Transformer-Maze

by

Annika Heuser

Submitted to the Departments of Brain and Cognitive Sciences and
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Computation and Cognition

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Author .....
Departments of Brain and Cognitive Sciences and Electrical Engineering
and Computer Science
August 5, 2022

Certified by .....
Edward Gibson
Professor, Brain & Cognitive Sciences
Thesis Supervisor

Certified by .....
Robert C. Berwick
Professor, Computational Linguistics
Thesis Supervisor

Accepted by .....
Mehrdad Jazayeri
Professor & Director of Education, Brain & Cognitive Sciences

Transformer-Maze

by

Annika Heuser

Submitted to the Departments of Brain and Cognitive Sciences and Electrical Engineering and Computer Science
on August 5, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Computation and Cognition

Abstract

Psycholinguists study online language processing to gain insight into both the different mental representations of various sentence types and the computational resources required to build those representations. Psycholinguists have a number of tools available to them, the most prevalent being eye-tracking and self-paced reading (SPR). However, a lesser-known tool called the Maze task, more specifically G(rammatical)-Maze, is arguably a better choice for detecting and localizing differences in processing difficulty from word to word. In G-Maze, a participant must choose between each successive word in sentence and a distractor word that does not make sense based on the preceding context. If a participant chooses the distractor as opposed to the actual word, then the trial ends and they may not complete the sentence. Like SPR, G-Maze can be cheaply run on a crowdsourcing platform, but it does a better job of localizing effects and filtering out noisy data. Still, the effort required to pick contextually inappropriate distractors for hundreds of words might cause an experimenter to hesitate before picking this method. Boyce et al. (2020) remove this hesitation with A(uto)-Maze, a tool that automatically generates distractors using a computational language model. In this thesis, we introduce the next generation of A-Maze: T(ransformer)-Maze. Transformer models are the current state of the art in natural language processing, and thousands, pretrained in a variety of languages, are freely available on the internet, specifically through Huggingface's Transformers package. In our validation experiment, T-Maze proves itself to be as effective as G-Maze with handmade materials, run in a lab. We are excited to provide psycholinguists with a new tool that allows them to easily gather high-quality online sentence processing data in many different languages.

Thesis Supervisor: Edward Gibson
Title: Professor, Brain & Cognitive Sciences

Thesis Supervisor: Robert C. Berwick
Title: Professor, Computational Linguistics

Acknowledgments

I am so privileged to have had the support of the incredible Abby Bertics. I doubt you realize it, Abby, but you were integral in this project getting off the ground in the first place. It was so kind of you to offer to help me with figuring out my package installation issue in Google Colab. I was beyond impressed by how quickly you figured it out after the hours I spent struggling with it. You are one of the smartest people I know. Knowing you had my back made this thesis, my greatest engineering endeavor to date, seem much more manageable. Your curiosity and our resulting discussions about AI and human cognition have not only enriched this thesis, but also me as a person. I am looking forward to many more in the future. Thank you, Abby, for always listening whenever I dove into random aspects of the project (often with absolutely no context, now that I think back on it). Thank you for always answering with thoughtful questions to help me better think things through and for always offering clever suggestions. Thank you so much for finding the time to help me edit this thesis. I know it must have been difficult while working a full-time job. It is so much clearer for your help and editing was much more enjoyable for me than it usually is because of your funny comments. Words nor ice cream have the power to allow me to express how grateful I am to you.

Thank you, Ted, for taking me under your wing and nurturing my interest in psycholinguistics and pursuing a career in academia. Neither this thesis nor my upcoming grad school adventure would have been possible without your support. I am really excited to keep working with you on this project in the future.

Thank you, Bob, for welcoming me into your lab when I was only a sophomore. Because of your trust in me, I became confident in my ability to conduct any research that interested me.

I am very lucky to have such a supportive family. Thank you, Dad, for your interest in my work. I really enjoyed being able to explain the project and my progress to you. Thank you for letting me read the finished sections to you in my excitement. Mom, thank you for all your encouragement throughout my time at MIT and for keeping me

sustained while I was writing this thesis. Your delicious cooking was instrumental to my productivity and morale. Thank you, Liam and Moose, for keeping me laughing while I was writing day in and day out.

Thank you, Kellie, for helping keep me sane throughout this entire thesis adventure. I am so lucky to have you as a friend.

Finally, thank you, Veronica, for your thoughtful responses to my questions about A-Maze. It was really nice to learn that such high quality work came from such high quality people.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent breakthroughs in artificial intelligence (AI), specifically in natural language processing (NLP), have led to the rise of computational models like OpenAI's GPT-3 (Brown et al., 2020), able to write academic papers about itself (Thunström and Steingrimsson, 2022). GPT-3 is so impressive that it has garnered mainstream media attention and prompted studies of the potential consequences of producing text indistinguishable from that written by humans (Floridi and Chiriatti, 2020; McGuffie and Newhouse, 2020). While some might argue that such AI advancements, especially of the NLP persuasion, are over-hyped, the fact remains that millions of dollars in funding are poured into building bigger and better models (Dale, 2022)[1]. A decent portion of this money must go towards the computational resources needed to train models with billions of parameters. These computational resources also have a steep environmental cost: training GPT-3, with its 175 billion parameters, was estimated to consume 1,287 MWh of electricity and produce 552.1 metric tons of CO_2 equivalent emissions (Patterson et al., 2021). For the sake of comparison, we will roughly estimate the amount of energy required for the human equivalent to GPT-3's training. Assuming an intake of 2000 calories a day[2], a human who just turned 20 years old would have consumed 14,600,000 calories over their lifetime. 14,600,000 calories is 0.017 megawatt-hours, or 0.0013% of the amount of energy needed to train

[1] See also: `https://www.nsf.gov/cise/ai.jsp`

[2] As recommended by the FDA, though we realize that this is an overestimate for the first several years of a human's life

GPT-3. It is also important to note that someone younger than 20 is capable of writing as eloquently as GPT-3. Additionally, writing is by no means the only skill, language-related or otherwise, that a human learns while growing up.

The human brain clearly reigns superior in language processing and production for the foreseeable future. The question then is how. In other words: how do our brains process and produce language so effectively? The overall vision of the fields of psycholinguistics and neurolinguistics is to answer this question. Plausible theories based on behavioral phenomena observed in speakers of various languages can guide neurolinguistic research. Online sentence processing relies on structure, meaning, attention, and memory, meaning that theories based on its observation can have far-reaching implications for linguistics and cognitive science.

Characteristics of sentence processing make it easier to observe. It is incremental, in that new linguistic information, whether the next phoneme or the next word, must be integrated into our understanding from the previous time step. Due to limited computational resources, the integration cost differs based on the context and properties of the new information. An increased integration cost is typically paid with more time, at the millisecond scale. By measuring how reading times change from word to word and sentence to sentence, researchers capture concise snapshots of online language comprehension and its computational constraints (Bartek et al., 2011; Gibson and Pearlmutter, 1998; Tanenhaus and Trueswell, 1995).

The most prevalent methods for collecting reading measures are eye tracking (Rayner, 1998) and moving-window self-paced reading (SPR; Mitchell, 1984). In eye-tracking, as the name suggests, an infrared camera *tracks* the movements of a participants eyes while they read a sentence projected onto a screen. It is relatively expensive because of the specialized equipment it requires. However, it results in high-quality, though complex, data, with several dependent measures to analyze. SPR, on the other hand, only has one dependent measure: a word's reading time (RT). In SPR, all but one word in a sentence are masked and the participant presses a button to re-mask the current word and reveal the next one. The time between button presses, during which a word is legible, is that word's RT. In an attempt to ensure

that a participant does not mentally check out while clicking through a sentence, each sentence is typically followed by a comprehension question. However, these questions tend to be so easy that an inattentive participant can often guess correctly on them. Additionally, analyses of SPR data often reveal what are known as "spillover effects", or greater average RTs after the expected source of the processing difficulty, not at it. Nonetheless, SPR's greatest advantage only became clear over the last decade: that researchers can run SPR experiments over crowdsourcing platforms (Enochson and Culbertson, 2015) like Amazon's Mechanical Turk (Paolacci et al., 2010) and Prolific (Palan and Schitter, 2018). Researchers can cheaply and quickly recruit much more diverse participant pools using crowdsourcing platforms. The only potential problem is that unsupervised participants paid per task are likely to optimize for speed, which means skimming for SPR. Because the comprehension question accuracy might not allow researchers to effectively filter out skimming participants, SPR can lose power over crowdsourcing platforms. In fact, a study confirmed this: the estimated power based on an SPR experiment run in-lab was greater for 2 out of 3 effects than the estimated power based on the same experiment run on Mechanical Turk (Boyce et al., 2020).

The Maze task (Forster et al., 2009) is another method for measuring incremental processing time differences that can be run on a crowdsourcing platform. In addition to better effect localization, the Maze task is much harder for an inattentive participant to complete. This is because participants are presented each word in the sentence, one at a time, along with a distractor, and to continue through the sentence, they must select the actual word over the distractor. A skimming participant is much more likely to pick a distractor, after which they are not allowed to finish the sentence. The trouble with the maze task is the effort required to pick out good distractors that obviously do not fit the context. Boyce et al. (2020) present a solution to this problem with A(uto)-Maze. A-Maze automatically generates distractors using a computational language model to determine which words are the least likely given the sentence's preceding context. Boyce et al. (2020) tested two versions of A-Maze, both based on recursive neural network (RNN) models, and found that they were

both better than SPR at detecting differences in processing difficulty at precisely the sentence region that the differences were expected.

RNNs have been supplanted by transformers as the state of the art in NLP. Recall that we started by discussing GPT-3, a massive language model that can generate text that is difficult to distinguish from that written by humans. GPT stands for "generative pretrained transformer". Many pretrained transformer models, including GPT-3's[3] predecessor GPT-2, are available on the internet for anyone to use. The accessibility of pretrained transformer models, especially those trained on different languages, motivated us to engineer Maze stimuli generation software based on the next generation of NLP technology. Inspired by Boyce et al. (2020)'s use of natural language processing technology as a means to better understand human language processing, we are excited to introduce T(ransformer)-Maze.

In the next chapter, we dive much deeper into the work upon which T-Maze was built, as well as discussing the pros and cons of alternative psycholinguistics methods for observing online sentence processing. Then in chapter 3, we explain the T-Maze system, specifically focusing on our design decisions. We describe our validation experiment in chapter 4. We find that T-Maze run on Prolific effectively detects and localizes processing difficulty differences. Finally, we discuss what we believe that T-Maze can bring not only to the field of psycholinguistics, but also artificial intelligence.

[3]GPT-3 is only available through OpenAI's API which dictates how a user can use it. Additionally, users need to pay for it after 3 months/enough usage. See: `https://openai.com/api/`

Chapter 2

Related Work

2.1 Online Reading Paradigms

Researchers have developed a number of experimental tasks to investigate how people comprehend written sentences. They have also developed a number of ways to evaluate these tasks. We will juxtapose three experimental tasks using five different criteria: 1) whether the task effectively reveals online sentence processing difficulty, 2) whether the difficulty is indicated precisely at the predicted word/sentence region, 3) the naturalness of the task, 4) the cost, and 5) the effort required to set up such an experiment.

We choose these five criteria because they give us insight into an experimental task's ability to contribute to the field's knowledge of sentence processing operations. Strong theories can be built off of clear and precise data of processing time differences across various sentence constructions. The argument in favor of natural tasks is that they should not distort typical sentence comprehension operations, and should therefore generalize to normal, unobserved reading (Witzel et al., 2012). The financial cost and experimenter effort of a task speaks to the replicability of the results found with it. Results found in an experiment that many different labs have the resources to replicate are much easier to validate.

The three experimental tasks that we consider are eye tracking, self-paced reading (SPR), and the maze task. Eye tracking and SPR, particularly non-cumulative,

moving-window, word-by-word SPR, are the most represented experimental tasks in the reading sentence comprehension literature. We will demonstrate how Transformer-Maze fits into the current state of the field by comparing and contrasting the maze task to the field's giants.

Witzel et al. (2012) compared the methods' performances on three different ambiguous sentence structures. Boyce et al. (2020) then compared Witzel et al. (2012)'s lab results to the results of web-based SPR and maze on the same structural ambiguities. We will accordingly use the same sentence types to be able to compare our results to Boyce et al. (2020)'s. These include relative clause (RC) attachment ambiguity, adverb attachment ambiguity, and noun phrase (NP) versus sentence (S) coordination ambiguity.

Here is an example from Witzel et al. (2012) of RC attachment ambiguity:

(1a) The son of the actress who shot *herself* on the set was under investigation. (*Low Attachment*)

(1b) he son of the actress who shot *himself* on the set was under investigation. (*High Attachment*)

(1a) is referred to as low attachment because the RC "who shot herself on the set" is attached to the local noun phrase immediately preceding it ("the actress"). The RC in (1b), on the other hand, is attached to the nonlocal NP further away ("the son"), resulting in the masculine reflexive pronoun. Witzel et al. (2012) and Boyce et al. (2020) found a strong preference for low attachment, meaning that (1b) resulted in more processing difficulty than (1a).

Here is Witzel et al. (2012)'s example for adverb attachment ambiguity:

(2a) Susan bought the wine she will drink *next week*, but she didn't buy any cheese. (*Low Attachment*)

(2b) Susan bought the wine she will drink *last week*, but she didn't buy any cheese. (*High Attachment*)

There are two verb phrases (VPs) to which the adverb can attach, the local "will drink" and nonlocal "bought the wine." The tense of the adverb indicates to which VP it is attached, matching the future local VP in the low attachment case of (2a) and the past nonlocal VP in the high attachment case of (2b). Witzel et al. (2012) and Boyce et al. (2020) also found a strong preference for low attachment.

