The G-Maze task in ambiguous pronoun resolution
(Bachelorproject)

Stijn Tromp, s1903500, s.tromp.1@ai.rug.nl,
Jennifer Spenader *

December 12, 2014

Abstract

This study uses the G-Maze task to study ambiguous pronoun processing, using materials from an earlier study by Taylor (2013) We expected the G-Maze task to show the effects of the Constraint-Based model without relying on such a comprehension question. The forced integration of the G-Maze task proved to not be enough to show these effects, a result similar to those found by Taylor when a comprehension question was ommited. Pronoun interpretation is a deep semantic process that requires active stimulation to see effects.

1 Introduction

When reading certain sentences, some pronouns may be ambiguous at the moment they are presented. Take for example the following sentence in Dutch, taken from the materials of Ryan Taylor (Taylor 2013):

(1) a. *Richard zag Roland de appelboom snoeien en kwam helpen.*
(Richard saw that Roland was trimming the apple tree, and came to help)
b. *Roland was blij dat Richard de ladder vast wilde houden.*
(Roland was happy that Richard offered to hold the ladder.)
c. *Hij verbaasde zich over de zwaarte van de ladder die rechtgehouden moest worden.*
(He was amazed at the weight of the ladder that had to be held straight.)

The example above (1) contains two characters, *Roland* and *Richard*, who alternate between the role of participant and object. The last sentence contains two words where processing difficulties may arise. At the start of the sentence, the pronoun *Hij* (he) is encountered. This pronoun is initially ambiguous, as there is not enough information present to determine whether it refers to *Roland* or *Richard*. Followed closely, the disambiguation word *ladder* is encountered, disambiguating the pronoun as referring to *Richard*.

Two factors that play a role in pronoun interpretation, are those of order-of-mention and recency. Gernsbacher et al. (Gernsbacher, Hargreaves, and Beeman 1989) studied two phenomena that can influence reaction times in experiments looking at pronoun interpretation. The First-Mention advantage states that the first participant mentioned in the story will form the foundation, and will be more readily available in memory. In the example (1), when the pronoun is read, the First-Mention advantage states that there is a tendency to disambiguate to Richard, since he is first mentioned in the story, and thus more readilly available from memory. The recent clause advantage states that words that take the role of participant in the most recent clause are more readilly available than those in earlier clauses. In the example (1), the most recent clause contains Roland as the participant, which makes it more readilly available. According to the research of Gernsbacher et al. (1989), this counterbalance makes preference to both Richard and Roland

*University of Groningen, Department of Artificial Intelligence

equal at the pronoun.

In earlier studies, two models have been proposed to describe the mechanics underlying the processing of pronoun ambiguity. In the Constraint-Based model (Trueswell and Tanenhaus 1994), all possible information available is considered in parallel, resolving the pronoun to the referent most strongly preferred in light of the information. In the above example, preference to both *Roland* and *Richard* is equal at the pronoun, and the Constraint-based model therefore predicts an increased reading time at the pronoun. If new information for disambiguating the pronoun is found later in the sentence, the pronoun is reconsidered, increasing processing time and resulting in slower reading.

In the Unrestricted Race model (Pickering, Traxler, and Crocker 2000), attempts are made to account for evidence where sentences with globally ambiguous pronouns (i.e., the pronoun is never resolved) are actually faster to process than sentences with unambiguous pronouns. In this model, at the point that the pronoun is encountered, a choice is made for one interpretation. Only when later evidence proves that the wrong choice was made, the interpretation is reconsidered. In sentences with ambiguous pronouns, compared to sentences with unambigous pronouns, both models predict longer reading times at the disambiguation word. In these sentences, the Constraint-Based predicts *longer* reading times at the pronoun, where the Unrestricted-Race model predicts *shorter* reading times at the pronoun.

To compare these two models, some kind of method is required to record the reading times at the words of interest. Ryan Taylor (2013) compared these two models using a Self-Paced Reading task. In his study, he found evidence in support of the Constraint-Based model underlying ambiguous pronoun resolution. As proposed by the model, reading times at the pronoun and the disambiguation words increased significantly. However, these effects were only found when a comprehension question was presented about what character the pronoun refered to.

To explain why effects were found in light of a comprehension question, and not without, the good-enough hypothesis was proposed. This hypothesis states that sentences aren't always processed in full, but only as much as is neccesary to meet current task demands. Under the good enough hypothesis, participants avoid processing the pronoun, when they realise that this information is not needed to meet the task demands.

