired annoyed the investor.

b. Did the professor hire the computer? Y

(17) a. The painter who the musician that the father missed sheltered cooked for the artist.

b. Did the father miss the musician? Y

(18) a. The painter who the hut that the father missed sheltered cooked for the artist.

b. Did the hut shelter the painter? Y

(19) a. The painter who the musician that the father missed cooked for the artist.

b. Did the painter cook for the artist? Y

(20) a. The painter who the hut that the father missed cooked for the artist.

b. Did the father miss the hut? Y

(21) a. The saxophonist who the trumpeter that the conductor brought along distracted thanked the violinist.

b. Did the conductor bring along the trumpeter? Y

(22) a. The saxophonist who the baton that the conductor brought along distracted thanked the violinist.

b. Did the baton distract the saxophonist? Y

(23) a. The saxophonist who the trumpeter that the conductor brought along thanked the violinist.

b. Did the saxophonist thank the violinist? Y

(24) a. The saxophonist who the baton that the conductor brought along thanked the violinist.

b. Did the conductor bring along the baton? Y

(25) a. The pharmacist who the optician that the stranger saw troubled questioned the customer.

b. Did the stranger see the optician? Y

(26) a. The pharmacist who the button that the stranger saw troubled questioned the customer.

b. Did the button trouble the pharmacist? Y

(27) a. The pharmacist who the optician that the stranger saw questioned the customer.

b. Did the pharmacist question the customer? Y

(28) a. The pharmacist who the button that the stranger saw questioned the customer.

b. Did the stranger see the button? Y

(29) a. The cleaner who the janitor that the doctor recognized hurt surprised the patient.

b. Did the doctor recognize the janitor? Y

(30) a. The cleaner who the ball that the doctor recognized hurt surprised the patient.

b. Did the ball hurt the cleaner? Y

(31) a. The cleaner who the janitor that the doctor recognized surprised the patient.

b. Did the cleaner surprise the patient? Y

(32) a. The cleaner who the ball that the doctor recognized surprised the patient.

b. Did the doctor recognize the ball? Y

(33) a. The dancer who the singer that the bystander admired hurt tipped the doorman.

b. Did the singer admire the bystander? N

(34) a. The dancer who the shoe that the bystander admired hurt tipped the doorman.

b. Did the shoe pinch the bystander? N

(35) a. The dancer who the singer that the bystander admired tipped the doorman.

b. Did the singer tip the doorman? N

(36) a. The dancer who the shoe that the bystander admired tipped the doorman.

b. Did the bystander admire the dancer? N

(37) a. The artist who the sportsman that the guard shouted at annoyed instructed the newscaster.

b. Did the sportsman shout at the guard? N

(38) a. The artist who the computer that the guard shouted at annoyed instructed the newscaster.

b. Did the computer annoy the guard? N

(39) a. The artist who the sportsman that the guard shouted at instructed the newscaster.

b. Did the sportsman instruct the newscaster? N

(40) a. The artist who the computer that the guard shouted at instructed the newscaster.

b. Did the guard shout at the artist? N

(41) a. The clerk who the bureaucrat that the visitor forgotten about helped annoyed the neighbor.

b. Did the bureaucrat forget about the visitor? N

(42) a. The clerk who the walking stick that the visitor forgotten about helped annoyed the neighbor.

b. Did the walking stick help the visitor? N

(43) a. The clerk who the bureaucrat that the visitor forgotten about annoyed the neighbor.

b. Did the bureaucrat annoy the neighbor? N

(44) a. The clerk who the walking stick that the visitor forgotten about annoyed the neighbor.

b. Did the visitor forget about the clerk? N

(45) a. The son who the father that the teacher saw disturbed visited the grandfather.

b. Did the father see the teacher? N

(46) a. The son who the loudspeaker that the teacher saw disturbed visited the grandfather.

b. Did the loudspeaker disturb the teacher? N

(47) a. The son who the father that the teacher saw visited the grandfather.

