Title
Using the Interpolated Maze Task to Assess Incremental Processing in English Relative Clauses

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 43(43)

ISSN
1069-7977

Authors
Vani, Pranali
Wilcox, Ethan Gotlieb
Levy, Roger

Publication Date
2021

Peer reviewed

Using the Interpolated Maze Task to Assess Incremental Processing in English Relative Clauses

Pranali Vani, Ethan Gotlieb Wilcox and Roger Levy

Abstract

In English, Subject Relative Clauses are processed more quickly than Object Relative Clauses, but open questions remain about where in the clause slowdown occurs. The *surprisal theory* of incremental processing, under which processing difficulty corresponds to probabilistic expectations about upcoming material, predicts that slowdown should occur immediately on material that disambiguates the subject from object relative clause. However, evidence from eye tracking and self-paced reading studies suggests that slowdown occurs downstream of RC-disambiguating material, on the relative clause verb. These methods, however, suffer from well-known *spillover effects* which makes their results difficult to interpret. To address these issues, we introduce and deploy a novel variant of the Maze task for reading times (Forster, Guerrera, & Elliot, 2009), called the *Interpolated Maze* in two English web-based experiments. In Experiment 1, we find that the locus of reading-time differences between SRCs and ORCs falls on immediate disambiguating definite determiner. Experiment 2 provides a control, showing that ORCs are read more slowly than lexically-matching, non-anomalous material. These results provide new evidence for the locus of processing difficulty in relative clauses and support the surprisal theory of incremental processing.

Keywords: Incremental Processing, Maze Task, Relative Clause, Surprisal Theory

Introduction

A large amount of evidence shows that, in English, Subject Relative Clauses (SRCs) are easier to process than Object Relative Clauses (ORCs), as measured in self-paced reading tasks (King & Just, 1991; Grodner & Gibson, 2005), eye-tracking (Staub, 2010; Traxler, Morris, & Seely, 2002) and other online incremental measures such as the Maze task (Forster et al., 2009). Subject RCs, as in (1-a) are so called because the head of the NP has been extracted from the subject position of the subordinated phrase, whereas for Object RCs, as in (1-b), the NP has been extracted from object position.

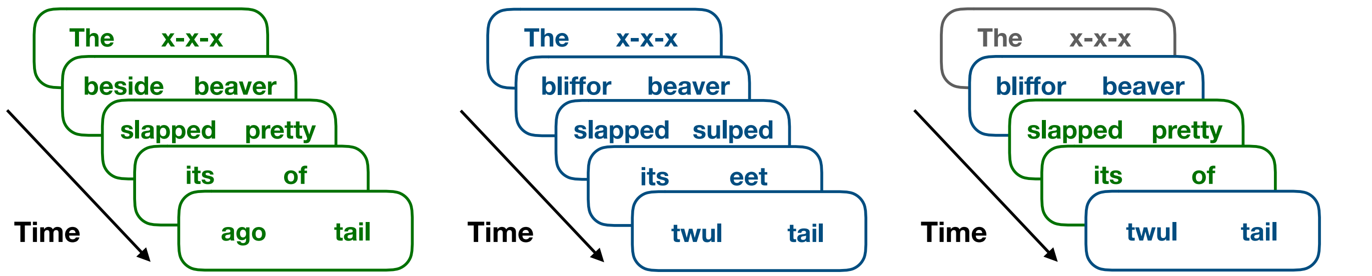
(1) a. The banker that ___ irritated the lawyer played tennis every Sunday.

b. The banker that the lawyer irritated ___ played tennis every Sunday.

One candidate hypothesis for the differential in processing times for SRCs and ORCs comes from *surprisal theory* of incremental processing (Hale, 2001; Levy, 2008), which has its roots in the information theory of Shannon (1948). According to surprisal theory, comprehenders use linguistic experience to make probabilistic predictions about upcoming material. The theory predicts that less probable material will be less expected and therefore more difficult to process. This theory accounts for many observed higher-level processing difficulties across a variety of grammatical constructions (Levy, Fedorenko, Breen, & Gibson, 2012) and for reading-times in naturalistic datasets (Smith & Levy, 2013; Wilcox, Gauthier, Hu, Qian, & Levy, 2020). Importantly, both experience and upcoming predictions can be modulated by syntactic structures, so expectations for the next word, following the context `The banker that...` are dependent on syntactic experience of the distribution of SRCs and ORCs. Consistent with this account, modulation of the SRC/ORC difficulty difference that is obtained by varying the animacy and/or pronominality of the head noun and/or RC NP is generally predictable from corpus frequencies (Traxler et al., 2002; Mak, Vonk, & Schriefers, 2002; Reali & Christiansen, 2007; MacDonald, 2013).