While the Self-Paced Reading task is a good method for tracking reaction times during reading, it does not guarantee words are fully read and intergrated at the moment a participant moves to the next word. In this study, we look at an alternative method to the Self-Paced Reading task. This method is called the Grammatical Maze task, or G-Maze for short, as proposed by Forster et al. (Forster, Guerrera, and Elliot 2009). The G-Maze task was designed to force complete integration of each word in the sentence, removing the need of a comprehension question to draw attention to the words of interest. The G-Maze task is designed as follows (Figure 1). The sentences of interest are presented to the participant one word at a time, with each word paired with a second natural distractor word, that doesn't form a grammatically correct continuation of the sentence. The participants are instructed to choose the word that correctly continues the sentence at every pairing. By presenting a choice at each word, instead of simply letting the participant decide "readiness" by pressing a button, the task forces the participant to pause reading untill the word is fully integrated in the context of the sentence. The choice also slows down processing, so there is no problem with synchronising keypress rate with processing time. Finally, Witzel et al. (Witzel (2012)) showed that that the forced integration serves as a way to reduce spillover effects, an effect where processing effects show up after the target word, adding the need to analyse the word after the target region to see the full effect.

Participants performed the G-Maze task applied to the materials used by Ryan Taylor (2013), in order to find out if the advantages proposed in the G-Maze theory work on ambiguous pronoun resolution. We found no effects on either the pronoun or the disambiguation word, suggesting that the forced integration introduced by the task was not enough to show the effects found in

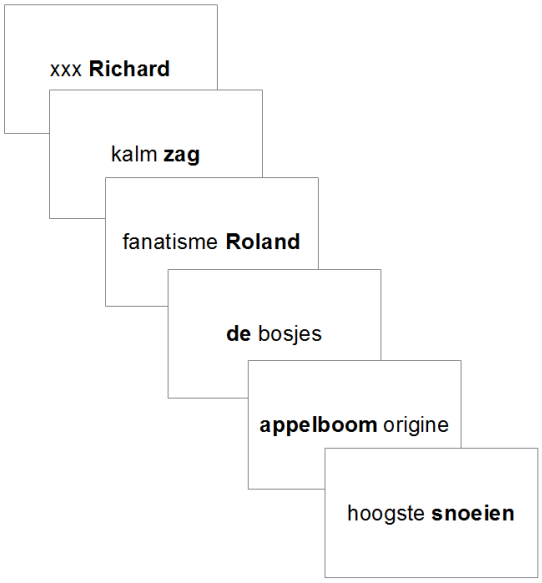

Figure 1: Example of frames in the G-maze task. The correct word is shown in bold

the self-paced reading task with comprehension question provided, as expected in the Constraint-Based model. This seems to be in line with the good-enough hypothesis, as outside the task, the information in the sentence is not used. No effect was found on words after either the pronoun or the disambiguation word, suggesting that the G-Maze task was successfull in localizing reading times, and avoiding spillover effects.

2 Method

2.1 Participants

16 native speakers of Dutch participated in the experiment. Mean:20, range 18-26, 8 female.

2.2 Design

The study was run on three macbooks, running at 50 Hz, projected on a large screen. Input was handled by an external keyboard. The experiment was set up using the DMDX software for measuring reaction times to visual and auditory stimuli.
The study consisted of 64 stories, of which 32 were target stories. All were written in Dutch. The target stories were taken from the experiment in Taylor (2013), manipulated by adding an adverb at the start of the last sentence, to ensure that it didn't start with the pronoun. The target stories consisted of 3 sentences. In the first sentence, two characters were introduced, one taking the role of subject, the other the role of object. In the second sentence, the role of the characters were reversed. The last sentence contained two points of interest: First, a pronoun is encountered, placed to be initially ambiguous in refering to either of the two characters in half of the sentences, and disambiguated on gender in the other half. A few words later a disambiguation word is encountered, containing key information to disambiguate the pronoun to one of the two referents. An example of the manipulated sentences:

(2) a. *Wouter huurde Jeroen in om een portret te schilderen.*
(Wouter hired Jeroen to paint a portrait.)
b. *Jeroen wist van Wouter dat het voor het nieuwe kantoor was.* (Jeroen knew from Wouter that it was for the new office.)
c. *Daarna ging hij (pronoun) elke dag om te (poseren/schilderen)(disambiguation word) naar het prachtige atelier.*
(Then he went to the beautiful studio every day to (pose/paint.))