Lastly, here is Witzel et al. (2012)'s example for NP vs S coordination ambiguity:

(3a) The robber shot the jeweler, and the salesman *reported* the crime to the police. (*Unambiguous - NP coordination*)

(3b) The robber shot the jeweler and the salesman *reported* the crime to the police. (*Ambiguous - S coordination*)

In (3b) the robber could have shot both the jeweler and the salesman, but when the reader gets to the verb of the conjoined sentence, it becomes clear that "the salesman" must be the subject as opposed to part of the direct object with "the jeweler". In (3a), however, the comma following "the jeweler" immediately shuts down the option of the the salesman being a part of the direct object. Boyce et al. (2020) found a preference for NP coordination with A-Maze but not with any other method while Witzel et al. (2012) only found one with eye tracking.

2.1.1 Eye tracking

In eye tracking, participants' eye movements are recorded by an infrared camera as they read full sentences on a screen. While our eye movements appear steady and continuous to us, they are actually made up of long (200-300ms) fixations broken up by rapid (about 30 ms) saccades. A participant's eyes have free rein to skid and stall over any part of the sentence at any time, creating an unconstrained space of possible results. Researchers have nonetheless found that lower skipping rates, longer overall looking times, and regressive as opposed to progressive saccades after the word are all indicators of greater processing difficulty. Which of these dependent measures and to what degree they will manifest for a given word can depend on that word's length,

frequency, and whether it is grammatically or semantically out of place in the context of the sentence. Eye tracking also permits variability from participant to participant, who are free to skim, cautiously read and reread a sentence multiple times to the point of memorization, or anything in between (Witzel et al., 2012). This freedom is the price for this task being the most natural of the three we discuss, and it is paid for by the experimenter, whose analysis is all the more complex for it. Participants must also come in to the lab to be outfitted with specialized equipment, which also makes eye-tracking experiments time-consuming before the data analysis even begins. Not all labs can afford the equipment and space to run an eye-tracking experiment, as well as the money to incentivize a diverse group of participants to come into the lab. Ultimately, the naturalness and the wealth of information that eye tracking provides comes at a steep price (Boyce et al., 2020).

2.1.2 Self-paced reading

In self-paced reading (SPR), participants press a button or computer key to click through a sentence on a screen. The time until the next button press, called the reading time (RT), is the only dependent measure. Researchers typically add comprehension questions after the sentence to discourage skimming. We focus on the most common type of SPR: non-cumulative, moving-window SPR. In this type of SPR, a participant can only see one word at a time while the rest are masked. While still very much in the spirit of reading, it is therefore significantly less natural than eye tracking, because of the forced focus on one part of the sentence. Additionally, once they click past a word, it becomes masked again and the participant cannot re-reveal it, even if they accidentally clicked past the word before they had finished integrating it into their mental model of the sentence. This often leads to participants slowing down on the subsequent words, despite these words not being the cause of the processing difficulty. The prevalence of these spillover effects means that SPR fares horribly on the second criteria of indicating the processing difficulty at precisely the predicted word. This is especially problematic when the researcher is conducting the experiment to determine the region of processing difficulty.

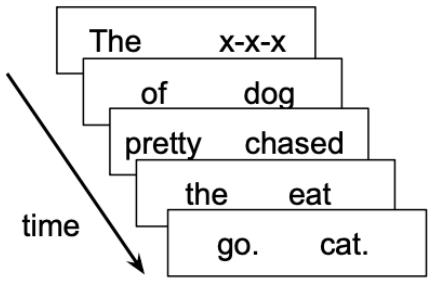
Attentive participants do tend to slow down, though a little late, when a sentence that they are clicking through causes them processing difficulty, meaning that it fares reasonably well with respect to the first criteria, at least in the lab. One of the major advantages of SPR though is that it does not require participants to come into the lab, they can be recruited on crowd-sourcing platforms like Amazon's Mechanical Turk (henceforth referred to as MTurk), where participants would click through the sentences and answer questions about them on their computers at home. Recruiting participants on crowd-sourcing platforms is significantly cheaper than recruiting them to come into the lab, and it allows researchers to recruit a much more diverse group of participants. A diverse group of participants increases the likelihood that any conclusions drawn from the study apply to all people, not just psychology students, for example. However, participants on crowd-sourcing platforms are incentivized to finish the task as quickly as possible in order to move on to another paying task. Accordingly, Enochson and Culbertson (2015) found that for SPR, MTurk workers had on average 180ms faster RTs than in-lab participants. Therefore, the comprehension questions may not guarantee that crowdsourced participants will read the sentences carefully enough to slow down when a region is more difficult to process.

Still Boyce et al. (2020) found a bias for low attachment in the reading times of crowdsourced participants reading sentences with adverb attachment ambiguity. The bias was apparent in the two words following the disambiguation region. For Witzel et al. (2012)'s in-lab participants, it was apparent at the predicted region in the reading times. Based on Boyce et al. (2020)'s Bayesian analysis; however, there was no effect for relative clause (RC) attachment disambiguation or noun phrase versus sentence coordination disambiguation in the SPR times of either the web-based participants or the Witzel et al. (2012)'s in-lab participants. However, it is important to note that Witzel et al. (2012) did find an effect for RC attachment disambiguation with their frequentist analysis. Still Boyce et al. (2020) found an effect through Bayesian analysis with the Maze task, both web-based and in-lab, for both RC and adverb attachment disambiguation. This suggests that SPR is the least effective of the 3 experimental methods that we discuss in detecting processing

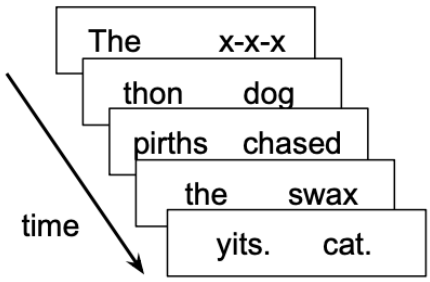
difficulty differences. Nonetheless, SPR is the easiest experiment type of the three we discuss for the researcher, as well as one of the cheapest, because it does not need to be run in the lab, researchers just need to think of simple comprehension questions to accompany their sentences, and there is only one dependent measure to analyze.

2.1.3 Maze

Forster et al. (2009) introduced the Maze task, in which participants are presented with the correct next word in the sentence alongside an obvious distractor word. In order to read the full sentence, they must always choose the correct word, otherwise the trial is terminated. They choose between the actual word and the distractor by pressing a button or computer key corresponding to the side of the screen of the word they are choosing. Like in SPR, the time between button/key presses, called the reaction time, is the dependent measure and is also abbreviated to RT. There are two flavors of Maze, based on what makes a distractor: L(exciality)-Maze, with nonce word distractors, and G(rammaticality)-Maze, with real word distractors that make little sense given the context. Figure 2-1 is an example of what a sentence might look like in either version of Maze.



(a) Sample G-maze



(b) Sample L-maze

Figure 2-1: For both G-Maze and L-Maze, the participant would choose "The" on the left, then "dog" on the right, "chased" again on the right, and so on (from Boyce et al. (2020)).

Maze is clearly the least natural of the 3 tasks—a reader never needs to bother with distractors. Nonetheless, Forster et al. (2009) found a garden-path effect at

the predicted region with both L-Maze and G-Maze. More specifically, Forster et al. (2009)'s L-Maze and G-Maze RTs were significantly faster for subject-extracted relative clauses (RCs) than for object-extracted RCs, corresponding with well-established eye-tracking results. When comparing Maze to eye-tracking and SPR, Witzel et al. (2012) found higher RTs at the disambiguating words for the RC and adverb high attachment sentences. The effect size was greater for G-Maze than for L-Maze. With Boyce et al. (2020)'s Bayesian analysis of Witzel et al. (2012)'s data, there was only a statistically significant effect for lab L-Maze in sentences with adverb attachment ambiguity. For lab G-Maze, however, the Bayesian analysis revealed an effect for both RC and adverb attachment disambiguation. This suggests that G-Maze is the stronger method with respect to the first criteria of the two Maze versions. Boyce et al. (2020)'s replication of Witzel et al. (2012) on MTurk also indicates that G-Maze is the strongest of the experimental methods that can be run online. Crowdsourced G-maze participants also had significantly greater RTs at the disambiguating words for both RC and adverb high attachment conditions, unlike crowdsourced L-maze and SPR participants. Therefore, the extra work that a participant must do for the maze task (i.e. reading another word, deciding which is the right word and which button/key to press) does not appear to distort a participant's typical sentence processing.

Because Maze can be run online, it is one of the most accessible methods to low-resource labs. However, researchers must pay for G-Maze's better effect detection and localization in the greater effort required to pick good distractors. For an SPR experiment, a researcher needs to think of one comprehension question per sentence, whereas for G-maze, a researcher needs to think of a distractor for each word in each sentence. In order to reduce the effort required to set up a G-Maze experiment, Boyce et al. (2020) introduce A(uto)-Maze, a method for automating the generation of distractor materials. They make A-Maze as well as the software for running the resulting Maze task on a crowdsourcing platform of the researcher's choice freely available online.

2.2 A-Maze

To generate distractor materials, A-maze runs potential distractors that are matched to the actual word by length and unigram frequency through a computational language model. A language model is defined as a probability distribution over sequences of tokens (Jurafsky and Martin, 2018). As such, a language model returns the probability of each of the potential distractors in the context of the sentence's preceding words. In A-Maze, the first potential distractor to have a conditional probability beneath a threshold, specifically 21 bits of surprisal, or about 4 in 10 million, is chosen as the distractor. If none of 100 potential distractors has a conditional probability under the threshold, then the distractor with the lowest conditional probability of the 100 is selected.

On their validation experiment, Boyce et al. (2020) ran A-Maze, as well as SPR, L-Maze, and G-Maze (using Witzel et al. (2012)'s materials for these experiments) on MTurk. They compared the results from all these web-run experiments to Witzel et al. (2012)'s lab results. Boyce et al. (2020) plugged two language models into their A-Maze distractor generation infrastructure: Gulordava et al. (2018)'s recurrent neural network (RNN) and Jozefowicz et al. (2016)'s RNN. They found that both A-Mazes detected the effects at the expected regions just as well as Witzel et al. (2012)'s in-lab G-Maze. Figure 2-2 shows that A-Maze Gulordova and A-Maze Jozefowics participants had significantly higher RTs for high attachment relative and adverb clauses, just like in-lab and online G-maze participants. Interestingly, the A-Mazes were the only methods with which Boyce et al. (2020) found a difference between the comma and no comma condition resulting in S v NP coordination ambiguity. The A-Maze method clearly passed its validation testing with flying colors, demonstrating that researchers can now much more easily create Maze experimental materials and run them on a crowdsourcing platform with Boyce et al. (2020)'s software.

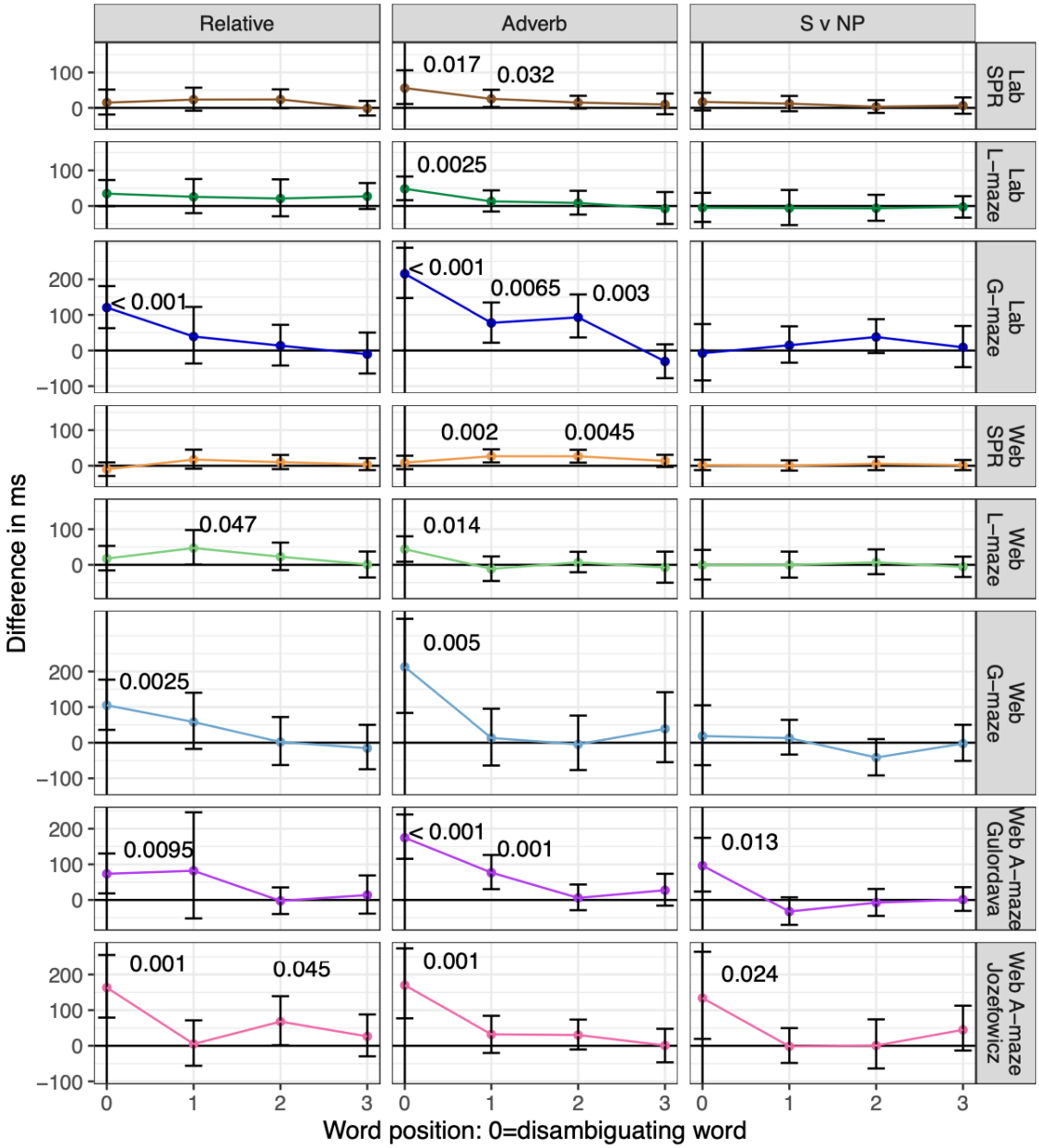


Figure 2-2: Results from Boyce et al. (2020)'s web-based validation experiments, compared to Witzel et al. (2012)'s in-lab results. These are plots of the 95% confidence intervals of the mean difference in RT between the dispreferred conditions (high attachment for relative and adverb clauses and no comma for S v NP ambiguity). Boyce et al. (2020) includes the Bayesian p-value equivalents that are < 0.05.