Surprisal theory and its competitors further makes specific predictions about the *locus* of processing difficulty. If comprehenders expect SRCs, then following the prefix `The banker that...` they should expect a tensed verb. Determiners, which unambiguously introduce ORCs, should be unexpected and processed more slowly. But previous experimental evidence from the most widely used behavioral measures for online processing difficulty, does not clearly follow these predictions. In eye tracking during reading, slowdowns for ORCs are largest at the RC noun and verb (Staub, 2010), and for self-paced reading, the slowdowns seem to be largest at the RC verb (Grodner & Gibson, 2005). These data raise questions about the reliability of surprisal theory to accurately predict incremental processing times for relative clauses. Notably, the competing Dependency Locality Theory (DLT; Gibson, 2000) predicts that the SRC-over-ORC difficulty differential will manifest at the RC verb. Thus, precisely *where* the difficulty of ORCs manifests is of considerable theoretical importance.

However, both eye-tracking and (especially) self-paced reading show spillover effects, where the difficulty posed by a a particular word manifests in downstream words (Ehrlich & Rayner, 1983). The Maze task (Forster et al., 2009), in contrast, is intended to minimize spillover effects. In the Maze Task, participants navigate through a sentence displayed on a monitor by pressing letters on their keyboard. At each sen-



Grammatical Maze (G-Maze) Lexical Maze (L-Maze) Interpolated Maze (I-Maze)

Figure 1: Three variants of the Maze Task. At each index, participants have to select the correct next-word in the sentence..

tence index, participants must choose between two possible continuations, one of which is a possible next-word of the sentence, the other of which is not. The time it takes to select the right continuation has been shown to be a good measure of incremental processing difficulty, and does not suffer from spillover effects (Boyce & Levy, 2020).[1] Forster et al. (2009) found that in both Grammatical and Lexical versions of the Maze task (Figure 1) for SRCs and ORCs, the difficulty differential shows up primarily at the RC NP's determiner. However, this study has never been replicated. Here, we extend these initial findings of Forster et al. in two ways. First, we replicate these findings using a novel Maze variant, the *Interpolated Maze*, which combines the advantages of the Lexical and Grammatical Maze tasks. The Interpolated Maze weaves together Grammatical and Lexical distractors, forcing participants to represent context (to succeed when the distractor is a word) but ensuring at theoretically critical positions in the sentence that the distractor is impossible to integrate with the context (because it is a non-word). Second, we use the Interpolated Maze to replicate a closely related result of Staub (2010) which addresses the word-order confound intrinsic to the SRC/ORC contrast.

Experiment 1, adapted from Traxler et al. (2002), compares processing times in SRCs vs. ORCs. We find that ORCs are processed more slowly than SRCs, and that the contrast is entirely due to differences in reading times on the RC determiners. Experiment 2, adapted from Staub (2010), provides a control for Experiment 1, by comparing ORCs to matching material in non-anomalous sentences that involve clausal embedding. We find, again, that ORCs are processed more slowly and that a large part of the incremental processing difficulty arises on the relative clause determiner. Both of these results are consistent with surprisal theory, and not with DLT.

General Methods

Interpolated Maze

We collected human incremental processing data from a novel implementation of the Maze Task (Forster et al., 2009) which we call the *Interpolated Maze*. In this task, experimental subjects read through a sentence on a monitor. At each sentence index they are presented with two possible continuations, and they must select the correct next-word by pressing a key on their keyboard. If participants select the wrong continuation, the trial ends and they are taken to the next sentence. Figure 1 shows a cartoon of this process for the three variants of the task which we will discuss below. In the first variant, Grammatical Maze (or G-Maze), distractors are words of English, but they do not constitute a grammatical continuation of the sentence. In the Lexical Maze (or L-Maze) variant, distractors are non-English nonce words. The time it takes participants to select the correct word has been shown to be a good measure of incremental processing difficulty, both for targeted syntactic testing (Boyce, Futrell, & Levy, 2020) and for naturalistic reading (Boyce & Levy, 2020). Crucially, processing times for maze data have been shown to appear on the target word itself, and not in downstream spillover regions.