Two factors were tested: ambiguity and order of mention.Ambiguity has two levels. To manipulate ambiguity, the gender of the second character was varied. In unambiguous sentences, the second character was female, allowing the sentence to disambiguate based on the gender of the pronoun (*hij* (he) if the character is male, *zij* (she) if female.). In case of the example sentence, Jeroen would be replaced with Jolien. In ambiguous sentences, both characters were male. In total, there were 4 types of stories, and each participant was presented with 8 of each.
Order of mention likewise has two levels: the prefered antecedent was either the subject or the object of the first sentence. This was manipulated by changing the disambiguation word. For example in (2) above, by using the verb 'poseren' (to pose), the pronoun disambiguates to Jeroen (Jeroen is posing for the painting), while using the verb 'schilderen'

(to paint), the pronoun disambiguates to Wouter (Wouter is painting a portrait of Jeroen).
In the G-maze task, each story was presented word-for-word, each word being paired with a second word that didn't form a grammatically correct continuation of the sentence in the current context. To select these distractor words, a list of bigrams and a list of unique words was generated from the Algemeen Dagblad portion of the CLEF corpus, a corpus containing articles from Dutch newspapers. Distractor words were randomly picked from the unique word list, and selected when all of the following conditions were matched:

- The distractor word didn't form a bigram with the current word in the sentence.
- The distractor word didn't form a bigram with the word that came before the current word in the sentence.
- The distractor word didn't form a bigram with the distractor word matched one word before it.
- The distractor word was no more than 4 characters longer than the current word in the sentence.
- The distractor word was no more than 4 characters shorter than the current word in the sentence.

The words were first generated automatically, and manually checked afterwards for correctness.

2.3 Procedure

The participants arrived at the test room in groups of three, and were given instructions before the start of the experiment. As stated earlier, the pairs of words were presented one at a time, starting with the first word of the sentence paired with three x characters. Selection of the correct word was performed by pressing either the right or left shift, corresponding to the left and right word. When the right word was selected, reaction time was recorded, and the next word pair was presented. When the wrong word was selected, the error was recorded as a negative reaction time, the participant was informed of their error, and the next story was presented. When no selection was made within 4 seconds (4000 miliseconds), an error was recorded and informed to the participant, before continuing to the next story.

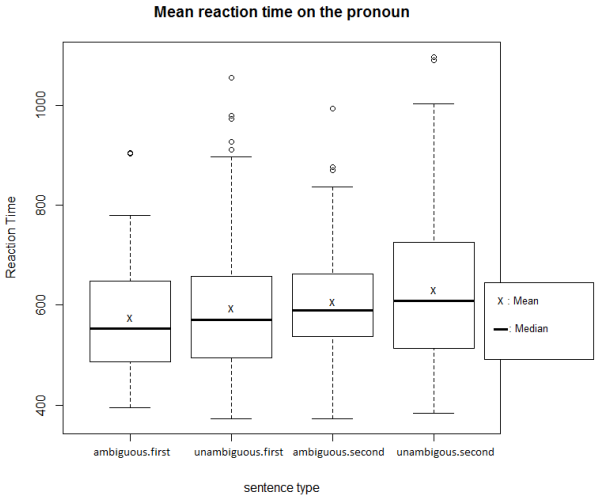

Figure 2: Reaction Times at the pronoun for all sentence types.

3 Results

3.1 Analysis

Analysis of the results showed that some test subjects failed to give any reply at all at the first words of the sentences. While a trial is terminated at a wrong reply, giving no reply simply causes the trial to continue after waiting 4000 ms, recording this as the reaction time. As a counter measure, and since these words didn't contain any relevant information, all first words of the sentences were removed. Furthermore, improbable reaction times in natural processing (smaller than 200 ms, and larger than 1100ms) were removed. 83.3% of the data was retained. As the distribution of reaction times was positively skewed, the reaction times were log-transformed to approximate a normal distribution. Linear Mixed Effect models were built to examine the relation between the sentence types, and the reaction times, at the pronoun and disambiguation word of the sentence, and one word there after. Furthermore, sequence was taken as a fixed effect to account for possible fatigue effects as the participants

