Caitlin Brady
Experimental Methods

Underspecification of syntactic ambiguities

1. Introduction

Any time we are exposed to language, we must interpret it in order to arrive at a meaning for what we've read or heard. However, it can be difficult to determine exactly what that process involves. This is particularly true of ambiguous sentences, which have commonly been assumed to be more difficult to process than unambiguous sentences, due to the fact that they have multiple possible interpretations in competition with each other. However, if we take reading time to be indicative of processing difficulty, as many researchers have done, there is evidence that ambiguous sentences may actually be less taxing to process, as studies have indicated that ambiguous sentences often have shorter overall reading times than their disambiguated counterparts.

There are two primary accounts for how this might be explained. Under the obligatory attachment account, exemplified by models like the race model, it is assumed that readers must always assign a specific interpretation to each sentence they encounter, and interpretation happens on-line as the sentence is encountered. Ambiguous sentences are processed more quickly because there is never an instance where whatever initial assumption was made regarding the interpretation early in the sentence is disrupted by contradicting later information. Unambiguous sentences may require reanalysis due to later information, resulting in ambiguous sentences having a global advantage in processing time when compared on average.

Another possibility is that readers do not need to arrive at a specific interpretation for each sentence (underspecification). This account posits that for a disambiguated sentence, the reader must arrive at a specific interpretation of either N1 or N2 being described by a relative clause (or some other type of ambiguous structure), while for an ambiguous sentence, the reader only needs to arrive at the fact that it applies to *someone* in the sentence.

This experiment is a replication of a study by Swet, et.al. (2008) that attempted to determine which approach is better able to capture patterns found when examining reading times of both ambiguous and unambiguous sentences. The goal is to determine whether the same patterns are found among a different group of participants to determine whether the patterns found in the prior study are generalizable. The experiment is almost the same as the previous study, but I did alter some of the experiment data to improve the clarity of the disambiguation for some experiment items. Because the disambiguation in the experiment items relies on gender agreement, some items from the original experiment were changed to be more clearly gendered, in order to avoid relying on participants to hold certain stereotypes or encounter unintended ambiguity.

2. Experiment

This experiment is designed to examine the effects of question type and sentence disambiguation type on reading times via a self-paced reading test. The goal is to compare the results to the predictions made by the two primary competing hypotheses, obligatory attachment or underspecification. Both approaches make different predictions about the patterns that will be seen in the reading times, so the results should be able to indicate which approach makes the more accurate predictions.

2.1 Methods

2.1.1 Design & Materials

This experiment consisted of a number of self-paced reading trials followed by questions for the participants to answer. The experiment items followed a 3x3 design and consisted of sentences containing two noun phrases with overt gender and a relative clause which contained a reflexive pronoun matching one (or both) of the two noun phrases in gender features. The basic template of the

sentences was "[The N1] of the [N2] [who [embedded verb] [reflexive pronoun] [modifier]] [matrix verb] [rest of matrix sentence]." This basic sentence template was manipulated to make sentences where the reflexive pronoun could disambiguate the attachment of the relative clause by matching the first or second noun, or could leave it ambiguous if both nouns shared the same gender feature. The modifier following the reflexive pronoun was also important due to the fact that effects on reading time often bleed into following parts of the sentence; by including a modifier after the reflexive made it possible to detect this before the introduction of the matrix verb. This sentence disambiguation manipulation was a within participant variable, as each participant encountered all three types.

In addition to the manipulation of sentence disambiguation via the genders of the nouns and reflexive pronoun, the experiment items were followed by comprehension questions falling into three different categories. These were between participant variables, as each participant was put into a group that only encountered questions of a single type. These question types were either related to the interpretation of what the relative clause was modifying (category Rel.), superficial questions that only tested basic sentence comprehension unrelated to the relative clause (category Sup.) or superficial questions that were only presented intermittently (category Int.) but were otherwise identical to the superficial category questions. Participants in the intermittent question category encountered questions 1/12 of the time, distributed randomly among the items they encountered.

A full experiment item with all conditions is as follows:

1. The nephew of the baroness who prepared himself very thoroughly came from the south.
2. The nephew of the baroness who prepared herself very thoroughly came from the south.
3. The niece of the baroness who prepared herself very thoroughly came from the south.