Between G-Maze and L-Maze, G-Maze has been shown to produce higher sensitivity (Boyce & Levy, 2020). However it poses a few interrelated methodological issues: because at each index, one distractor must constitute a possible next-word of the sentence, it cannot be deployed to test ungrammatical sentences (or ungrammatical sentence variants, such as, say featural mismatches), and may not be the best choice for sentences with highly unexpected continuations. In the unexpected or ungrammatical region, participants would be presented with the 'true' unexpected continuation alongside the 'false' distractor (ungrammatical), and would not know which continuation to select. Additionally, recent implementations of G-Maze (Boyce et al., 2020) use a neural-network based language model to automatically generate a low-probability distractor. Because the language model generates distractors based on their conditional probability, there are certain cases in which the G-Maze distractor may be a grammatical, but very unlikely continuation. These grammatical continuations could pose an additional problem for participants especially when the correct continuation is also unlikely. For example, following the continuation `The tenant that...` the two possible continuations generated from from the model in (Boyce & Levy, 2020) are `despised` and `wrestled` (for the G-Maze). While one of these contin-

[1] See, especially Table 1 of (Boyce & Levy, 2020) which shows a significant effect of current-word surprisal on reading time, but no significant effect for previous words on reading-time in the Maze task.

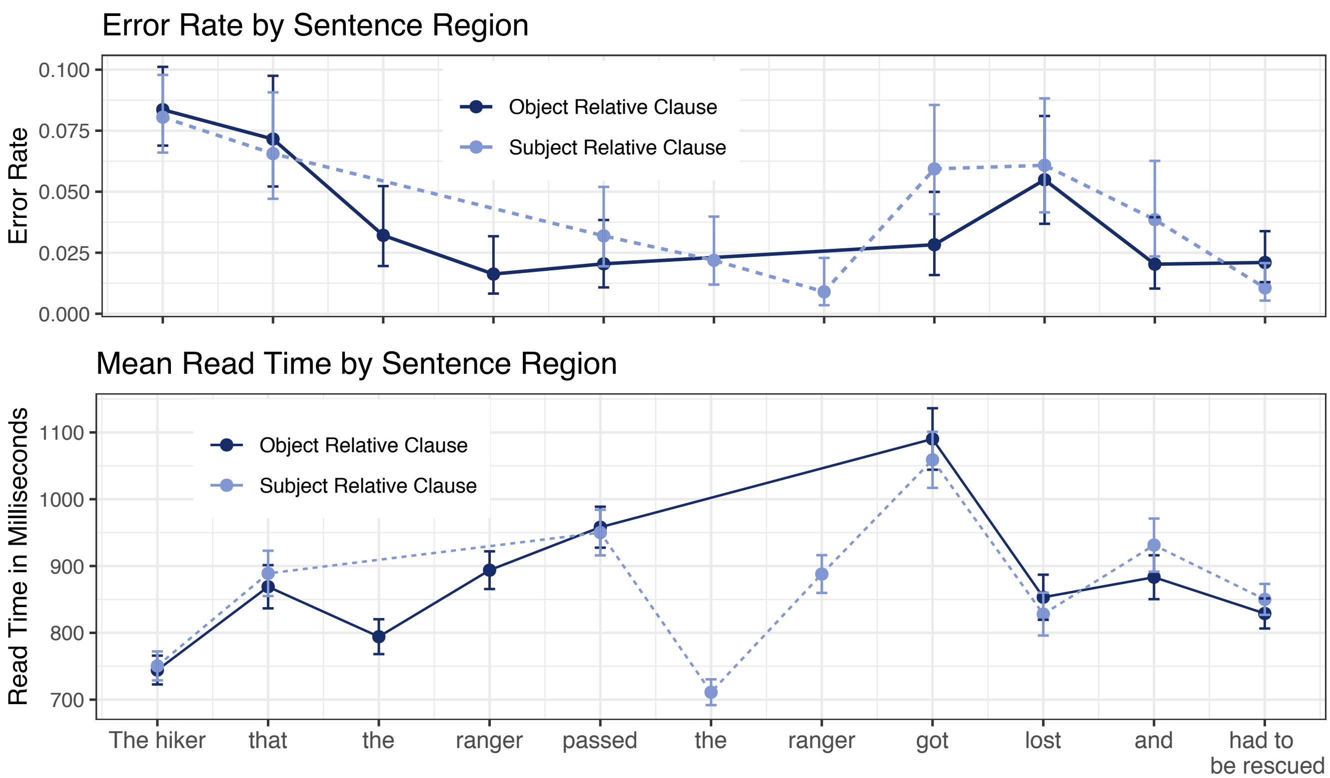


Figure 2: Experiment 1. Top Panel: Mean error rate by sentence region. Error bars are 95% binomial confidence intervals. Bottom Panel: Mean reading time by sentence region. Error bars are 95% confidence intervals.