b. Did the father visit the grandfather? N

(48) a. The son who the loudspeaker that the teacher saw visited the grandfather.

b. Did the teacher see the son? N

(49) a. The conductor who the choirmaster that the worker ignored hit berated the musician.

b. Did the choirmaster ignore the worker? N

(50) a. The conductor who the sponge that the worker ignored hit berated the musician.

b. Did the sponge hit the worker? N

(51) a. The conductor who the choirmaster that the worker ignored berated the musician.

b. Did the choirmaster berate the musician? N

(52) a. The conductor who the sponge that the worker ignored berated the musician.

b. Did the worker ignore the conductor? N

(53) a. The defence who the prosecutor that the spy looked at surprised convinced the judge.

b. Did the prosecutor look at the spy? N

(54) a. The defence who the knife that the spy looked at surprised convinced the judge.

b. Did the knife surprise the spy? N

(55) a. The defence who the prosecutor that the spy looked at convinced the judge.

b. Did the prosecutor convince the judge? N

(56) a. The defence who the knife that the spy looked at convinced the judge.

b. Did the spy look at the defence? N

(57) a. The cousin who the brother that the peasant described pleased hated the uncle.

b. Did the brother describe the peasant? N

(58) a. The cousin who the diamond that the peasant described pleased hated the uncle.

b. Did the diamond please the peasant? N

(59) a. The cousin who the brother that the peasant described hated the uncle.

b. Did the brother hate the uncle? N

(60) a. The cousin who the diamond that the peasant described hated the uncle.

b. Did the peasant describe the cousin? N

(61) a. The painter who the musician that the friend liked disturbed admired the poet.

b. Did the musician like the friend? N

(62) a. The painter who the film that the friend liked disturbed admired the poet.

b. Did the film disturb the friend? N

(63) a. The painter who the musician that the friend liked admired the poet.

b. Did the musician admire the poet? N

(64) a. The painter who the film that the friend liked admired the poet.

b. Did the friend like the painter? N

Appendix B
Experiments 3 and 4 stimuli (German SPR and eyetracking)

(1) a. Der Anwalt, den der Zeuge, den der Spion betrachtete, schnitt, überzeugte den Richter.
b. Hat der Zeuge den Spion betrachtet? N

(2) a. Der Anwalt, den der Säbel, den der Spion betrachtete, schnitt, überzeugte den Richter.
b. Hat der Säbel den Spion betrachtet? N

(3) a. Der Anwalt, den der Zeuge, den der Spion betrachtete, überzeugte den Richter.
b. Hat der Zeuge den Spion betrachtet? N

(4) a. Der Anwalt, den der Säbel, den der Spion betrachtete, überzeugte den Richter.
b. Hat der Spion den Anwalt betrachtet? N

(5) a. Der Beamte, den der Bürokrat, den der Besucher vergass, stützte, verärgerte den Nachbarn.
b. Hat der Bürokrat den Besucher vergessen? N

(6) a. Der Beamte, den der Tisch, den der Besucher vergass, stützte, verärgerte den Nachbarn.
b. Hat der Tisch den Besucher gestützt? N

(7) a. Der Beamte, den der Bürokrat, den der Besucher vergass, verärgerte den Nachbarn.
b. Hat der Bürokrat den Nachbarn verärgert? N

(8) a. Der Beamte, den der Tisch, den der Besucher vergass, verärgerte den Nachbarn.
b. Hat der Besucher den Beamten vergessen? N

(9) a. Der Bräutigam, den der Schwiegervater, den der Musiker trug, ablenkte, begrüsste den Pfarrer.
b. Hat der Schwiegervater den Musiker getragen? N

(10) a. Der Bräutigam, den der Bilderrahmen, den der Musiker trug, ablenkte, begrüsste den Pfarrer.
b. Hat der Bilderrahmen den Musiker abgelenkt? N

(11) a. Der Bräutigam, den der Schwiegervater, den der Musiker trug, begrüsste den Pfarrer.
b. Hat der Schwiegervater den Pfarrer begrüsst? N