2.3 Transformers

Transfomer models replaced RNNs and the related long short-term memory (LSTM) and gated RNN architectures as the state of the art approach to many natural language processing applications (Vaswani et al., 2017). In language modeling specifically, transformer models are at the top of the WikiText-103 (Merity et al., 2016) and word level Penn Treebank (Marcus et al., 1994) language modeling leaderboards[1] (Brown et al., 2020; Shoeybi et al., 2019). In addition to greater performance, thousands of transformer models trained on hundreds of languages, are available to anyone online via Hugginface's open-source library, Transformers[2] (Wolf et al., 2020). This makes the Transformers library a one-stop-shop for any researcher, no matter their computational resources, wanting to use A-Maze for a different language.

However, many of the transformers in that library, including every BERT-based (Devlin et al., 2018) model, are not language models by definition, because they are trained on a masked language modeling objective. Unlike sequential recurrent architectures, transformers encode an entire sequence at once, add on a positional encoding to keep track of where the sequence tokens are in relation to each other, and then compute a representation of the whole sequence via self-attention (Vaswani et al., 2017). Aside from more parallelization, this also means that a token early in a sequence given to a transformer encoder, like "hungry" in "The hungry hippo wasn't satisfied after 20 watermelons." can attend to tokens later in the sequence like "20" and "watermelons." Masked language modeling, where a random word in the sentence is masked and then the transformer must predict what it is based on the surrounding context before and after the word, takes advantage of this. Because of the bidirectional nature of a masked language model (MLM), it does not define a probability distribution over a sentence's preceding context. This is problematic when the concept of A-Maze depends on such a probability distribution. We could limit ourselves to decoder-only transformers trained on a language modeling objective

[1] We look to the language modeling leaderboards here: https://paperswithcode.com/task/language-modelling. Most of the top models on these leaderboards are transformers.

[2] `https://huggingface.co/`

like GPT (Radford et al., 2018), where tokens are prevented from attending to tokens in subsequent positions. However, while such transformers dominate the language modeling leaderboards, the test datasets all consist of English text. The best models in other languages could be bidirectional transformers. In fact, the only monolingual German model (i.e., exclusively trained on German text) of which we are aware is an MLM called GottBERT (Scheible et al., 2020). It is possible that German GPT-2 (Schweter, 2020) might outperform GottBERT in language-modeling, but we failed to find any performance comparison. Regardless, Salazar et al. (2019)'s MLM scoring package allows researchers to plug many additional models from Huggingface's Transformers package into A-Maze.

Salazar et al. (2019)'s `mlm.scorers` Python package computes a sentence's pseudo-log-likelihood (PLL). The PLL of a sentence W is defined as

$$PLL(W) \coloneqq \sum_{t=1}^{|W|} \log P_{MLM}(w_t | W_{\setminus t}; \Theta)$$

where t is the token index, w_t is the token at index t, $W_{\setminus t}$ is the set of tokens not at index t, and Θ is the set of the MLM's parameters. The PLL score then is the sum of the log probability of each sentence token conditioned on every other token, whereas a language model's log probabilities can only be conditioned on the preceding tokens, i.e., $W_{<t} \coloneqq (w_1, ..., w_{t-1})$. Salazar et al. (2019) found that PLL scores could effectively predict which of two sentences was more acceptable according to human judgements. The tested MLMs all outperformed GPT-2 on the Benchmark of Linguistic Minimal Pairs (BLiMP) (Warstadt et al., 2020). This suggests that for our purposes of selecting the least acceptable distractor, PLLs could even prove more useful than the log probabilities A-Maze hinges on. In any case, we can also use Salazar et al. (2019)'s package for log probability sentence scoring based on some English GPT-2 models[3]. We build T(ransformer)-Maze to be compatible with any

[3]While we have not yet implemented this alternative, it should not be too difficult to build a functionally equivalent log probability sentence scoring object using other GPT-based models in the Transformers package with the functions from the `AutoModelWithLMHead| class at our disposal.`

scorer from `mlm.scorers`, maximizing the types of transformer models that can be plugged into it. For T-Maze's debut, we use BERT to calculate distractors' PLLs for direct comparison (Devlin et al., 2018). We are excited to explain how Salazar et al. (2019)'s MLM scoring package fits into our T-Maze implementation.

Chapter 3

Transformer-Maze

In this chapter, we will discuss our design decisions in engineering T(ransformer)-Maze. We first explain our decisions pertaining to setting up T-Maze to produce English materials. Then we discuss our general algorithm for collecting and evaluating potential distractors and how it differs from Boyce et al. (2020)'s A-Maze. Finally, we explain how to transform the resulting T-Maze materials into a web-hosted experimental interface that crowdsourced participants can navigate.

3.1 Language-specific setup

A word should be matched with a distractor that is roughly as easy to read. All potential distractors should therefore approximately match the word in length and frequency. We discuss how we define "roughly match" in section 3.2. For the sake of efficiency, we only ever evaluate well-matched distractors. We therefore create a data structure called `freq_bins` for quick access to the distractors best matched to a word. `freq_bins` is a dictionary whose keys are Zipf frequencies to the second decimal place (i.e. 4.89) and whose values are a list of words that all have that Zipf frequency. The Python package `word_freq` returns each word's Zipf frequency, which Speer et al. (2018) define as the base-10 logarithm of the number of times the word appears per billion words. Figure 3-1 shows part of the `freq_bins` dictionary for the replication of Boyce et al. (2020). Each language requires us to use a different

`freq_bins` dictionary, whose values are lists of words from only that language. While building `freq_bins`, we can save ourselves work later by already excluding inappropriate words and nonwords and inserting words that are typically capitalized as their capitalized forms. Constructing the `freq_bins` for new languages will always be time consuming. Here, I detail how we did this specifically for English.

```
5.55: ['already', 'anything', 'case', 'nothing', 'person', 'today'],
5.56: ['business', 'care', 'start', 'system', 'times', 'week'],
5.57: ['getting', 'god', 'government', 'group', 'looking', 'public', 'women'],
5.58: ['done', 'however'],
5.59: ['called', 'different', 'having', 'thought'],
5.6: ['company', 'doing', 'few', "he's", 'let', 'real'],
5.61: ['city', 'days', 'lot', 'name', 'night', 'play', 'until'],
5.62: ['away', 'left', 'number'],
5.63: ['free', 'second', 'someone'],
5.64: ['money'],
5.66: ['ever', 'family', 'keep', 'might', 'please', 'put'],
5.67: ['big', 'feel', 'sure', 'team'],
5.68: ['against', "didn't", 'end', 'found', 'must', 'show'],
5.69: ['each', 'without'],
5.7: ['again', 'next'],
5.71: ['give', 'house', 'place', 'school'],
5.72: ['during', 'game', "I've", 'thing'],
5.73: ['under'],
5.74: ['another', 'does', 'things'],
5.75: ['help', 'high', 'little', 'old', 'since'],
5.76: ['always', 'better', 'find'],
5.77: ['around', 'between'],
5.78: ['am', 'come', 'part', 'state', 'three'],
5.79: ['both', 'every'],
5.8: ["can't", 'same', 'used'],
5.81: ['home', 'long', 'look', 'something', 'use'],
```

Figure 3-1: `freq_bins` dictionary, where the keys are zipf frequencies and the values are list of words with those frequencies

3.1.1 Excluding nonwords

We first build a set of nonwords that should not be considered as potential distractors, unimaginatively called `nonwords_set`. The first time T-Maze was run on a sentence, there was no established set of nonwords. It was only after taking issue with some of T'Maze's chosen distractors that we added and began filling `nonwords_set`. We noticed in both German and English that several of the distractors, especially those matched to short words, were acronyms. Because we did not have any acronyms

in our actual sentences, we scraped the Wikipedia pages on English and German acronyms for their respective nonwords sets. We focus on the process of building the English `nonwords_set` for the remainder of this section because it is the one we used while creating the materials for the experiment discussed in this thesis.

The English acronyms Wikipedia pages contained over 4,000 acronyms, some of which were also actual words, like "care" and "are." To prevent adding actual words, we did not add acronyms that had more than 4 characters and a non-zero zipf frequency while scraping Wikipedia. Now two problems remain: 1) words like "care" are not added, but shorter words like "are" still would be. 2) Some longer and commonly used acronyms like "lmao" are now not added. To address the first problem, we hand-cultivated a set of actual words that we also checked before adding an acronym while scraping Wikipedia. The set comprises of the only English 1-letter words, "a" and "i"[1], common 2- and 3-letter words, including "no," "to," "can," and "are," as well as other under 3 letter words that we noticed doubled as acronyms while looking through the Wikipedia pages.

To address the second problem, we initialized the English `nonwords_set` with common acronyms such as "lmao" and "NCAA" that we realized had not made it into the set after our scraping process. In addition to common acronyms, the English `nonwords_set` is also initialized with potentially offensive words, common typos one often finds on the internet such as "ofthe," month abbreviations, and interjections like "um" or "ahh." We also add all the words in Boyce et al. (2020)'s `exclude.txt` file before scraping Wikipedia for acronyms to add to the `nonwords_set`. This file includes nonwords of the same flavor as already described, as well as some Spanish words and all single letters except for "a." However, we delete the word "i" from the file, because it is a common pronoun and we make sure to properly capitalize it so that it is recognized as such.

We decide to exclude all these word types despite their prevalence on the internet, because our experimental and filler sentences are representative of more formal, written American English and because it is in our best interest not to offend any

[1]All words are initially lowercase even if their dominant form is uppercase.

participants with potentially offensive distractors. While it should be obvious that the distractor is not a word in the sentence that a participant is reading, it should only be obvious because the distractor does not make sense given the preceding context. Participants should not be able to differentiate a distractor from the actual word without context because the distractor is an acronym while none of the actual words have been acronyms or because it is only found in less formal contexts.

The words that make up the lists in `freq_bins` are from the `wordfreq` python package (Speer et al., 2018). We chose this package because it allows for easy iteration over all the word frequencies and the corresponding words that have those frequencies. It also supports 44 different languages, making it easier to set up T-Maze for many other languages than just the ones we tested. Their data comes from 8 different domains of sources, including Wikipedia, subtitles, news, as well as Twitter and Reddit, which explains the inclusion of internet slang.

In addition to compiling a `nonwords_set` we also have a language-general (at least general to most, if not all, languages with Latin alphabets) nonwords check and an English-specific check. Other language-specific checks should be easy to add with the English-specific check as an example. This check simply results in the exclusion of words without vowels[2] while we iterate over the English `wordfreq` frequency dictionary. The language-general check makes sure that words in the `nonwords_set`, words with numbers in them, and words containing a period or comma are excluded.

The definition of a nonword depends on the context of the specific language, as well as flavor of that language, as dictated by the experiment. We will provide the sets of nonwords that we compiled for English and German. However, researchers wanting to use T-Maze for other languages will need to compile entirely different sets of nonwords, perhaps using a similar process or perhaps establishing a starkly different one. Even researchers wanting to use T-Maze for English and German may want to consider compiling different nonwords sets, depending on their experimental materials. For example, if a researcher's experimental sentences contain acronyms, then acronyms should probably also be considered as potential distractors.

[2]We count "y" as a vowel.

3.1.2 Capitalizing words

The keys in the `wordfreq` frequency dictionary are all lowercase words. Each key is assigned a value that corresponds to the sum of the frequencies of that word's capitalized and lowercase forms. Looking up the capitalized form of a word returns the same value as its lowercase form, meaning that `wordfreq` returns a Zipf frequency of 7.09 for both "i" and "I". While sentences in which the pronoun are not capitalized are relatively common in informal English, the word's dominant written form is capitalized. Conversely, a word like "difficult" would only be capitalized in specific contexts, such as at the beginning of a sentence or as part of a title. We argue that for almost all words, there is a dominant form, either capitalized or lowercase, that eclipses the other form in terms of frequency. The only counterexamples we could think of were the German pair "sie" and "Sie", both pronouns with different meanings and therefore both rather common, and the English pair "may", the auxiliary, and "May", the month and name. Despite these counterexamples, we maintain the assumption that every word has a strong dominant form, and we only consider that dominant form as a potential distractor.