Sentence 1 – Disambiguated to N1, Sentence 2—disambiguated to N2, Sentence 3—ambiguous

1. Was the nephew thoroughly prepared?
2. Did someone come from the south?

Question 1—Related to relative clause attachment, Question 2—superficial question

In addition to the 36 experimental items, participants also encountered 36 filler items. These varied somewhat in form but were roughly the same length as the experimental items and also contained two noun phrases, though they did not have an ambiguous relative clause to be disambiguated by a reflexive pronoun. Participants all encountered the same filler items, which were accompanied by different questions according to their question type group. Those who were in the Rel. group had comprehension questions that required a somewhat closer reading of the sentence, those who were in the Sup. and Int. groups received similarly simple questions relating to basic comprehension. An example of a filler item is as follows:

Sentence: The nephew of the princess sneezed during the ceremony and was terribly embarrassed.

Rel. question: Did the nephew sneeze?

Sup. Question: Was somebody proud?

Questions were evenly divided between yes and no correct answers (with any ambiguous rel. questions being correct with either answer). Experiment and filler items were randomized and interspersed so participants never encountered two experiment items consecutively.

2.1.2 Procedure

The experiment was set up on PCIbex and distributed to participants via an email link. All participants performed the experiment online on a computer, a fact that was enforced by the fact that PCIbex's self-paced reading test is incompatible with mobile devices due to its reliance on having a space bar to proceed to the next word. This made it somewhat less convenient for people to complete the experiment, so several participants completed the experiment on the same device one after another.

Participants were presented with instructions describing the basic flow of the experiment consisting of reading a sentence and then answering a question about it. They were told that they would

see only one word at a time and could proceed to the next word in the sentence by pressing the space bar on their keyboard. They were asked to read the sentences at a natural pace and to avoid rushing. Before the experiment began, they were presented with five practice sentences in order to familiarize themselves with the procedure. These sentences were the same for all participants, similar to the filler items, and were followed by questions of a similar type to those that accompanied the filler data for their respective question type groups, just as was done in the original study. All participants were also informed in the instructions that it was possible they would not encounter questions after every sentence. This was included to ensure that those in the intermittent group, who only rarely encountered questions, would not be confused at the questions' absence.

2.1.3 Participants

Participants were my friends and neighbors, mostly located in Ithaca, NY. The participants were primarily in their 20s and 30s. There were 15 participants total and all participants were native speakers of English from the United States of America. All of the participants were female, though this information was not recorded in the experiment data collection itself. They were recruited by being asked if they were interested in helping with the experiment and were not compensated for their participation with anything other than gratitude. The only discard criterion was that participants needed to be native English speakers. Since none of the participants were non-native speakers, no participants were discarded from the data pool.

2.2 Predictions

Because this is a replication of an existing study, the main prediction is that the results will match those of the previous study. Specifically, it is predicted that the underspecification approach will better fit the data. This should be evident in how the question type interacts with the reading times, as obligatory attachment predicts that ambiguous sentences should always be read fastest and question type should not affect the reading time patterns, while the underspecification approach predicts that there will be a difference between the relative clause question group, which will require readers to arrive at a definitive attachment in order to answer the question; in particular, it is expected that the global advantage for ambiguous sentence reading times predicted by obligatory attachment will not be found in the relative clause question group.

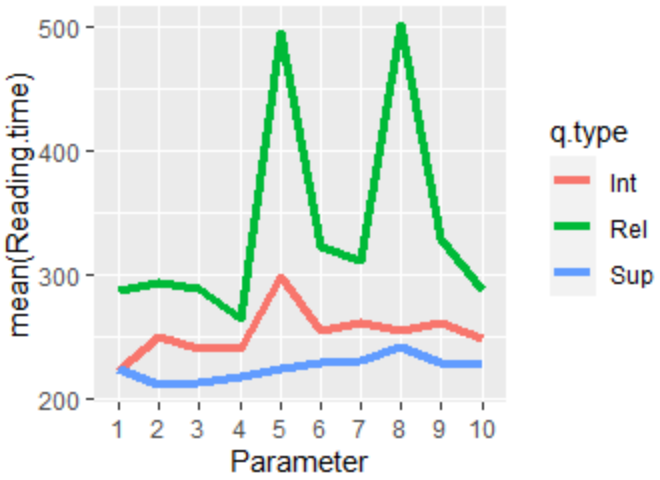
The study being replicated found that the reading times at the point of disambiguation (the reflexive pronoun) showed main effects of both question type and sentence disambiguation type, but no interaction between the two, while the postdisambiguation region (the modifier immediately following the reflexive pronoun) showed both main effects for question type and sentence disambiguation type and a significant interaction between the two, suggesting that underspecification is the better explanation and that readers answering questions about relative clause attachment spend longer recovering from the disambiguation toward N1 than toward N2 or ambiguous sentences.