uations is more unlikely than the other, neither is truly ungrammatical and thus poses a problem to the normal G-Maze setup.[2]

To solve these problems, we introduce Interpolated Maze (I-Maze). In I-Maze, we interweave L-Maze and G-Maze distractors. In critical sentence regions where one of the conditions is ungrammatical or highly unlikely we use only L-Maze distractors. Participants are instructed to select grammatical continuations over ungrammatical ones, and true English words over nonce distractors, thus making their choice unambiguous at each sentence index. In order to spread L-Maze distractors evenly through the sentence we randomly sample ∼15% of all words in non-critical regions and render them as L-Maze choices. We render these additional L-Maze distractors by changing 2 consecutive distractors, to match the 2-3 word sequence of L-Maze distractors participants encounter in critical regions. Because I-Maze makes fewer assumptions about the grammatical status of the target sentence, it is a fully generalizeable variant of the Maze task.

For the experiments reported in this paper, G-Maze distractors were produced with the scripts provided in Boyce and Levy (2020). Nonce words were generated with Wuggy (Keuleers & Brysbaert, 2010). For all of our experiments participants were recruited on Amazon Mechanical Turk and experiments were hosted on Ibex Farm (Drummond, 2013). 50 participants were recruited to take part in each experiment. Subjects were not allowed to participate in Experiment 2 if they had completed Experiment 1. Each participant was required to be accessing the internet using an IP address located within the United States, have a lifetime M-Turk approval rating of above 95%, and have graduated from a US high school. The results of participants who indicated that they were not native English speakers were filtered out. Participants were compensated $1.00 for their participation in the study.

Experiment 1

This experiment paired subject relative clauses with object relative clauses. The pairing allowed for a direct word-by-word comparison between the two types of sentences, which was intended to highlight the precise location of slowdowns in each relative clause type.

Design and Materials

The experiment followed the same design as Experiment 1 from (Traxler et al., 2002), with two conditions, given in (2).

(2) a. The child that chased the babysitter squealed with delight at the game. [SUBJECT RELATIVE CLAUSE]
b. The child that the babysitter chased squealed with delight at the game. [OBJECT RELATIVE CLAUSE]

The *Subject Relative Clause* condition consisted of a sentence which included a subject relative clause attached to an NP that served as the subject of the sentence. Likewise, the *Object Relative Clause* condition consisted of a sentence with an object relative clause also attached to a subject NP. The

[2] Other examples include `The plumber that...[ helped / claims ]` and `The golfer that...[ liked / opens ]`

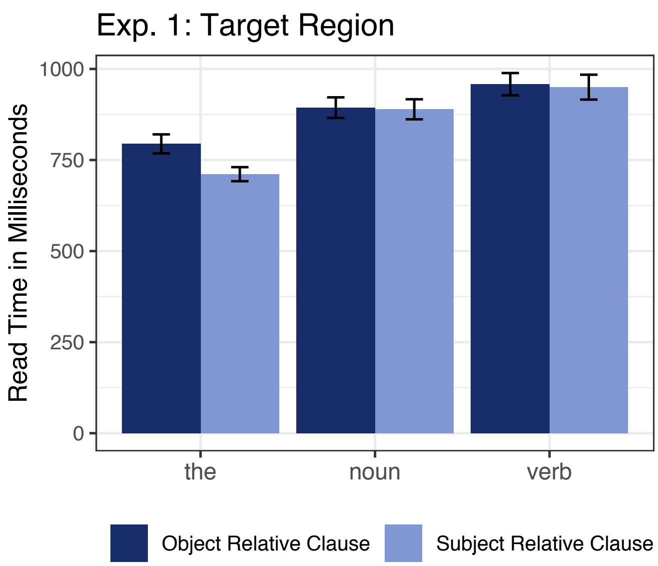


Figure 3: Results for the target region in Experiment 1 showing the contrast at *the*, and not in the other regions.