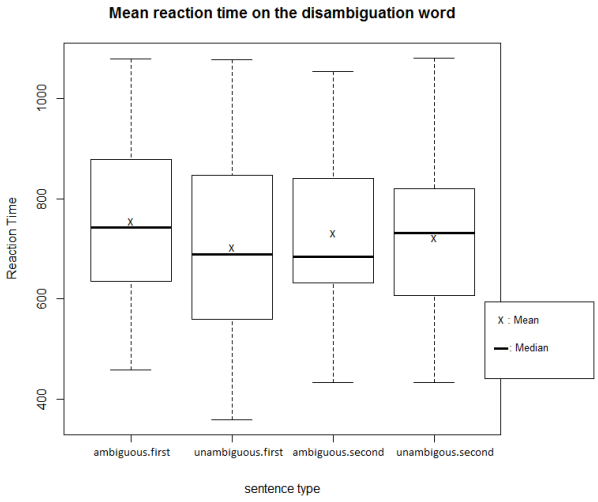

Figure 3: Reaction Times at the disambiguation word for all sentence types.

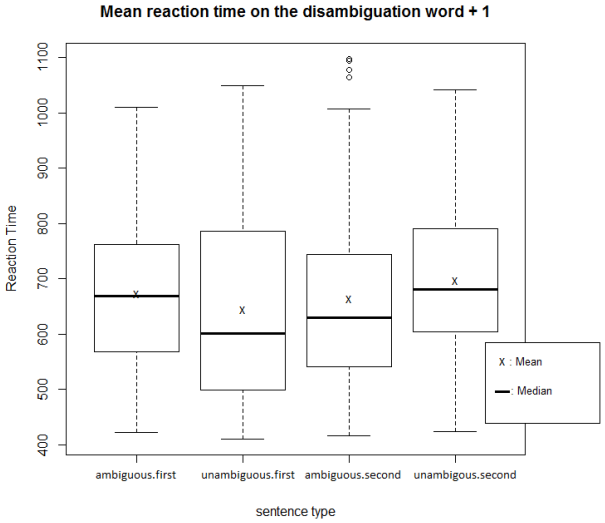

Figure 4: Reaction Times at one word after the disambiguation word for all sentence types.

got further into the experiment. Sequence describes the order in which participants read the sentences, so a higher sequence number corresponds to a sentence further into the experiment.

3.2 Pronoun

A null Linear Mixed Effects model was built for the effect of sentence type and sequence on the reaction time (RT) at the pronoun. The table of this model is shown in the appendix (Table 1). This model took ambiguity, order of mention and sequence as fixed effects. Three additional models were built, each omiting one of the fixed effects. Analysis of Variance was performed between the null model and the three models to see if the ommited fixed effects had an actual effect. Neither ambiguity nor order of mention had an effect on RT (ambiguity: $\chi^2 = 0.0067$, $p = 0.9346$ order of mention: $\chi^2 = 1.1278$, $p = 0.2882$). An effect of sequence was found on the RT at the pronoun ($\chi^2 = 42$, $p < 0.001$), with participants reacting faster as the experiment progressed (t = -6.77).

3.3 Disambiguation

A null Linear Mixed Effect model was build for the effect of sentence type and sequence on the RT at the disambiguation word. The table of this model is shown in the appendix (Table 2). Analysis of Variance was performed between the null models, and three models each ommiting one of the fixed factors, as was done at the pronoun. None of the factors had a significant effect on the RT (ambiguity: $\chi^2 = 1.4705$,$p = 0.2253$, order of mention: $\chi^2 = 0.0011$,$p = 0.9738$, sequence: $\chi^2 = 2.5894$, $p = 0.1076$).

3.4 Spillover

Linear Mixed Effects models were built for the effect of sentence type and sequence on one word after the pronoun, and one word after the disambiguation word, in the same manner as the models built for the Pronoun. The table of this model is shown in the appendix (Table 3 and Table 4). Analysis of Variance was performed between the null models, and three models each omiting one of the fixed factors, as was done at the pronoun.