We developed two algorithmic approaches for determining every word's dominant form. For the first, we rely on a spaCy (Honnibal et al., 2020) pipeline for part of speech (POS) tagging. We capitalize all words tagged as proper nouns, and we check for personal pronouns. If a personal pronoun starts with "i'" as we'd expect for words such as "I'm" or "I'll," then we capitalize the "i". spaCy has 4 pretrained English pipelines, 3 of which have a POS tagging accuracy of 97%. `en_core_web_trf` ("trf" for "transformer") is slightly better, with a 98% tagging accuracy. While the higher accuracy and the matching transformer method were appealing, using `en_core_web_trf` to compile our `freq_bins` dictionary took over 5 times as long as using the lightweight `en_core_web_sm`. We argue that the additional time and computing resources required for `en_core_web_trf` are not worth the 1% improved accuracy. We find that `en_core_web_sm` tends to overgeneralize. For example, it tags "democratic" and "national" as proper nouns even though they are adjectives,

probably because they are often capitalized in the context of organization titles. We collected the overgeneralizations that we observed in a set to prevent their capitalization with this approach.

The second approach relies on the `language-tool-python` package[3], a wrapper for LanguageTool, an industry spelling and grammar checker. The tool returns a list of suggested changes if it is passed text with any errorsWe check the first three suggestions. The first of them made all lowercase that matches the `word_freq` key is the one we insert in the `freq_bin` dictionary as a word's dominant form. We are able to properly capitalize words like "iPhone" with this approach. Additional handling is required for proper nouns with apostrophes like "London's" and "O'Malley". We must call LanguageTool on the apostrophe's prefix and then in the case of Irish names beginning with "O'", we call it again on the suffix. Contrary to spaCy's `en_core_web_sm`, the `language-tool-python` package seems to only undergeneralize, in that it will not correct lowercase names that are also words, no matter how much more common the name is. We were overjoyed to learn the meanings of "tony", "rick", "sally", "batman", and many more. For those wondering: "batman" is a dated term for an officer's personal servant in the British military (Stevenson, 2010). We also collected these names in a set to automatically capitalize them despite LanguageTool taking no issue with their lowercase forms.

While we did not do any side-by-side formal testing of the two approaches for our purposes of single English word capitalization, we suppose that the `language-tool-python` package makes fewer mistakes than the spaCy pipelines. Based on our observations of the approximately 100 most frequent word lists, the only types of mistakes LanguageTool makes are failing to capitalize names that have dictionary definitions. Meanwhile both `en_core_web_sm` and `en_core_web_trf` are much less consistent in their error patterns. However, the better performance for our purposes costs more time (approximately 2.5 hours) and computing resources.

Because we already created three different `freq_bins` dictionaries to test the different tools at our disposal, we decided to create an ensemble dictionary. `en_core_web`

[3]https://github.com/jxmorris12/language_tool_python

`_sm`, `en_core_web_trf`, and `language-tool-python` are all given equal voting power, so if two of the three indicated that a word should be capitalized, then it is capitalized in the ensemble dictionary. Despite both the spaCy pipelines tending to overcapitalize, we noticed that `en_core_web_sm` and `en_core_web_trf` often made mistakes on different words. Therefore, while we are sure that some words that should not be capitalized might still be capitalized in our ensemble dictionary because both the spaCy pipelines voted for that, we hope that this happens less often than names that also have dictionary definitions being capitalized because both spaCy pipelines overpowered the LanguageTool. However, we have no evidence supporting our optimism so other researchers may find the `language-tool-python` more effective, at least for English. We also definitely would not recommend building an ensemble dictionary from scratch because of the time required. Both spaCy and LanguageTool support many other languages. LanguageTool even supports different regional flavors of languages like English, German, and Portuguese, so we hope both approaches will help guide researchers interested in using T-Maze for experiments in different languages.

This is by no means a perfect process and we concede that our English `freq_bins` dictionary still contains nonwords and incorrectly capitalized words. For this reason, we built in a parameter for how many of the top distractors to save. We frequently ran T-Maze with this parameter set to 3, meaning that we saved the 3 best distractors for each word in each sentence. Then, in the case that the top distractor is not a word that makes sense in the context of our experimental materials, we can replace it with the second best distractor, assuming we would not also consider it a nonword. We wrote a function to aid researchers with this process[4]. However, we must note that this reduces the number of distractors evaluated because while the top distractor does not count, we are not evaluating another distractor in turn. We also wrote functions allowing us to delete a list of nonwords from an already existing `freq_bins` dictionary and switch the capitalization of a list of words in it. These functions, as well as the discussion of our language-specific processes should save researchers a lot of time in building `freq_bins` that result in distractors that are well-matched to the

[4]We did not use this function for our validation experiment.

experimental materials in terms of language flavor. We expect that this language- and even experiment-specific setup will always be the most time-consuming step in using T-Maze, especially in the case of a new language. We therefore fantasize that other researchers who use T-Maze will provide their `freq_bins` dictionaries, especially for new languages, to help future users.

3.2 Collecting potential distractors

We allow the user to decide how many potential distractors to evaluate for each word. This gives the user the power to make computational trade-offs and language-specific adjustments. Behind the scenes, we grab the number of desired distractors that are within a range of word lengths and Zipf frequencies. We will now discuss this process and how we compute those ranges in greater depth.

The user-defined parameter for how many potential distractors to evaluate per word is called `num_eval`. If `num_eval` is set to 10,000, then we would consider potential distractors with a greater variety of Zipf frequencies and lengths. If we set it to 100, we consequently evaluate distractors within a much stricter range. With our frequency bins, we already have an intuitive way to collect distractors with a similar unigram frequency to the word we want to replace. We simply need to define the number of bins above and below the starting bin, corresponding to the Zipf frequency of the word we want to replace, from which we can collect. We arbitrarily initialize the upper bound Zipf frequency bin that we can check to be 0.1 greater than the starting bin. This means that we can check the 10 frequency bins above the starting bin and the 10 below it for potential distractors, giving us 20 to search in total.

For word length, we could also think of each word length as a bin and only consider words in the two bins above and below, for example. This is problematic when you have a short word, like "is." Our starting bin would then be 2, and the bins below would be 1 and 0. There are only 2 1-character words in English and obviously no 0-character words. We are less likely to be able to evaluate the number of specified distractors when left with just 2 decent-sized bins (the 1- and 2- character bins are

comparatively very sparse). Instead, we approximate the distribution of distinct word length as a normal distribution. We know a normal distribution is not the best approximation but we chose it over the double exponential found to closely approximate the empirical distinct word frequency because of its simplicity (Smith, 2012). All we need to describe a normal distribution is a mean and standard deviation. The mean word length (in graphemes) of distinct English words is approximately 7.26, with a standard deviation of 2.28 (Marian et al., 2012). There is less probability density at the tails of a normal distribution, reflecting the problem with shorter words.

The goal is then to find the bounds, centered at the length of the word we are replacing, for 20% of the cumulative probability density. In other words, our goal is to solve this equation for a:

$$\int_{l-a}^{l+a} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma}} dx = 0.2$$

where l is the length of the word we want to replace with a distractor, μ is the mean word length, and σ is the word length standard deviation. For English, $\mu = 7.26$ and $\sigma = 2.28$. Figure 3-2 includes an illustration of the idea behind our word length bounds calculation. We relied on the SciPy (Virtanen et al., 2020) `stat.norm` module to practically carry this out. First, we determine the value of the cumulative distribution function (CDF) for the length of the word we are replacing. Then, we find the CDF bounds by adding and subtracting 0.1 to this value. This means that we get 0.2 if we subtract our upper CDF bound from our lower CDF bound. We ensure that the lower and upper bound do not exceed 0.01 or 0.99 respectively. We then inverse the CDF bounds back to word length bounds with `stat.norm`'s percent point function (ppf) and convert the resulting values to integers. It is important to note that because of the rounding down, the resulting bounds do not encompass exactly 20% of the cumulative distribution, always less, and they are often not perfectly centered around the length of the word we wish to replace. Because we do not extend the CDF upper bound when we hit the lower limit of 0.01 and vice versa, bounds for word lengths at the distributions tails can encompass much less than 20%, not

fully solving the original problem. However, we were hesitant to further off-center the CDF bounds in the case that we hit the upper or lower limit. It is possible that extending the upper bound when we hit the lower limit and vice versa would still result in length bounds that still reasonably match the original word length, but we did not investigate this. Nonetheless, our method results in bounds that fluctuate with distance from the distinct word length mean. For example, in English, the bounds for a word of length 3 would be [1,4], ranging 5, while the bounds for a word of length 6 would be [5,6], only ranging 2. We therefore consider this an improvement upon fixed bound ranges, with which we would need to reiterate over distractors more frequently at the tail ends of word lengths. The most frequent words tend to also be the shortest (Zipf, 1949), so this should save us a lot of computational work in the long run.

We iterate through the word lists mapped to by the Zipf frequencies within the frequency bounds and collect the potential distractors also within the word length bounds until we have `num_eval` distractors. Punctuation counts as a character so "I've" is considered a 4-character word, for example. We start at the frequency bin of the actual word we are trying to pair with a distractor and then search the bins directly above and below. Next, we search the bins directly above and below those, and so on. That way, if there are enough distractors within our bounds specified by the initial parameters, the top `num_eval` that we select are the best matched in terms of frequency. In the opposite case of not exceeding `num_eval` even after collecting all the potential distractors within the bounds, then we increase our initial parameter values. These parameters define the greatest Zipf frequency bin we can search and the CDF range, so they are initially 0.1 greater than the Zipf frequency of the word we are replacing and 0.2 respectively. They are both arbitrarily increased by 0.1. Because we also increase our word length bounds[5], we must iterate through

[5]When we produced our validation experimental materials, we did not realize that we were not actually increasing our word length bounds as we increased our frequency bounds. After fixing this, we found that 20.22% of the distractors would have been different. We repeated our analysis after excluding these distractors and found the same results. The analysis we present in this thesis includes these distractors because we believe that they would have been indistinguishable from the distractors that we would have used instead in the results.

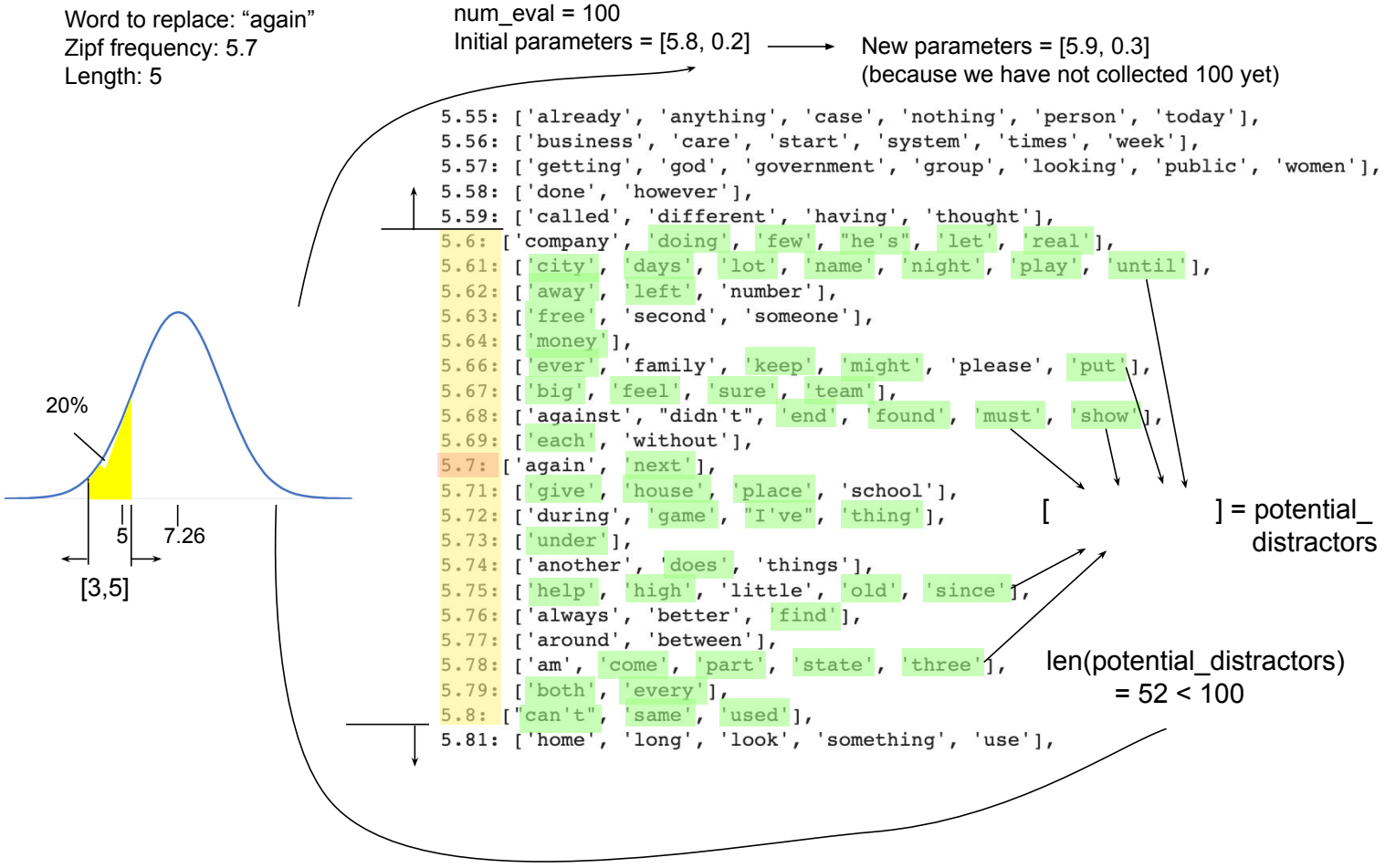


Figure 3-2: The process of collecting distractors. All the words highlighted in green are considered as potential distractors because they are within the yellow highlighted zipf frequency and word length bounds. These bounds are widened if there are not enough potential distractors within them to evaluate the number that the user specified (`num_eval` = 100), which is the case here because there are only 52 potential distractors in the bounds specified by the initial parameters.

the bins included in the initial frequency bounds again in case they have longer or shorter words that we could not consider during the initial pass. We maintain a set of distractors that we have already evaluated so that we do not reconsider any distractor within the frequency and word length bounds for more than one pass. We repeat the process of finding all the words within the bounds specified by the parameters and then increasing the parameters to consider more words until we have collected as many distractors as the user specified via the `num_eval` parameter.