(12) a. Der Bräutigam, den der Bilderrahmen, den der Musiker trug, begrüsste den Pfarrer.

b. Hat der Musiker den Bräutigam getragen? N

(13) a. Der Bruder, den der Cousin, den der Bauer fand, entzückte, hasste den Onkel.

b. Hat der Cousin den Bauer gefunden? N

(14) a. Der Bruder, den der Schmuck, den der Bauer fand, entzückte, hasste den Onkel.

b. Hat der Schmuck den Bauer entzückt? N

(15) a. Der Bruder, den der Cousin, den der Bauer fand, hasste den Onkel.

b. Hat der Cousin den Onkel gehasst? N

(16) a. Der Bruder, den der Schmuck, den der Bauer fand, hasste den Onkel.

b. Hat der Bauer den Bruder gefunden? N

(17) a. Der Zauberer, den der Akrobat, den der Zuschauer beschrieb, ärgerte, besuchte den Zirkusdirektor.

b. Hat der Akrobat den Zuschauer beschrieben? N

(18) a. Der Zauberer, den der Hut, den der Zuschauer beschrieb, ärgerte, besuchte den Zirkusdirektor.

b. Hat der Hut den Zuschauer geärgert? N

(19) a. Der Zauberer, den der Akrobat, den der Zuschauer beschrieb, besuchte den Zirkusdirektor.

b. Hat der Akrobat den Zirkusdirektor besucht? N

(20) a. Der Zauberer, den der Hut, den der Zuschauer beschrieb, besuchte den Zirkusdirektor.

b. Hat der Zuschauer den Zauberer beschrieben? N

(21) a. Der Einbrecher, den der Dieb, den der Mann beschützte, bezauberte, beschuldigte den Komplizen.

b. Hat der Dieb den Mann beschützt? N

(22) a. Der Einbrecher, den der Stein, den der Mann beschützte, bezauberte, beschuldigte den Komplizen.

b. Hat der Stein den Mann bezaubert? N

(23) a. Der Einbrecher, den der Dieb, den der Mann beschützte, beschuldigte den Komplizen.

b. Hat der Dieb den Komplizen beschuldigt? N

(24) a. Der Einbrecher, den der Stein, den der Mann beschützte, beschuldigte den Komplizen.

b. Hat der Mann den Einbrecher beschützt? N

(25) a. Der Neurotiker, den der Exzentriker, den der Psychiater bezahlte, ängstigte, versetzte den Berater.

b. Hat der Psychiater den Exzentriker bezahlt? Y

(26) a. Der Neurotiker, den der Dolch, den der Psychiater bezahlte, ängstigte, versetzte den Berater.

b. Hat der Dolch den Neurotiker geängstigt? Y

(27) a. Der Neurotiker, den der Exzentriker, den der Psychiater bezahlte, versetzte den Berater.

b. Hat der Neurotiker den Berater versetzt? Y

(28) a. Der Neurotiker, den der Dolch, den der Psychiater bezahlte, versetzte den Berater.

b. Hat der Psychiater den Dolch bezahlt? Y

(29) a. Der Arbeiter, den der Monteur, den der Vorarbeiter vergass, verletzte, beschimpfte den Passanten.

b. Hat der Vorarbeiter den Monteur vergessen? Y

(30) a. Der Arbeiter, den der Eimer, den der Vorarbeiter vergass, verletzte, beschimpfte den Passanten.

b. Hat der Eimer den Arbeiter verletzt? Y

(31) a. Der Arbeiter, den der Monteur, den der Vorarbeiter vergass, beschimpfte den Passanten.

b. Hat der Arbeiter den Passanten beschimpft? Y

(32) a. Der Arbeiter, den der Eimer, den der Vorarbeiter vergass, beschimpfte den Passanten.

b. Hat der Vorarbeiter den Eimer vergessen? Y

(33) a. Der Banker, den der Kreditgeber, den der Kunde mochte, nervte, bestahl den Vermieter.