3.4.1 One word after the pronoun

Neither ambiguity nor order of mention had a significant effect on the RT one word after the pronoun (ambiguity: $\chi^2 = 0.1217$,$p = 0.7272$, order of mention: $\chi^2 = 0.3143$,$p = 0.5751$). An effect of sequence on RT was found ($\chi^2 = 10.057$,$p < 0.01$),

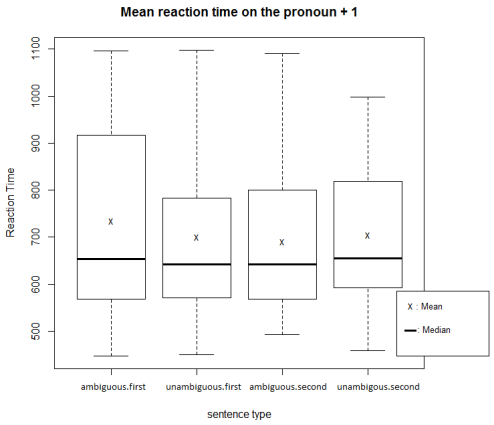

Figure 5: Reaction Times at one word after the pronoun for all sentence types.

with participants reacting faster as the experiment progressed (t = -3.24).

3.4.2 One word after the disambiguation word

Neither ambiguity nor order of mention had a significant effect on the RT one word after the pronoun (ambiguity: $\chi^2 = 0.0.947$,$p = 0.7583$, order of mention: $\chi^2 = 0.0947$,$p = 0.7583$). Once again, an effect of sequence on RT was found($\chi^2 = 18.93$,$p < 0.001$), with participants reacting faster as the experiment progressed (t = -4.51).

4 Discussion

In this paper, we set out to see if the G-Maze task proposed by Forster et al. would be capable to show the effects of pronomial ambiguity. In the end, no difference between ambiguous and non-ambiguous pronouns were found. To contemplate what might have caused this, we turn to an earlier study of the G-Maze task applied to different temporal ambiguities.

4.1 Earlier tests on G-Maze

In their paper, Witzel et al (Witzel (2012)) compared four different psycholinguistic methodologies, among which was the G-Maze task. Like in the current study, no sentences in the G-Maze trial were followed by comprehension questions. These methodologies were tested against three temporal ambiguous sentence structures different from pronoun resolution. We will consider the two where the G-Maze task showed the expected effects. In Relative Clause Attachment ambiguity, modification of a relative clause would cause it to refer to either the global, or local component noun of a complex noun. eg:

(3) a. The son of the actress who shot *herself* (local, low attachment)
b. The son of the actress who shot *himself* (global, high attachment)

In this structure, longer reading times were expected in sentences with high attachment, which was found in the G-Maze task trial.

In adverb attachment ambiguity, it was unclear whether an adverb referred to one of two verbs. Attachment was modifier by changing the tense of the verb. eg:

(4) a. Susan bought the wine she will drink *next week* (attachment to will drink, low attachment)
b. Susan bought the wine she will drink *last week* (attachment to bought, high attachment)

In this structure, once again, longer reading times were expected in sentences with high attachment, which was also found in the G-Maze task trial.

The study of Witzel et al. showed that the G-Maze task can indeed be a powerful tool in showing local difficulties in processing temporal ambiguities. Since no effects were found in the current study, there might be some fundamental difference between the phenomena tested by Witzel et al., and those studied in this paper. In both the relative clause attachment ambiguity and adverb attachment ambiguity, the role of the word or structure refered to by the word of interest (the relative clause and the adverb) was defined in a purely *syntactic* manner. In RC attachment, this would be either the global noun (the son) or the local noun embedded in the complex noun (the actress). In

Adverb attachment, the pronoun referred to either the verb in the main predicate (Susan *bought*) or the verb phrase modifying the object (The wine she will drink). In the current study, both possible referents of the pronoun have been subject and object, in either of the two preceding sentences. Since there is no syntactical difference between the possible referents, disambiguating the pronoun requires use of deeper semantic understanding of the sentence, whether this is the gender of the possible referents, or the world knowledge imparted by the disambiguation phrase. It could be this dependance on contextual meaning that causes processing time on the pronoun to not differ when it's initially ambiguous.