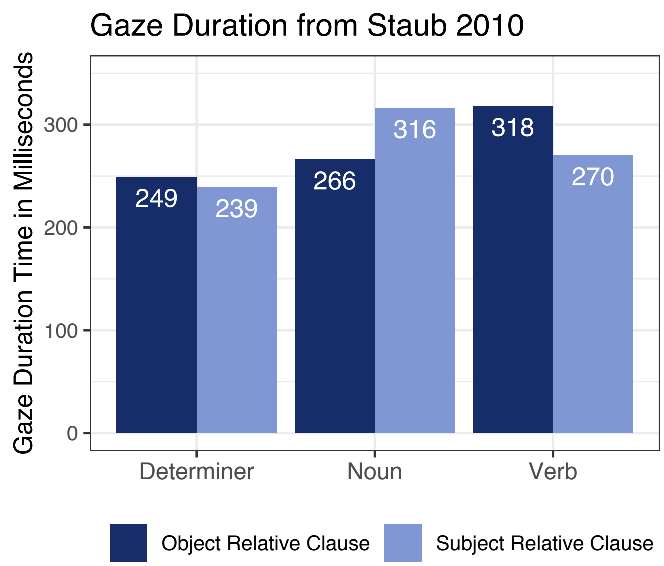


Figure 4: Gaze duration data as reported in (Staub, 2010), showing the differences in the noun and verb regions.

only difference between these two conditions was the relative clauses within them. In this experiment the critical regions—which were rendered as L-Maze decisions to participants—were the relative clause and the matrix verb of each sentence, which are underlined in (2). Target sentences were the same ones used in Experiment 1 of Traxler et al. (2002), and consisted of 24 target items with 20 filler items. If the surprisal theory of incremental parsing is an adequate explanation for the differences observed, we would expect processing slowdowns on the determiner of the Object RC compared to the Subject RC. Large contrasts in processing difficulty in other regions would be inconsistent with surprisal theory.

Results and Discussion

If participants select the incorrect continuation, they are automatically taken to the next sentence in the experiment and do not see the rest of the words in the sentence. Therefore, the proportion of words that subjects fail to reach in the Maze task due to previous incorrect answers gives a good sense of how difficult the task is. We found that participants reached 74.31% of items, averaging across participants. This indicates that they understood the task and were engaged. We filtered out 205 responses (0.9% of total responses) for being more than 2 standard deviations away from the mean read-time.

The top panel of Figure 2 shows the by-word error rate for this experiment, with error bars indicating 95% binomial confidence intervals. The error rate is the proportion of participants who make a mistake in a given region. Boyce et al. (2020) found that, for G-Maze and L-Maze tasks, error rates tend to be under 10%, but can be higher at the beginning of sentences. We find that for all regions, error rates are under 10%, providing further evidence that participants understood and were engaged in the task.

The bottom panel of Figure 2 shows mean reading-time by region, with the regions on the x-axis and the average read-time for all words in the region on the y-axis. As sentences had varying numbers of words in their post-RC continuation regions, the last points indicate mean reading times for words with sentence index 12 or higher. Error bars are 95% confidence intervals. We test the difference in the total time it takes to read the relative clause by fitting a linear mixed effects model with RC-type as the sole predictor variable and by-participant and by-item random slopes. We discard data from participants who make a wrong choice somewhere in the relative clause. We find that, for the whole critical region, condition was a near-significant predictor of total read-time ($p = 0.059$), where the ORC condition was associated with higher reading times. Visually, it seems that the locus of this reading-time slowdown is the disambiguating determiner.

Figure 3 shows the reading-time by condition with matched critical-region words. This figure shows more clearly that the locus of difficulty is the disambiguating determiner; both the noun and verb varied very little in processing time across the two conditions. To test whether these differences were significant, we fit a mixed effects linear regression model for the data in each paired RC region. We added experimental conditions as predictor variables, and included by-item and by-participant random slopes. We find that the condition is a significant predictor of the reading time for the determiner (`the`) ($e = 70ms$, $p < 0.001$), where an ORC condition predicted a higher reading time. We do not find a significant difference in reading time across the two conditions for the noun ($p = 0.4$) nor for the verb ($p = 0.7$).

To contrast our results with previous findings, in Figure 4 we include a graph rendered from the data presented in Staub (2010), for the same experimental conditions. The x-axis is