3.3 Evaluating potential distractors

To evaluate a distractor, we add it after the sentence's preceding context and then score the resulting sentence with an `MLMScorer` from Salazar et al. (2019)'s Python package that we discussed in section 2.3. For example, to evaluate the potential distractor "similar" for "washed" in "The raccoon washed its hands." then we would evaluate "The raccoon similar". While the scorer might be using a bidirectional transformer, it is only given the preceding context because that is all a participant has access to with the G-Maze paradigm when deciding between the actual word and the distractor. We match the actual word's punctuation, like Boyce et al. (2020), but not the capitalization, unlike Boyce et al. (2020). This means that to evaluate the potential distractor "climb" for "Steve" in "Our neighbor, Steve, lent us some flour." we would score "Our neighbor, climb,".

Also unlike Boyce et al. (2020), we do not define a threshold PLL value such that the first potential distractor with a lower PLL is chosen. PLLs cannot be interpreted like log probability and we found that different models score the same sentences differently so we do not think that a PLL threshold makes sense. We must therefore evaluate all `num_eval` potential distractors and while maintaining a list with the best so far. This list consists of tuples with the distractor's pseudloglikelihood (PLL) score (or conditional log probability if the `MLMScorer` is using GPT-2) and the distractor itself. The user defines the length of this list. A list of length 1 is enough to pair all of an experimenter's sentences with distractors. If the experimenter decides that

the list will be length 3, then the top 3 distractors for each word are saved in a pandas (McKinney et al., 2010) DataFrame. Along with the top n distractors and the necessary information for an experimenter to determine where each distractor belongs (i.e. word position, item number, and ranking among the top n), we also store the value of `num_eval`, the time taken to generate the distractor, and the part of speech (POS) tag of the word to which the distractor it matched. The experimenter can also choose to record the POS tag of the distractors themselves. The DataFrame is ultimately saved in a comma-separated values (csv) file. The experimenter can use the csv file to replace any problematic distractors with the second or third best distractor or to run analyses on the distractors generated by T-Maze. Through such an analysis, an experimenter could find that distractors that have the same POS tag as the words with which they are being matched tend to be harder to distinguish as distractors, for example. They could then adjust their materials or T-Maze's distractor generation accordingly.

Sometimes an experimental item might consist of two minimally differing sentences. The critical region contains the only difference, defining the two conditions. This is the case for Witzel et al. (2012)'s materials. They used the same distractors for both conditions of an experimental item when producing both their G-Maze and L-Maze materials. It is therefore impossible to attribute any of the difference in RTs between the two conditions to the distractors. For their A-Maze materials, Boyce et al. (2020) also only generated one distractor per word position of each item. A researcher can also match distractors across experimental items with T-Maze. We use the same distractor matching approach as Boyce et al. (2020). At the critical region, where the actual words the distractor is being paired with are different, we determine the bounds for potential distractors based on the average length and frequency of the actual words. After the critical region, the preceding context differs based on the condition. To determine the best distractor for both conditions' preceding context, we score potential distractors twice. We insert the potential distractors after the first condition's preceding context and score the resulting text, and then repeat with the second condition's preceding context. The potential distractor's score is the average

of the PLLs from the two preceding contexts. We then choose the distractor with the best average PLL of all those evaluated. For now, T-Maze can only match distractors across items with just two conditions. Experimenters can of course also match each sentence in an item with its own distractors with T-Maze.

3.4 From distractor generation to experimental web-interface

We rely on Boyce et al. (2020)'s Ibex module[6] to transform our generated materials into an experimental interface for crowdsourced participants. PCbex Farm (Zehr and Schwarz, 2018), a server for web-based experiments, allows us to easily load the module from Github. When generating our distractors, we format the original sentences and the distractors into javascript that can be inserted directly into the `sample.js` file in Boyce et al. (2020)'s Ibex module. We expect that researchers will find it very easy to host T-maze experiments on a web-server because of our automatic javascript formatting and the open-source experimental tools from Boyce et al. (2020) and Zehr and Schwarz (2018).

[6] `https://github.com/vboyce/Ibex-with-Maze`

Chapter 4

Validation Experiment

In T-Maze's maiden voyage, we replicate the experiment Boyce et al. (2020) used to demonstrate A-Maze's efficacy.

4.1 Methods

4.1.1 Materials

Our experiment is based on the same materials as Witzel et al. (2012) and Boyce et al. (2020). We reformatted the experimental items in `g_maze.js` in the Boyce et al. (2020)'s A-Maze Github repository [1] to generate new distractors for them with T-Maze. For easy reference, we will reiterate the sentence structures meant to elicit processing effects that we discussed in section 2. The sentence structures set up 3 types of syntactic attachment ambiguity:

- relative clause (RC) attachment ambiguity
- adverb attachment ambiguity
- noun phrase (NP) versus sentence (S) coordination ambiguity

Table 4.1 contains examples of each type of ambiguity. These sentences represent each condition. Participants read these sentences as part of Witzel et al. (2012) and

[1] `https://github.com/vboyce/Maze`

Table 4.1: Example stimuli for each condition. The disimbiguating words are italicized.

Relative clause - Low attachment:
(4a) The niece of the fisherman who got *himself* a sailboat learned to sail.
Relative clause - High attachment:
(4b) The niece of the fisherman who got *herself* a sailboat learned to sail.
Adverb clause - Low attachment:
(5a) Robert will meet the friend he phoned *yesterday*, but he doesn't want to.
Adverb clause - High attachment:
(5b) Robert will meet the friend he phoned *tomorrow*, but he doesn't want to.
Sentence vs noun phrase (S v NP) coordination - With comma:
(6a) The crowd cheered for the model, and the designer *took* a bow after the show
Sentence vs noun phrase (S v NP) coordination - No comma:
(6b) The crowd cheered for the model and the designer *took* a bow after the show.

Boyce et al. (2020)'s experiments, as well as our own. Based on their results, the (a) sentence types are supposed to be easier for native English speakers to process. We therefore generally expect the low attachment and comma conditions to have mean lower RTs.

We chose to generate our distractors with `bert-base-uncased`. `bert_base_cased` and `bert-base-uncased`, both with only 110 million parameters (Devlin et al., 2018), are among the models with the fewest parameters that Salazar et al. (2019) tested. Nonetheless, `bert-base-cased` outperformed GPT-2 (with 345 million parameters) on the Benchmark of Linguistic Minimal Pairs (Salazar et al., 2019). Still, the general trend with transformers that Salazar et al. (2019) also observed with PLLs is that the greater the number of parameters, the greater the performance. We hope to determine T-Maze's baseline performance in generating distractors with `bert-base-uncased` because we would expect larger models to also outperform it in producing distractors. `bert-base-uncased` is trained on lowercase English text while `bert_base_cased` is trained on the same English text but maintaining its punctuation[2]. We chose `bert-base-uncased` over `bert_base_cased` because we know that our capitalization procedure (refer to section 3.1.2) is not perfect and we did not want the model assign-

[2]We found this information in the Huggingface model cards: `https://huggingface.co/bert-base-cased` and `https://huggingface.co/bert-base-uncased`. These pages recommend citing Devlin et al. (2018), however, we could not find this information in the paper itself.

ing a capitalization mistake a lower PLL than it would have assigned the distractor had it been properly capitalized.

Like Boyce et al. (2020), we matched distractors across sentences in the same item (see section 3.3) and we did not quality control our distractors after they were generated. The random distractor side assignment, the instructions, the debriefing, and the welcome page were all the same as for Boyce et al. (2020) because they were based on those in `header.js` in the A-Maze repository. We include the `sample.js` that we plugged into Boyce et al. (2020)'s Ibex Maze module in the appendix.

4.1.2 Participants

We recruited 50 participants on Prolific (Palan and Schitter, 2018). We used Prolific's prescreening to ensure that we only got participants located in the United States who indicated that they have American nationality and that their first language is English. We paid each of our participants $4.38 because we expected the study to take them on average 25 minutes. Prolific ensures that a participant can only complete the study once. Because of some technical difficulties[3], we only ended up with data from 49 participants. We were satisfied with this because Prolific's prescreening allowed us to include all their data in our analysis. In comparison, Boyce et al. (2020) needed to exclude some participants for indicating that they were not native English speakers, not United States citizens or not currently located in the United States. As a result, Boyce et al. (2020)'s A-Maze analyses were based on 46 and 42 participants respectively.

4.1.3 Procedure

Eligible participants who wanted to do our task clicked the link we provided in Prolific that led to our experiment hosted on PCIbex Farm (Zehr and Schwarz, 2018). Just like in Boyce et al. (2020), participants were first told how we would use their data

[3]We did not realize that the Maze Ibex module requires a participant to have a physical keyboard. We hope that admitting this mistake will allow researchers to run web-based Maze experiments more smoothly in the future.

and then asked to give their informed consent. Participants know that they can retract their consent by returning their submission on Prolific if they so choose. Next, participants were asked to enter their Prolific ID. Then they were given instructions and 8 practice sentences. Afterwards they worked through 96 sentences: 24 sentences of each type of ambiguity (i.e. RC attachment, adverb clause attachment, and S v NP coordination), and 24 filler sentences. There was a progress bar at the top of the screen to let participants know approximately how much they had left. This matches Boyce et al. (2020) and Witzel et al. (2012) in spirit, but we rely on the progress bar to serve the same purpose as arranging the sentences into 8 blocks of 12 items and informing participants after each block of how many they have left. At the end of the study, participants all received the same code that they entered in Prolific's interface to demonstrate that they completed the experiment. We estimated that the experiment would take on average 25 minutes, which was about accurate—the median completion time did not exceed this estimate. For those interested in what it was like to participate in our T-Maze validation experiment, please refer to this PCIbex Farm demonstration: `https://farm.pcibex.net/r/PFuPTr/`.

4.1.4 Data analysis

We excluded the data of one participant who returned their submission after working through some of the sentences, leaving us with the data of 49 participants[4]. If a participant makes a mistake, the trial ends at that word, and we cannot collect that participant's data for the rest of the sentence. We removed 2.3% of the data for being mistakes and 15% for being blank because of a mistake earlier in the sentence. This leaves us with 83%. Surprisingly, this is greater than the amount of data Boyce et al. (2020) had left for L-Maze (75%), which is an easier task than G-Maze. It's also greater than the percentage of RTs left for web-based G-Maze (64%) and both A-Mazes (64% for Gulordava and 55% for Jozefowics). We attribute this to our Prolific

[4]If a participant returns their submission on Prolific, then Prolific allows you to recruit another participant to take their place. Therefore, even after excluding data from participants who returned their submissions, we would have ended up with 50 participants had it not been for the aforementioned technical problem

participants being more careful or attentive, as indicated by their greater average time spent on the task (25 vs. 15 minutes).

To make our results directly comparable with Boyce et al. (2020)'s, we conduct our analysis with their R (R Core Team, 2022) code. This code calculates the difference in RT between the two conditions at each word position, starting from -5 with respect to the critical word's position (or 5 words before the critical word) to +5. In the case of a two-word critical region (e.g. "next week" in sentence (2a) in section 2), the RT for position 0 is calculated by averaging the RTs of both the words in the region (e.g. 1/2*RT for "next" + 1/2*RT for "week"). We use the same mixed effects Bayesian model, fit via `brms` (Bürkner, 2017):

$$\texttt{log(RT)} \sim \texttt{condition + (condition|subject) + (condition|item)} \quad (4.1)$$

We also re-run the analysis on the Boyce et al. (2020)'s A-Maze (Gulordava and Jozefowicz) and G-Maze data (web-based and from Witzel et al. (2012)'s lab) for comparison. Because p-values belong to frequentist inference (Wagenmakers et al., 2008), we discuss two-sided p-value equivalents. In this case, the p-value equivalent is defined as $1 - q$, where q is the probability mass contained in the largest symmetric interval that does not include 0 on the posterior distribution for 'condition'. We also report estimated effect sizes as the mean difference in RT between the dispreferred conditions, represented by the (b) sentences in table 4.1, and the preferred conditions, represented by the (a) sentences (Boyce et al., 2020).