b. Hat der Kunde den Kreditgeber gemocht? Y

(34) a. Der Banker, den der Geldautomat, den der Kunde mochte, nervte, bestahl den Vermieter.

b. Hat der Geldautomat den Banker genervt? Y

(35) a. Der Banker, den der Kreditgeber, den der Kunde mochte, bestahl den Vermieter.

b. Hat der Banker den Vermieter bestohlen? Y

(36) a. Der Banker, den der Geldautomat, den der Kunde mochte, bestahl den Vermieter.

b. Hat der Kunde den Geldautomat gemocht? Y

(37) a. Der Pianist, den der Cellist, den der Hausmeister sah, traf, ersetzte den Violinisten.

b. Hat der Hausmeister den Cellist gesehen? Y

(38) a. Der Pianist, den der Ball, den der Hausmeister sah, traf, ersetzte den Violinisten.

b. Hat der Ball den Pianist getroffen? Y

(39) a. Der Pianist, den der Cellist, den der Hausmeister sah, ersetzte den Violinisten.

b. Hat der Pianist den Violinisten ersetzt? Y

(40) a. Der Pianist, den der Ball, den der Hausmeister sah, ersetzte den Violinisten.

b. Hat der Hausmeister den Ball gesehen? Y

(41) a. Der Anwohner, den der Wanderer, den der Pförtner suchte, störte, verarztete den Verletzten.

b. Hat der Pförtner den Wanderer gesucht? Y

(42) a. Der Anwohner, den der Stuhl, den der Pförtner suchte, störte, verarztete den Verletzten.

b. Hat der Stuhl den Anwohner gestört? Y

(43) a. Der Anwohner, den der Wanderer, den der Pförtner suchte, verarztete den Verletzten.

b. Hat der Anwohner den Verletzten verarztet? Y

(44) a. Der Anwohner, den der Stuhl, den der Pförtner suchte, verarztete den Verletzten.

b. Hat der Pförtner den Stuhl gesucht? Y

(45) a. Der Tänzer, den der Artist, den der Zuschauer bewunderte, drückte, beobachtete den Einlasser.

b. Hat der Zuschauer den Artist bewundert? Y

(46) a. Der Tänzer, den der Schuh, den der Zuschauer bewunderte, drückte, beobachtete den Einlasser.

b. Hat der Schuh den Tänzer gedrückt? Y

(47) a. Der Tänzer, den der Artist, den der Zuschauer bewunderte, beobachtete den Einlasser.

b. Hat der Tänzer den Einlasser beobachtet? Y

(48) a. Der Tänzer, den der Schuh, den der Zuschauer bewunderte, beobachtete den Einlasser.

b. Hat der Zuschauer den Schuh bewundert? Y

Appendix C
A note on the questions in Experiment 6

Regarding Experiments 1-5, an anonymous reviewer expressed the concern that the questions that followed the ungrammatical conditions may have provided a cue to participants to process the ungrammatical sentences in a certain way. In order to investigate this possibility, in experiment 6 we manipulated the patterns in the questions as follows: half the participants were presented with items involving questions that had the following biased pattern (condition a is the grammatical sentence, b the ungrammatical):

```
YES
a) Did N3 V3 N2?
b) Did N1 V1 N4?
NO
a) Did N2 V3 N3?
b) Did N2 V1 N4?
```

The other half of the participants were shown questions with the following unbiased pattern:

```
YES:
sentences 1-4
a) N1 V1 N4?
b) N3 V3 N2?
sentences 5-8
a) N3 V3 N2?
b) N1 V1 N4?
NO:
sentences 9-12
a) N2 V3 N3?
b) N3 V1 N4?
sentences 13-16
a) N3 V1 N4?
b) N2 V3 N3?
```

We reasoned that if question-response strategy were a factor affecting the pattern of the data, then we would see an interaction between question type and the grammaticality manipulation; we found no such interaction.