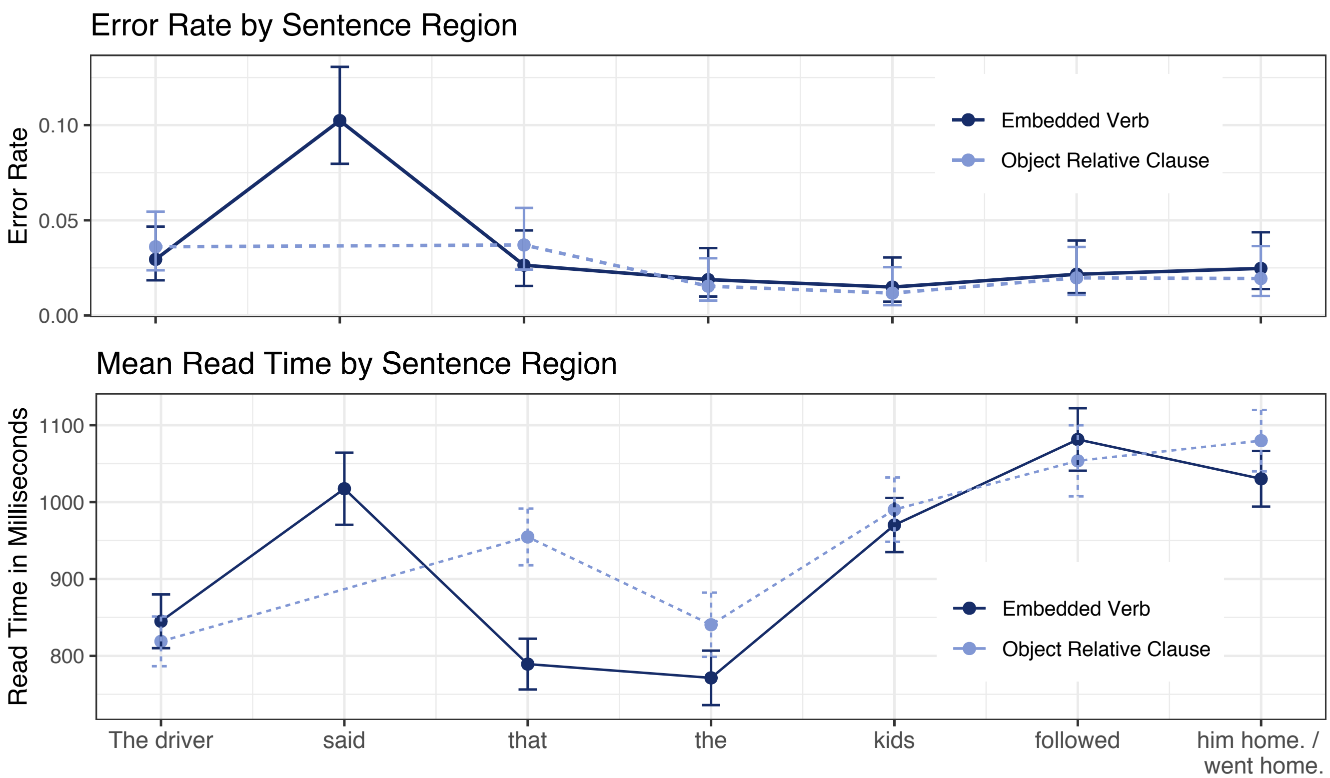


Figure 5: Experiment 2. Top Panel: Mean error rate by sentence region. Error bars are 95% binomial confidence intervals. Bottom Panel: Mean reading time by sentence region. Error bars are 95% confidence intervals.

again sentence region, but the y-axis is now mean gaze duration time across items in an eye-tracking experiment, which produces smaller processing times than the Maze task. The crucial contrast between these results and ours are the differences in gaze duration between conditions on the verb and on the noun. Each of these regions is read with about a ∼50ms difference between the conditions, whereas the determiner is read at practically the same pace in both ORCs and SRCs. Our results are opposite, with large differences for reading times of the determiner between conditions, and almost no difference at the noun and the verb.

The results from our experiment indicate that the locus of processing slowdown in Object Relative Clauses is on the disambiguating determiner. This is consistent with the results presented in Forster et al. (2009), who find similar effects using a purely G-Maze implementation of this experiment. Furthermore, these results are consistent with a surprisal theory approach to incremental processing, which predicts that participants should experience slowdown as soon as they recognize that the current input does not match their expectations.

One possible explanation for the difference between Maze and eye-tracking experiments is spillover effects. In Staub (2010), regions that were read more slowly in the ORC condition immediately follow the disambiguating determiner. Slowdown could be due to difficulties at the determiners that manifest in downstream regions. Supporting this analysis is the fact that the determiner was shown to be skipped frequently in both conditions, and that a large number of regressive saccades occurred at the determiner (and the noun) in the ORC condition. These regressive saccades may indicate a processing failure and suggest that the locus of the processing difficulty lies at either the determiner or the noun, even though longer reading times are observed only on the verb. As noted by Staub (2010), longer verb reading times could be consistent with a theory which assigns processing slowdown to difficult memory retrieval at the verb. However, the results from our experiment fail to identify any processing difficulty at the verb, suggesting that these observed slowdowns are more likely the result of spillover effects. However, we note that though spillover effects exist in eye-tracking and self-paced reading, they can be mitigated through design decisions and post-processing such as deconvolutional techniques (Shain & Schuler, 2018).