4.2 Comprehension Questions

Taylor (2013) studied pronomial ambiguity using a self-paced reading paradigm. He found an increase in reading times at ambiguous pronouns only when focus on semantic content was increased when a secondary task concerning the comprehension of the pronoun. No effect was found when this question was omitted. In Witzel's study, the assumption was made that due to the nature of the task, a comprehention question was unneccesary for the G-Maze task to show the expected effect. Due to the nature of the sentences discussed in the last section, it is possible that the type of sentences used in this study are simply processed too little for an effect to arise, even in the G-maze task. An inclusion of the aforementioned comprehension questions might help to increase processing enough for the G-maze task to show the expected effect.

4.3 Sequence

An effect of sequence was found on all words of interest, except for the disambiguation word, suggesting a learning effect where participants got better at the task as time went by. This is opposite of the fatigue effect that was expected, where reaction time would slow down due to the lengthy experiment. However, sequence had no effect on the reaction time at the disambiguation word. What might have caused this is unknown.

4.4 Conclusion

Though the G-Maze task shows great promise in showing effects of temporal ambiguities in a localised manner, a lot of questions remain unanswered. While it is clear that the restrictions the G-Maze task places on processing isn't enough to evoke the effects expected of pronomial ambiguity, at least in it's current setup, it isn't fully clear whether this lack of an effect is due to the lack of comprehention questions, or some other inherent feature of the G-Maze task. An interesting setup for future research would be to include these comprehention questions, and see if the expected effects are found.

References

Kenneth I Forster, Christine Guerrera, and Lisa Elliot. The maze task: Measuring forced incremental sentence processing time. *Behavior research methods*, 41(1):163–171, 2009.

Morton Ann Gernsbacher, David J Hargreaves, and Mark Beeman. Building and accessing clausal representations: The advantage of first mention versus the advantage of clause recency. *Journal of Memory and Language*, 28(6):735–755, 1989.

Martin J Pickering, Matthew J Traxler, and Matthew W Crocker. Ambiguity resolution in sentence processing: Evidence against frequency-based accounts. *Journal of Memory and Language*, 43(3):447–475, 2000.

R. Taylor. *Tracking Referents: Markedness, World Knowledge and Pronoun Resolution.* PhD thesis, University of Groningen, 2013.

J Trueswell and M Tanenhaus. Toward a lexical framework of constraint-based syntactic ambiguity resolution. *Perspectives on sentence processing*, pages 155–179, 1994.

Witzel J. & Forster K. Witzel, N. Comparisons of online reading paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research*, 41(2):105–128, 2012.

5 Appendix

The following tables show fixed-effects factors in the model fitted to the reaction time (RT) data at the pronoun, the disambiguation word, and one word after these regions. Estimated Coefficients, standard errors, and t values for the mixed-effects regression model fitted to the log transformed reaction times of the G-Maze task experiment. Positive coefficients for ambiguity reflect an increase in RT for unambiguous sentences, while positive coefficients for order of mention reflect an increase in RT when the pronoun disambiguates to the second mentioned character.

Table 1

	estimate	std. error	t value
(Intercept)	6.482485	0.038776	167.18
ambiguity	-0.001830	0.022306	-0.08
order of mention	0.024401	0.022937	1.06
sequence	-0.006955	0.001027	-6.77

Linear Mixed Effects model at the pronoun.
ambiguity + order of mention + sequence + $(1|Participant) + (1|item) + \varepsilon$

Table 2

	estimate	std. error	t value
(Intercept)	6.611790	0.053150	124.40
ambiguity	-0.040436	0.033216	-1.22
order of mention	0.001139	0.034516	0.03
sequence	-0.002449	0.001487	-1.65

Linear Mixed Effects Model at the disambiguation word.
ambiguity + order of mention + sequence + $(1|Participant) + (1|item) + \varepsilon$

Table 3

	estimate	std. error	t value
(Intercept)	6.599243	0.041022	160.87
ambiguity	-0.009397	0.026862	-0.35
order of mention	0.015503	0.027600	0.56
sequence	-0.003758	0.001158	-3.24

Linear Mixed Effects Model at one word after the pronoun.
ambiguity + order of mention + sequence + $(1|Participant) + (1|item) + \varepsilon$

Table 4

	estimate	std. error	t value
(Intercept)	6.592478	0.052246	126.18
ambiguity	0.009175	0.029761	0.31
order of mention	0.002559	0.031120	0.08
sequence	-0.006092	0.001352	-4.51

Linear Mixed Effects Model at one word after the Disambiguation Word.
ambiguity + order of mention + sequence + $(1|Participant) + (1|item) + \varepsilon$