We also run the same post-hoc power analysis as Boyce et al. (2020) to quantify T-Maze's sensitivity and compare it to the G-Mazes and A-Mazes. Boyce et al. (2020) estimate the probability of finding a significant effect (i.e., p-value equivalent ≤ 0.05) as a function of the number of participants via Monte Carlo simulation. Each participant's data loss rate was sampled from a normal distribution specified by the experimentally observed mean and variance. This data loss rate then determines the probability that any line of data should be eliminated from the simulated participant's

data. The RTs were simulated by sampling parameter values from the fitted brms model's (4.1) posterior. We then run the same analysis (with again the same model) that we ran on the real data, though this time only on the critical region and using `lme4` (Bates et al., 2015) because it is faster. Finally, we estimate the statistical power as the proportion of simulations that had a significant effect size. Following Boyce et al. (2020), we stimulate groups of participants of size 10, 20, 30, and so on to 60, running 500 Monte Carlo simulations per method/simulated participant count combination and report the estimated statistical power.

4.2 Results

In figure 4-1, we compare T-Maze's estimated effect sizes of each type of attachment ambiguity with the methods Boyce et al. (2020) found to be the most effective. Table 4.2 also includes the results for word positions before the critical region and positions +4 and +5, as well as all the p-value equivalents, not just those less than 0.05. In figure 4-2, we quantify and compare T-Maze's sensitivity for each type of attachment ambiguity.

In brief, figure 4-1 indicates that T-Maze reveals large localized effects as well as lab- and web-based G-Maze, and arguably also both A-Mazes. While A-Maze and G-Maze sometimes have some spillover effects that are smaller than the effect at the critical region, T-Maze tends to have them more often. Figure 4-2 also generally demonstrates that T-Maze is approximately as powerful as lab and web G-Maze. We now dive into the results specific to each type of attachment ambiguity.

4.2.1 Relative clause disambiguation

For relative clause attachment disambiguation, T-Maze reveals a significant effect of 137 ms ($p < 0.001$). This is most similar to Lab G-Maze's effect of 121 ms, whose p-value equivalent is also < 0.001. The 105 ms web G-Maze effect, the 74 ms A-Maze Gulordava effect, and the 163 ms A-Maze Jozefowicz effect, all with p-value equivalents < 0.01, are also in the general ball park. T-Maze has a spillover

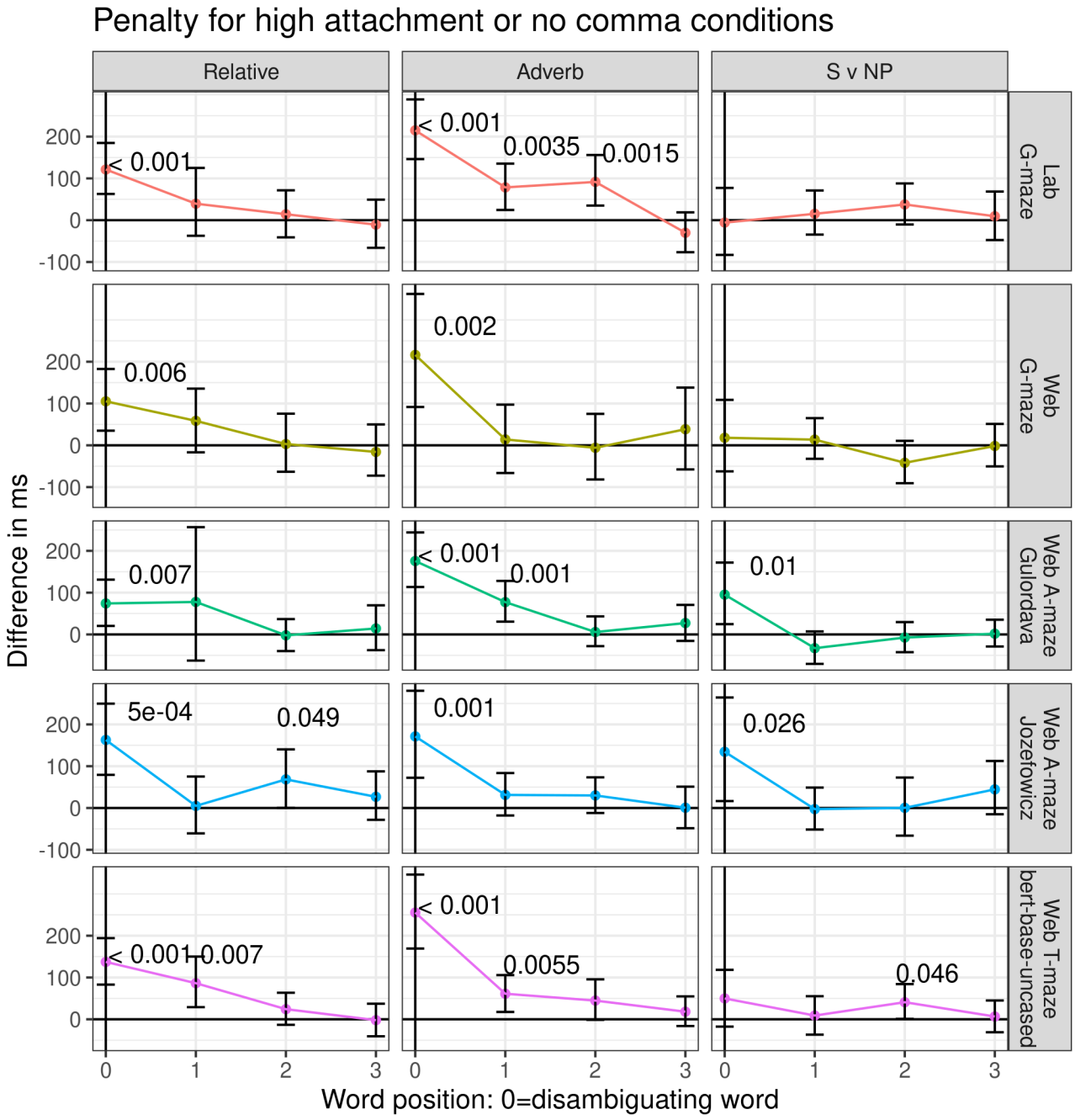


Figure 4-1: Estimated effect sizes with error bars indicating the 95% confidence intervals and p-value equivalents when $p < 0.05$. We include Boyce et al. (2020)'s data for comparison.

Word Position	Lab G-maze	Web G-maze	Web A-Maze Gulordava	Web A-Maze Jozefowicz	Web T-Maze bert-base-uncased
			Relative Clause		
-5	3 (0.89)	-4 (0.89)	10 (0.53)	-11 (0.54)	6 (0.82)
-4	-3 (0.89)	8 (0.69)	-9 (0.65)	-8 (0.53)	1 (0.98)
-3	39 (0.081)	-9 (0.72)	-14 (0.39)	-20 (0.48)	19 (0.34)
-2	-32 (0.31)	0 (0.97)	15 (0.43)	-2 (0.92)	31 (0.13)
-1	-33 (0.33)	-42 (0.26)	34 (0.2)	-14 (0.69)	19 (0.4)
0	**121 (0)**	**105 (0.006)**	**74 (0.007)**	**163 (5e-04)**	**137 (0)**
1	39 (0.3)	58 (0.12)	78 (0.29)	5 (0.88)	**86 (0.007)**
2	14 (0.6)	3 (0.93)	-2 (0.91)	**68 (0.049)**	24 (0.21)
3	-11 (0.7)	-16 (0.6)	14 (0.58)	27 (0.35)	-2 (0.9)
4	-50 (0.072)	-52 (0.19)	17 (0.47)	17 (0.51)	-18 (0.37)
5	-9 (0.76)	-22 (0.61)	-16 (0.49)	20 (0.57)	-4 (0.84)
			Adverb Clause		
-5	-13 (0.53)	-27 (0.55)	-9 (0.64)	3 (0.94)	19 (0.32)
-4	-11 (0.54)	-23 (0.46)	-15 (0.4)	11 (0.62)	12 (0.46)
-3	13 (0.66)	-65 (0.16)	10 (0.68)	15 (0.61)	-6 (0.73)
-2	33 (0.27)	0 (0.99)	32 (0.2)	16 (0.56)	11 (0.6)
-1	-17 (0.4)	-3 (0.91)	18 (0.46)	-26 (0.34)	8 (0.57)
0	**215 (0)**	**216 (0.002)**	**176 (0)**	**171 (0.001)**	**255 (0)**
1	**78 (0.0035)**	14 (0.71)	**77 (0.001)**	31 (0.21)	**61 (0.0055)**
2	**92 (0.0015)**	-6 (0.89)	6 (0.76)	30 (0.15)	45 (0.055)
3	-30 (0.21)	38 (0.39)	27 (0.2)	1 (0.96)	18 (0.28)
4	22 (0.4)	40 (0.19)	0 (0.98)	13 (0.64)	23 (0.25)
			S v NP		
5	22 (0.37)	31 (0.38)	-15 (0.41)	-42 (0.12)	4 (0.82)
-5	-6 (0.77)	-31 (0.3)	-2 (0.93)	-11 (0.72)	15 (0.32)
-4	3 (0.89)	2 (0.94)	-27 (0.43)	-55 (0.13)	-28 (0.13)
-3	-18 (0.4)	-13 (0.63)	-6 (0.72)	4 (0.86)	-16 (0.34)
-2	**-46 (0.047)**	-24 (0.37)	-13 (0.38)	5 (0.85)	-2 (0.9)
-1	-5 (0.85)	-38 (0.077)	-11 (0.54)	-32 (0.29)	-18 (0.21)
0	-6 (0.89)	18 (0.69)	**95 (0.01)**	**134 (0.026)**	50 (0.15)
1	15 (0.58)	14 (0.57)	-33 (0.1)	-3 (0.91)	9 (0.69)
2	38 (0.12)	-42 (0.11)	-7 (0.69)	0 (1)	**41 (0.046)**
3	9 (0.73)	-2 (0.94)	2 (0.92)	45 (0.15)	7 (0.73)
4	-28 (0.36)	-26 (0.42)	4 (0.84)	-6 (0.91)	8 (0.71)
5	-13 (0.71)	-2 (0.98)	28 (0.29)	-29 (0.54)	10 (0.78)

Table 4.2: Mean RT differences between the dispreferred conditions (high attachment or no comma/(b) sentences) and preferred conditions, followed by p-value equivalents in parentheses. We include Boyce et al. (2020)'s results for comparison. Values with $p < 0.05$ are bolded.

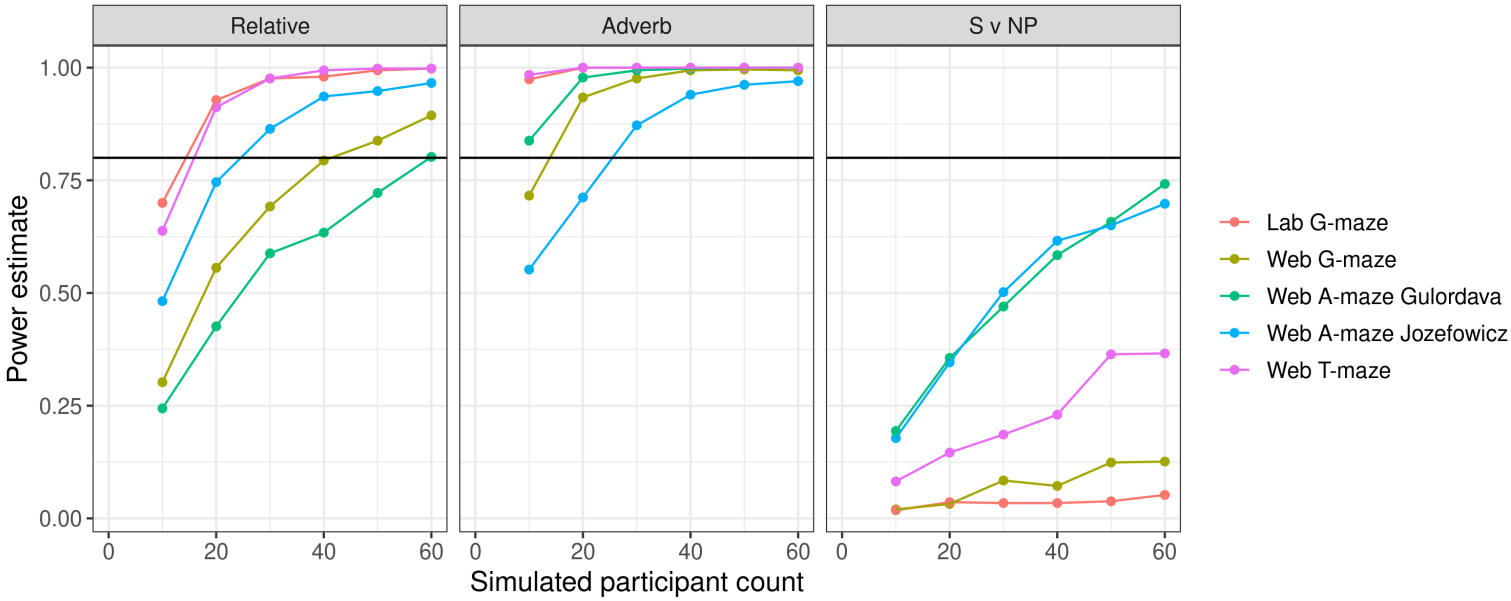


Figure 4-2: Estimated power for different numbers of participants based on observed data from different methods

effect of 86ms at position +1, that is greater than A-Maze Gulodava's effect at the disambiguating region in terms of milliseconds, and equal in terms of significance. A-Maze Jozefowicz also has a spillover effect of similar size in terms of milliseconds (68 ms) at +2 but it is barely significant ($p = 0.049$). T-Maze more or less matches lab G-Maze in power, both being the only two methods to exceed an estimated power of 0.8 with just 20 simulated participants.