Experiment 2

In the previous experiment, we found that SRCs were processed more quickly than ORCs, and that the locus for the difference was on the relative clause's determiner. However, because the previous experiment tested variation in processing within relative clause types, it does not tell us whether participants were experiencing difficulty when processing these structures in general. In order to test the reliability of the results, and to provide a non-relative clause baseline, in this experiment we compare processing times of ORCs to lexically identical material in syntactic environments that should cause no processing slowdown.

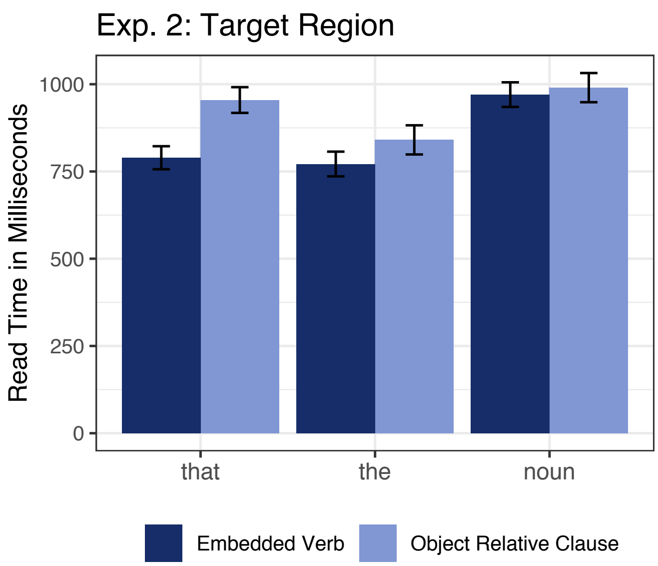


Figure 6: Results for the target region in Experiment 2 showing the largest contrasts at *that*.

Design and Materials

The experiment followed the same design as Experiment 2 from Staub (2010) and consisted of two conditions, outlined in Example (3). The *Object Relative Clause* condition consisted of an object relative clause attached to a subject. The *Embedded Clause* condition consisted of a clausal-complement taking the verb with the same subject. Because the complementizer *that* can serve to introduce both an embedded clause and a RC, there was direct lexical overlap between the beginning of the RC and the beginning of the embedded clause, which is in bold in the example below.

(3) a. The bus driver said **that** the kids followed him all the way home. [EMBEDDED CLAUSE]

b. The bus driver **that** the kids followed wondered about the location of the hotel. [OBJECT RELATIVE CLAUSE]

For this experiment, the critical regions which were rendered as L-Maze choices to participants were the determiner, noun and verb, which are underlined in (3). If participants are experiencing processing difficulty for relative clauses, then we expect the complementizer to be read more slowly in the *Object Relative Clause* condition than in the *Embedded Clause* condition. Furthermore, we expect the determiner, which unambiguously introduces the RC as an *object* RC to be read more slowly in the *Object Relative Clause* condition as well. Materials were the same ones used in Experiment 2 of Staub (2010), with 24 target items with 20 filler items.

Results and Discussion

We found that participants reached 78.76% of the words in the experiment (averaging across all participants). This is similar to Experiment 1, and we take it to indicate participants understood the task and were engaged. We filtered out 45 words (0.18% of the data) for being more than two standard deviations away from the mean reading time.

The top panel of Figure 5 shows the mean error rate per sentence region. (Error rate is the proportion of participants who make a wrong choice in the region, which sends them to the next sentence in the trial.) We find a low error rate in all of our regions, with slightly higher error rates for the embedding verb (e.g. *said*) in the *Embedded* sentences. For all regions that are paired, error rates are not significantly different between conditions.

The bottom panel of Figure 5 shows the mean time it took participants to read words in each region in milliseconds. Error bars are 95% confidence intervals. Figure 6 offers a closer look at the target regions, pairing the words that correspond across the two conditions. To test for differences in reading times between conditions, we fit a linear mixed effects model with the condition (ORC vs. Embedding) as the sole predictor and with by-participant and by-item random slopes. Summing over the entire critical region, we find that ORC sentences were read more slowly than Embedding sentences ($p < 0.001$). Turning to the individual words, we find that condition is a significant predictor of reading times for the complementizer (`that`) ($e = 78ms$, $p < 0.001$) and approached significance for the determiner (`the`) ($e = 26ms$, $p = 0.07$) where the ORC condition was associated with higher reading times for both regions. The differences are not significant in the other regions.