The expected results of this experiment are that the same patterns of main effects and interactions will be seen in the data collected here.

2.3 Results

The experiment results were cleaned up into a more manageable form using the results cleaning script available on PCIbex's website. This allowed for easier processing of the data using R Studio.

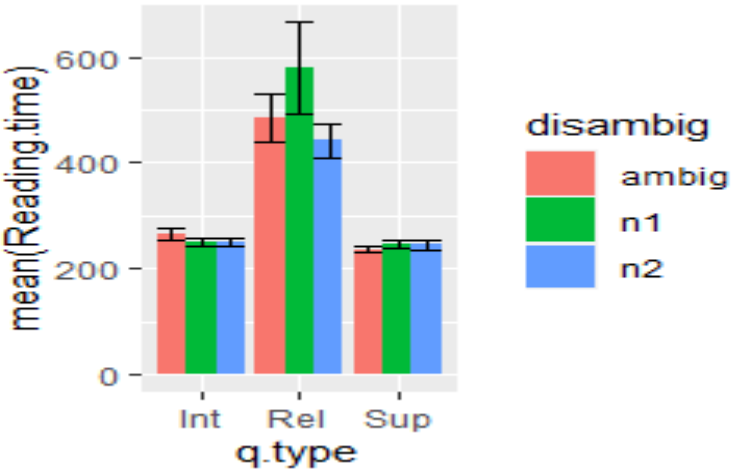
The word-by-word reading times for each of the three question type groups were averaged and plotted on the following chart to demonstrate the pace at which the readers progressed through the sentences they encountered. The y axis represents the mean reading times and the x axis represents relevant section of the sentence (some sections contained multiple words in some items so those were averaged down by category, as the previous study also did). The numbers on the x axis are as follows: 1 = The, 2 = N1, 3 = of, 4 = the, 5 = N2, 6 = who, 7 = verb, 8 = Reflexive pronoun, 9 = modifier, 10 = matrix verb. Of these, categories 7, 9, and 10 had variable numbers of words. Elements following the matrix verb were not charted on this graph.



From this, we can see that the different question types followed slightly different patterns. Readers in the relative clause question group had the longest reading times, and they slowed down the most at the second noun and the reflexive pronoun. The other groups had smaller changes overall, and their patterns differed somewhat, with the intermittent group slowing more at N2 and the superficial group slowing more at the reflexive, though neither had such extreme changes of pace. This pattern differed from that of the previous study, which found that the three sentence types followed much more similar patterns and rates, though relative clause question type was still the slowest overall. This is very likely to be related to the comparatively smaller sample size of my data, as individual variation in reading speed and rate adjustment patterns is far more noticeable with only a few participants per group. This does, however, support the idea that the reflexive pronoun is a relevant part of the sentence to examine.

In line with the original study, both the reflexive pronoun reading times and the post-reflexive modifier reading times were examined. The reflexive is important because it is the point of disambiguation and it was a noticeable point of change in two of the three charted sentence types above. The post-disambugation region is also important because it has been noted that reading pace changes may bleed over into following elements and this was an area that the original studies authors found supported their conclusion that underspecification was the better explanatory account.

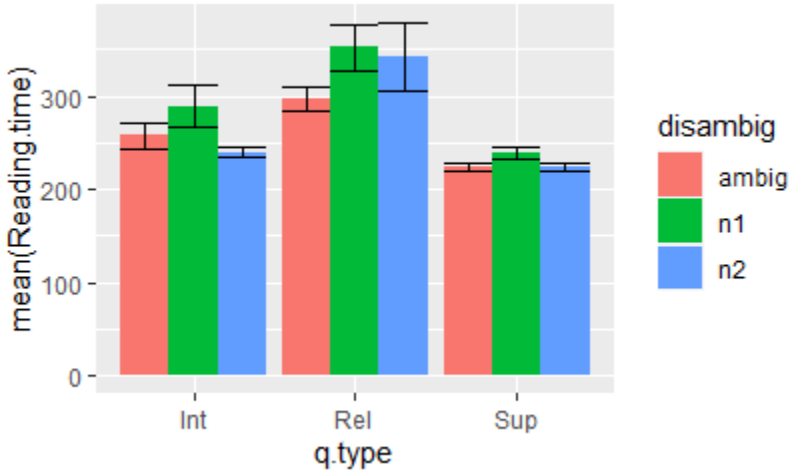
The reading times for the reflexive pronoun as a function of question and sentence disambiguation type are displayed on the following chart:



The reading times for ambiguous sentences are not consistently lower than the other types, and in fact, none of the types display a clear consistent pattern across the question type groups. Sentences disambiguated in favor of N1 are expected to take the longest, assuming the bias toward N2 attachment that the previous study found is true, but this was only clearly the case in the relative clause question group. The other groups showed little variation overall and not much difference between the different types of ambiguity.