4.2.2 Adverb clause attachment disambiguation

For adverb clause attachment disambiguation, the effect sizes are greater across the board. T-Maze, with 255 ms ($p < 0.001$), has the greatest effect size at the critical region in terms of milliseconds and significance, though tied with lab G-Maze (125 ms, $p < 0.001$) and A-Maze Gulordava (176 ms, $p < 0.001$) for significance. Web G-Maze and A-Maze Jozefowicz, with effect sizes of 216 ms ($p = 0.002$) and 171 ms ($p = 0.001$) respectively, are by no means far behind. The majority of methods also had a smaller but still significant spillover effect at position +1. More specifically, lab G-Maze had an effect of 78 ms ($p = 0.0035$) at +1, A-Maze Gulordava an effect of 77 ms ($p = 0.001$)

and T-Maze an effect of 61 ms ($p = 0.0055$). Lab G-Maze's spillover continues into position +2, with an effect size of 92 ms ($p = 0.0015$), which is greater than lab G-Maze's effect at +1 in terms of both milliseconds and significance, surprisingly enough. The spillover almost continues into position +2 for T-Maze as well, but the effect of 45 ms barely misses significance with a p-value equivalent of 0.055. T-Maze matches G-Maze in power more closely for adverb attachment disambiguation. Having detected such strong effects, they both start with an estimated power near 1 with just 10 simulated participants.

4.2.3 S v NP coordination disambiguation

Lastly, for S v NP coordination disambiguation, only the A-Mazes have significant effects at the critical region. A-Maze Gulordava has an effect of 95 ms ($p = 0.01$) and A-Maze Jozefowicz an effect of 134 ms ($p = 0.026$). Notably these effects are much smaller than those found with every method we discuss for the other types of attachment disambiguation. At position +2, T-Maze detects a barely significant ($p = 0.046$) effect of 41 ms. It is possible that this is a real effect, that our participants slowed down after the critical region because the sentences without commas were harder for them to process but they integrated the disambiguating verb late. It is possible, but we doubt it. In table 4.2, we can see that G-Maze detects an effect of almost the same significance at position -2, but this is almost certainly a spurious effect. Position -2 is before the disambiguating region, and the sentences and distractors for the two conditions are exactly the same for all word positions preceding 0, the disambiguating region. Therefore, the effect cannot be attributed to the stimuli. This suggests that other effects of similar size, like the effect T-Maze detects at +2, could very well be spurious too.

Out of 5 groups of 50 participants reading the S v NP sentences in a G-Maze paradigm, only 2 slowed down at the critical region. These 2 groups also made the greatest number of mistakes while working through the sentences, as is apparent in figure 4-3. When a participant makes a mistake, the trial ends and no more data is collected for the rest of the sentence. This means that all the errors A-Maze

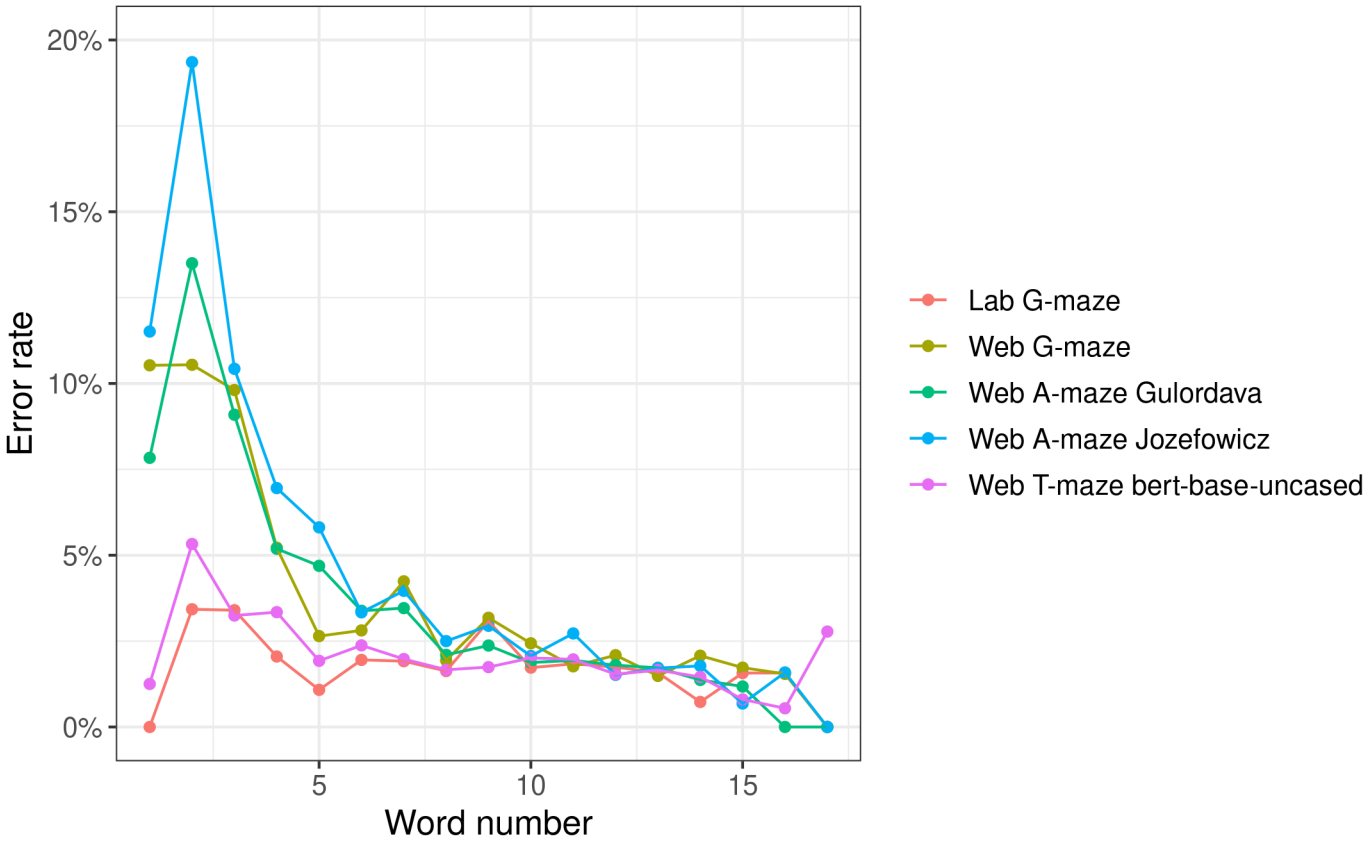


Figure 4-3: Participant error rate at each word position, where word 1 is the first word in the sentence (always paired with the distractor "x-x-x"). Lab G-Maze participants could not make an error at word 1 because they simply pushed a button to continue to word 2.

participants made at the beginning of sentences resulted in less data being collected.

We must also point out in figure 4-3 that T-Maze participants make a comparable amount of mistakes at the beginning of the sentence and otherwise to Witzel et al. (2012)'s in-lab G-Maze participants. Perhaps we can attribute this to `bert_base_uncased` generating better distractors than the A-Mazes' RNNs, even at the beginning of the sentence where the models get very little context to take into account. However, we think it is more likely due to T-Maze's Prolific participants being more attentive than the MTurk participants. We think this for 2 reasons: 1) Lab G-Maze participants and web G-Maze participants were faced with the same distractors, but web G-Maze participants had a similar error rate as A-Maze Gulordava participants. This suggests that the distractors themselves are not the main driving factor behind the error rate.

2) Prolific has been found to produce higher quality data for online behavioral studies than Mechanical Turk (Peer et al., 2021).

Before we continue to argue that the S v NP coordination disambiguation effects found by the A-Mazes might not replicate, we first turn to Boyce et al. (2020)'s explanation as to why no effect was found with lab and web G-Maze. How difficult a distractor is to discern from the actual word affects how much time a participant will take to select a word. Uncertainty about which word is the distractor will slow participants down, which can confound the delay of the dispreferred condition's greater processing difficulty. Boyce et al. (2020) demonstrated with figure 4-4 (but without T-Maze) that there could very well be such a confound obscuring a possible effect for lab and web G-Maze. For both S v NP coordination conditions, lab and web G-Maze have the greatest error rates at the disambiguating region, suggesting that their distractors were more difficult for participants to discern than they were for the other methods. However, it is also important to note that for most of the other methods, the error rate was also greater at the S v NP conditions' critical regions than at the surrounding word positions. Additionally, the error rate at the critical region is lowest for T-Maze, which does not find a S v NP coordination disambiguation effect. This suggests that perhaps lab and web G-Maze had no underlying effect that was obscured by difficult-to-discern distractors.

Moreover, the A-Mazes, even though they are the only methods to detect a significant effect at the critical region, they are still arguably under-powered for S v NP coordination disambiguation (see figure 4-2). Nonetheless, they are unsurprisingly the the most powerful for this type of attachment disambiguation. However, even with the maximum 60 simulated participants, both A-Mazes fall short of the .80 standard for power in behavioral sciences (Cohen, 1992).Even based on just the significant A-Maze results, only about 65% of simulated replications with 50 participants would find another significant effect. Taking into account the insignificant results of the other methods, which were often higher powered than the A-Mazes for the other types of attachment disambiguation, the actual rate of replication with the A-Maze materials is likely less than 0.65. However, we must note that α, the established risk of a type I

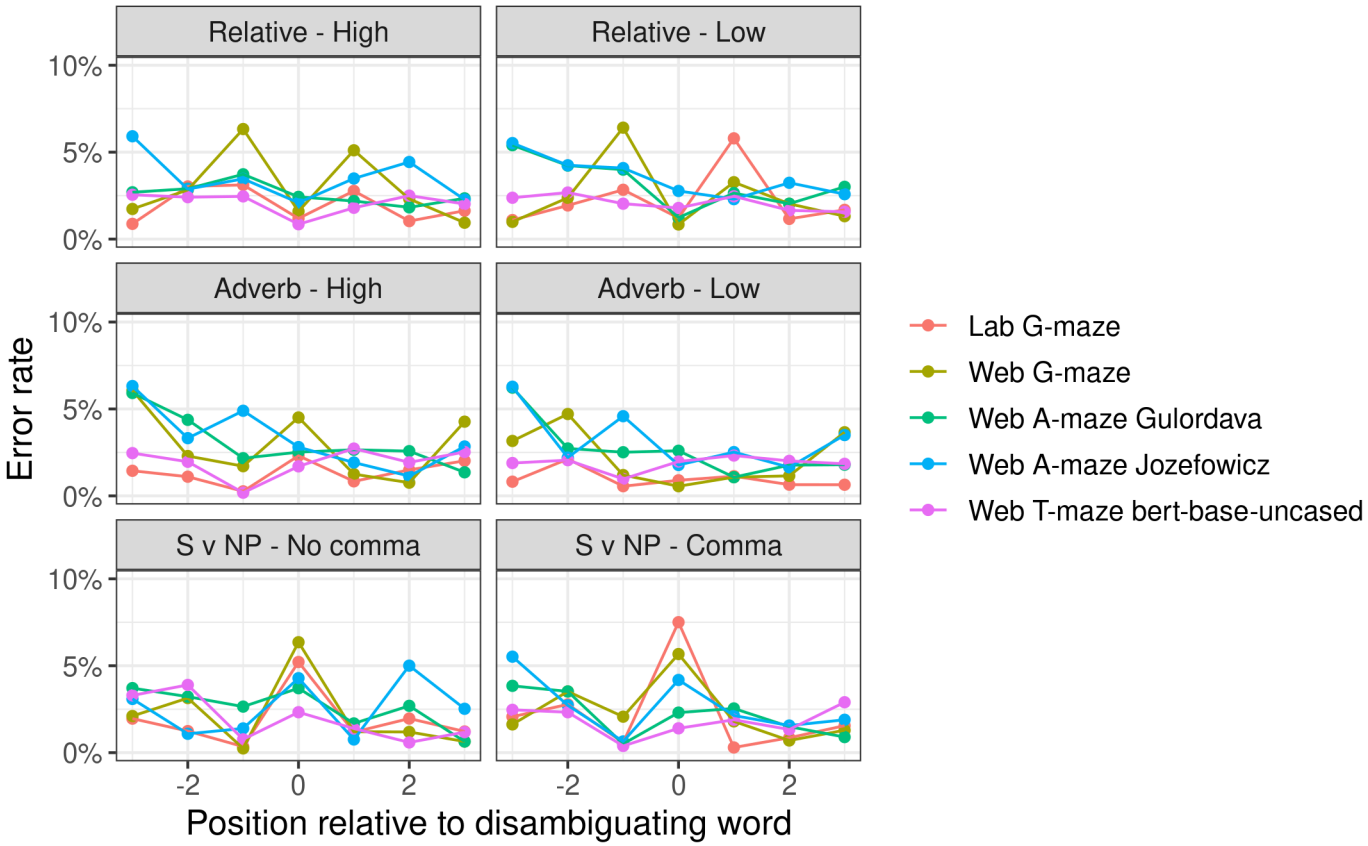


Figure 4-4: Error rates for each condition's critical region. T-Maze tends to have some of the lowest error rates, especially at the critical regions.

error (i.e., finding an effect even though the null hypothesis is true), is two-sided for these analyses. Defining it as one-sided would increase power across the board and it would make sense to do so, because we hypothesize the dispreferred condition (the no comma condition) to have greater RTs than the preferred comma condition because it should cause greater processing difficulty. If we treated our p-value equivalents as frequentist p-values, then the one-sided p-value would be half that from a two-sided test, assuming the difference is in the expected direction (Jones, 1952). Nonetheless, halving all the p-value equivalents at the critical regions still does not result in G-Maze (lab nor web) nor T-Maze having significant effects, though T-Maze would only be 0.025 off.