In Experiment 1, we cannot be confident that order effects, or interference effects at the verb are not causing the observed differences in reading time. In this experiment, we match the conditions precisely in terms of material and (except for one word) position in the sentence. The observed higher reading times in the ORC condition, therefore can help eliminate these potential confounds and provides additional evidence that ORC determiners are the main locus of processing difficulty when reading these structures.

General Discussion

The result of the two experiments supports a surprisal based theory to explain the difference in processing between ORCs and SRCs. Experiment 1 shows that ORC sentences were processed slower than SRC sentences and that the ORC determiner was the locus of the slowdown. In Experiment 2 we found that the only regions to manifest slowdown between conditions were the determiner and the complementizer. These data are consistent with an interpretation that slowdowns are due to readers encountering unexpected material, rather than difficult word ordering or other interference effects. One major difference between the Maze task and previously used eye-tracking experimental design is that the Maze task does not allow for re-analysis of words, as participants are not able to look back at earlier parts of the sentence. Instead, they must rely on memory traces for re-analysis. This means processing failures are not observed through regressive saccades as in eye-tracking but instead through inflated reading times of unexpected words.

One high-level worry with the Maze task is that it does not resemble behaviors experimental subjects engage in on a daily basis, and is therefore more "artificial" (Forster et al., 2009) than eye-tracking, and may reinforce top-down processing strategies. Although we do observe differences between Maze and eye-tracking results in our studies, we believe that overall comparison between the two methods indicate these worries are unfounded. First, both Maze and eye-tacking produce the same shape between processing time and by-word conditional probability (log-linear) indicating that these two methods are capturing the same behavior *in aggregate* (Boyce & Levy, 2020). Second, in other controlled experiments, Maze and eye-tracking have produced reading time patterns that match qualitatively. Therefore, while the artificiality of the Maze task is a potential concern, the evidence suggests that it does not pose significant challenges to it as a valid methodological tool.

Comparing the results of Experiment 1 to the results obtained by Forster et al. (2009), we find similar effect sizes. Both our results and theirs show marginal ($< 15ms$) differences between conditions on the relative clause noun and verb. They find a $137ms$ difference at the determiner, whereas we find an $83ms$ difference. These slight differences are sensible in light of the data presented in Boyce et al. (2020), who find that lab-based G-Maze experiments of the type deployed by Forster et al. tend to have the highest sensitivity. While the data presented here suggests that web-based I-Maze is slightly less sensitive than lab-based G-Maze, they show that this experimental technique is powerful enough to capture crucial contrasts and shed light on fine-grained processing of syntactic structures.

Acknowledgements

RPL would like to acknowledge NSF grant BCS-1551866, a Google Faculty Research Award, and funding from the MIT–IBM AI Lab.

References

Boyce, V., Futrell, R., & Levy, R. P. (2020). Maze made easy: Better and easier measurement of incremental processing difficulty. *Journal of Memory and Language*, *111*, 104082.

Boyce, V., & Levy, R. (2020). A-maze of natural stories: Texts are comprehensible using the maze task. *Proceedings of the Architectures and Mechanisms for Languages Processing Conference*.

Drummond, A. (2013). Ibex farm. *Online server: http://spellout. net/ibexfarm*.

Ehrlich, K., & Rayner, K. (1983). Pronoun assignment and semantic integration during reading: Eye movements and immediacy of processing. *Journal of verbal learning and verbal behavior*, *22*(1), 75–87.

Forster, K. I., Guerrera, C., & Elliot, L. (2009). The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, *41*(1), 163–171.

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

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input for sentenial complexity. *Cognitive science*, *29*(2), 261–290.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*(3), 627–633.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, *30*(5), 580–602.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Levy, R., Fedorenko, E., Breen, M., & Gibson, E. (2012). The processing of extraposed structures in english. *Cognition*, *122*(1), 12–36.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. , *4*, 226.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of memory and language*, *47*(1), 50–68.

Reali, F., & Christiansen, M. H. (2007). Processing of relative clauses is made easier by frequency of occurrence. , *57*(1), 1–23.

Shain, C., & Schuler, W. (2018). Deconvolutional time series regression: A technique for modeling temporally diffuse effects. In *Proceedings of the 2018 conference on empirical methods in natural language processing.*

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*(3), 379–423.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Staub, A. (2010). Eye movements and processing difficulty in object relative clauses. *Cognition*, *116*(1), 71–86.

Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of memory and language*, *47*(1), 69–90.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.