A mixed ANOVA was performed to determine whether any significant effects were present. A significant main effect was found for question type ($F=17.22$, $p<0.001$), but not for sentence disambiguation type ($F=0.46$, $p>0.5$). This test also found no significant interaction between the two factors. This differed from the previous study which did find a significant effect of sentence type as well, though it also found no interaction between the two at the reflexive.

The same analysis was applied to the postdisambiguation modifier. The results can be seen in the following chart:



Here we can see that the results are much more similar across the different disambiguation types and the patterns are slightly more consistent. The relative clause question type once again has the longest reading times, but not by nearly as wide a margin. In each question type, sentences which disambiguated to N1 had the longest reading times.

A mixed ANOVA shows that for this area, there is again a significant effect of question type on reading time ($F=22.88$, $p<0.001$), as well as a small effect of disambiguation type ($F=2.63$, $p<0.1$). Once again, there did not appear to be a significant interaction between the two.

The data was also divided into early and later trials to see if the effects changed with more exposure to their group's question type, but the results were not different. If there was any adjustment to question type, it must have happened very quickly rather than gradually over the course of the trials.

Question answers for both experiment trials and fillers were examined to determine whether participants were rushing through without reading or favoring certain answers, but there were very few incorrect answers overall and no consistent patterns in favoring any answer type over another.

2.4 Discussion

The results of this experiment differed noticeably from the original study, in that this study did not find a significant effect of disambiguation type on reading time for the relative clause and only found a minor one for the postdisambiguation modifier. It also did not find any interaction between the two factors, while the original study found an interaction in the postdisambiguation region. This study also didn't find any significant patterns in question answers or errors, while the original study did find some minor patterns in both.

As there did not appear to be evidence of answering randomly, it would appear that the participants were reading the sentences properly rather than rushing through. As such, it is difficult to determine why the results found here are so different from those in the previous study. However, one observation that may be important is that the average per-word reading times overall and in the two highlighted areas were much faster for my participants than those in the previous study. My results on the chart of average reading times per word range between 200 and 500ms, while those in the previous study ranged from 300 to 600. This faster reading pace, combined with the small sample size may possibly be obscuring effects simply because at a certain point the pace of the reading becomes limited by the rate at which the participant can press the space bar to proceed, rather than the rate the words are being read. This could be supported by the fact that the slowest-read category, the relative clause question category, was the only one to match the previous study's pattern in the reflexive pronoun reading times.

As the results stand, this experiment did not find enough evidence to determine which hypothesis was correct as the results could be consistent with either.

3. Conclusion

Determining whether processing of ambiguous structures requires full analysis of the ambiguous structure or whether it can be left unspecified is an important goal if we are to understand how human language is processed. The original study indicated that there is evidence to support the underspecification account. However, this attempted replication failed to find the same evidence. This may have been the result of the small sample size skewing toward fast readers who read at a pace similar or faster than their button pressing pace. Without a larger and more diverse sample of participants, it is difficult to draw firm conclusions. It is clear that the type of questions asked affect the rate at which the readers proceed through the sentence, but that alone isn't enough to conclude whether underspecification is in play or not.

While this experiment did not arrive at a clear conclusion, it does demonstrate an important finding about this type of test. While the importance of larger sample sizes is widely considered in experiments, the demographics and any language-related skills they may possess is also important to consider. The participants of this study were faster readers overall than those in the previous study, and this fact may have led to problems with the experiment methodology in that the self-paced reading test relies on the assumption that the reader will always be reading the words slowly enough to keep pace with their button pressing. If the button pressing rate and the reading rate converge, it could potentially obscure the data.

4. References

Swets, Benjamin, Timothy Desmet, Charles Clifton, Jr., and Fernanda Ferreira (2018). Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition, 36*(1), 201-216.