While we cast doubt on whether the A-Maze S v NP results would replicate if we were to use the same materials, we do not argue that the disambiguation phenomenon itself is not real. Two studies found an effect with SPR, which Boyce et al. (2020) found to generally be less powerful A-Maze and G-Maze: Frazier and Clifton (1997), investigating English, and Frazier (1987), investigating Dutch, which is typologically similar to English. Unlike relative clause attachment disambiguation (for which we observed much stronger effects), Witzel et al. (2012) mention no studies that fail to find an effect[5]. We instead suspect there might be a confound in Witzel et al. (2012)'s S v NP sentences.

In table 4.3, we highlight some no comma condition sentences that might be about as easy for readers to process as their comma condition counterparts for various reasons. In the first sentence (part of item 28), a reader would probably expect parallelism in the case that both the cat and dog are in the conjoined direct object. In other words, a reader would probably instead expect "The little girl fed her pet cat and dog." or "...fed her pet cat and pet dog." or even "...fed her pet cat and her pet dog." The hierarchical imbalance of "The little girl fed her pet cat and her dog" could tip the reader off that "her dog" is the start of a new clause. In the second sentence, the fact that the guitarist is presumably a part of the band might

[5]It is possible that such studies were conducted but not published because of publishing bias (Neuliep, 1990; Tincani and Travers, 2019).

Item number	No comma sentence
28	The little girl fed her pet cat and her dog wanted a can of food, too.
31	The audience applauded the guitarist and the band cheered for him very loudly.
34	The woman dressed her baby and her son got his clothes from the dresser.
37	The teacher praised the girl and her family was proud of her good grades.

Table 4.3: No comma condition sentences that might be easier to process

let the reader know early that "the band" will be followed by an action. In the third sentence, both the woman's baby and her son are her children, so a reader might expect someone wanting to refer to them both for the direct object to instead write, "The woman dressed her children." Additionally, the word choice of "baby" next to "son", suggests that the woman's son is significantly older than his "baby" sibling. Therefore, a reader might assume before the disambiguating verb that "the son" is not apart of the object because it is unlikely that an older child would still be dressed by their mother. In the last sentence, it is also unlikely for a teacher to praise "the girl and her family" because the majority of the time that a teacher is with their students, the students' families are not there at the same time. It is therefore much more likely the teacher would just praise the girl and that "and her family" is starting a new clause about their positive reaction, which is exactly the case. There are 24 S v NP items, and we present 4, or 16.66% , as potentially problematic. There may be more, we simply picked out the ones for which we could immediately think of reasons that the no comma condition might not be any more difficult to process than the comma condition.

Even though T-Maze does not detect a significant effect for S v NP coordination disambiguation, the results still validate T-Maze as a powerful method for detecting localized sentence processing effects. In this section, we demonstrated that T-Maze compares to A-Maze as well as in-lab G-Maze in terms of effect sizes at the critical region and estimated power. T-Maze's participant error rate is similar to lab G-Maze's and better (i.e. smaller) than A-Maze's, but this in all likelihood because we tested T-Maze on Prolific, which is known to have more attentive participants than MTurk (Peer et al., 2021).

Chapter 5

Contributions

In this thesis, we reviewed methods for detecting online sentence processing effects. We argue that G-Maze, while an unnatural task, has more potential than eye-tracking and self-paced reading because it can be run over crowdsourcing platforms, automatically filters out data from inattentive participants, and effectively localizes differences in processing difficulty. Also in its favor is A-Maze, which uses a sequential language model to automate distractor pairing. This removes one of the greatest hurdles for researchers interested in running G-Maze experiments: the time and effort required to think of good distractors for hundreds of words. We make running G-Maze experiments, especially in other languages, that much easier with the development of T(ransformer)-Maze. Transformer models are the current state of the art, and many transformers, pretrained on a variety of languages, are available online, for anyone to use. Huggingface's Transformers library is a testament to this.

Through our validation experiment, we demonstrated that T-Maze is as effective as G-Maze run in a lab, with handmade materials, at localizing differences in processing difficulty due to syntactic attachment disambiguation. We will make T-Maze as easy for other researchers to use as possible, as an open-source python package accompanied by thorough documentation. After testing T-Maze on other languages, we will also make the frequency bins available to spare researchers the trouble of the language-specific setup. We hope that T-Maze enables and encourages more labs to collect the empirical evidence they need to develop and test theories of online language

understanding.

That generally explains how T-Maze could contribute to the field of psycholinguisitcs. In chapter 1, we also promised to explain how it could contribute to the field of artificial intelligence. In this thesis, we do not add to the field's understanding of transformer models, we only use them as an engineering tool. To summarize the work's relationship with AI from chapter 1, especially because we thought ourselves quite clever in writing this: we make "use of natural language processing technology as a means to better understand human language processing." This statement is a 180 of a common concluding sentiment of research investigating some aspect of human intelligence. The idea is that with a better understanding of our own intelligence, we should be able to computationally implement more efficient and generalizable learning algorithms inspired by our own. After all, McCulloch and Pitts (1943) modelled their artificial neuron, the building block of the Perceptron (Rosenblatt, 1958) and many neural networks to follow, after biological neurons. This same idea lets us travel the remaining 180 degrees to come full circle. In other words, with T-Maze, we use natural language processing technology as a tool to better understand human language processing, which in turn, could help us build better natural language processing systems.

Appendix A

T-Maze Validation Experiment

sample.js

```
//for G-maze

var shuffleSequence = seq("code", "setcounter", "welcome", "prolific_id", "intro-gram
    ", "intro-practice", followEachWith("sep", "practice"), "end-practice",
    followEachWith("sep",randomize(anyOf(startsWith("rel"),startsWith("and"),
    startsWith("adverb"), startsWith("filler")))), "topic","debriefing");



//for L-maze

//var shuffleSequence = seq("code", "setcounter", "welcome", "intro-lex", "intro-
    practice", followEachWith("sep", "practice"), "end-practice", followEachWith("sep
    ",randomize(anyOf(startsWith("rel"),startsWith("and"), startsWith("adverb"),
    startsWith("filler")))),  "explanation","instructions2", anyOf("questionnaire"),"
    topic","debriefing");



var showProgressBar =true;



var defaults = [

                "Question", {
```

```
                                as: ["yes", "no"],

        presentAsScale: false,

        presentHorizontally: false,

    },

];



//var code = Math.floor(Math.random()*100000000);

//var sendingResultsMessage = "The results are now being transferred.  Please wait.";

//var completionMessage = "Thank you for your participation.  The results were
    successfully transmitted.  Your participation code is: " + code.toString();

//var completionErrorMessage = "The transmission of the results failed.  Please
    contact online_experiment@mit.edu and retry the transmission again by clicking
    the link.  Your participation code is: " + code.toString();

var code = "1EB8F1E6"; //replace with prolific code

var completionMessage = "Thank you for your participation! Your completion code is: "
     + code + ". Please copy and enter this code when you return to Prolifc to
    demonstrate that you completed the experiment.";



var items = [

    //["code", "DashedSentence", {s:code.toString(), mode:"speeded acceptability",
        wordTime:1}],

        ["setcounter", "__SetCounter__", { }],

    ["welcome", "Message", {html:'<table width="100%"><tr><td valign="top" align="
        right">Department of Brain and Cognitive Sciences<br>Massachusetts Institute
        of Technology<br>77 Massachusetts Avenue<br>Cambridge, MA 02139-4307, USA</tr
        ></table>\
<h2>Thank you very much for your participation!</h2><p>This is part of a MIT
    scientific research project. Your decision to pariticpate in this study is
    voluntary. There is no way for us to identify you. The only information we will
```

```
have, in addition to your responses, is the time at which you completed the
survey. The results of the research may be presented at scientific meetings or
published in scientific journals. Clicking on the link below indicates that you
are at least 18 years of age and agree to participate in this survey voluntarily
.'}],

["prolific_id", "Form", {html: 'Please enter your Prolific ID in the box below:<
    br/><textarea name="prolific_id" rows="1" cols="50" autofocus="true"></
    textarea>'}],
["explanation", "Form", {html:'How was your experience doing this task? What did
    you think of its length?<br/><textarea name="explanation" rows="3" cols="50"
    autofocus="true"></textarea>'}],

    ["instructions2", "Message", {html:'Now please answer a couple of questions
        about your background.  In accordance with the ethics guidelines of the
        Massachusetts Institute of Technology, this information will be stored in
         anonymous form and it will be impossible to link it to you.'}],

["questionnaire", "Question", {q:"Are you a native speaker of English?"}],

//["questionnaire", "Question", {q:"In what country did you learn to speak
    English?:", as:["United States of America", "United Kingdom", "Canada", "
    Australia", "New Zealand","Other"]}]

// ["questionnaire", "Form", {html:'How old are you? <input type="text" name="age"
    size="2" maxlength="2" autofocus="true"/>'}],

//["questionnaire", "Question", {q:"Please select the highest level of education
    you have attained:", as:["Less than high school", "High school graduate", "
    Some college", "2-year college degree", "4-year college degree", "
    Professional degree", "Doctorate"]}],

["questionnaire", "Question", {q:"Are you a citizen of the United States?"}],

["questionnaire", "Question", {q:"Do you currently reside in the United States
    ?"}],

//["topic", "Form", {html:'Very briefly, what do you think this study is about?<
    br/><textarea name="topic" rows="3" cols="50" autofocus="true"></textarea
    >'}],

["debriefing", "Message", {html:'<p>Thank you.  You will receive the
    participation code on the next page.</p>\n\n<p>Purpose of this study (feel
    free to skip):  W e re  generally interested in how the human brain processes
    language. The present study is testing out a new method for studying what
```

```
        types of sentence constructions are easier or harder to read. Your data will
        help us to answer these questions.</p>'}],

        ["intro-lex", "Message", {html: "<p>For this experiment, please place your
            left index finger on the 'e' key and your right index finger on the 'i'
            key. </p><p>On each screen you will see two options: one will be a word
            and one will be on a non-word. Select the real word by pressing 'e' (left
            -hand) for the option on the left or pressing 'i' (right-hand) for the
            option on the right.</p><p>The words will make a sentence.</p>" }],

        ["intro-gram", "Message", {html: "<p>For this experiment, please place your
            left index finger on the 'e' key and your right index finger on the 'i'
            key.</p><p> You will read sentences word by word. On each screen you will
             see two options: one will be the next word in the sentence, and one will
             not. Select the word that continues the sentence by pressing 'e' (left-
            hand) for the word on the left or pressing 'i' (right-hand) for the word
            on the right.</p><p>Select the best word as quickly as you can, but
            without making too many errors.</p>"}],

        ["intro-practice", "Message", {html: "The following items are for practice."
            }],

        ["end-practice", "Message", {html: "End of practice. The experiment will
            begin next."}],

        ["sep", "MazeSeparator", {normalMessage: "Correct! Press any key to continue
            ", errorMessage: "Incorrect! Press any key to continue."}],

        ["done", "Message", {html: "All done!"}],

[["adverb_high", 72], "Maze", {s:"Kim will display the photos she took next month,
    but she won't show all of them.", a:"x-x-x all granted made you'd year white use
    knew him than easy left been get way."}],
[["adverb_low", 72], "Maze", {s:"Kim will display the photos she took last month, but
     she won't show all of them.", a:"x-x-x all granted made you'd year white use
    knew him than easy left been get way."}],
[["adverb_high", 71], "Maze", {s:"Bob will complete the project he started next month
    , but Fred won't finish his.", a:"x-x-x much towards was you've off enough need
    hear two pole fact passed go."}],
[["adverb_low", 71], "Maze", {s:"Bob will complete the project he started last month,
     but Fred won't finish his.", a:"x-x-x much towards was you've off enough need
    hear two pole fact passed go."}],
[["adverb_high", 70], "Maze", {s:"John hired the clerk he will promote last month,
    but he fired another employee.", a:"x-x-x races it's adapt are it's staying care
    fun work way Iran matter purposes."}],
```

[["adverb_low", 70], "Maze", {s:"John hired the clerk he will promote next month, but he fired another employee.", a:"x-x-x races it's adapt are it's staying care fun work way Iran matter purposes."}],
[["adverb_high", 68], "Maze", {s:"Cathy will burn the wood she gathered next week, but she will save some of it.", a:"x-x-x all bible made owned day parallel use feel one than these tried say way year."}],
[["adverb_low", 68], "Maze", {s:"Cathy will burn the wood she gathered last week, but she will save some of it.", a:"x-x-x all bible made owned day parallel use feel one than these tried say way year."}],
[["adverb_high", 67], "Maze", {s:"Mark will answer the email he got next week, but he doesn't know what to write.", a:"x-x-x want areas made Italy why ever use using them time example been up us ahead."}],
[["adverb_low", 67], "Maze", {s:"Mark will answer the email he got last week, but he doesn't know what to write.", a:"x-x-x want areas made Italy why ever use using them time example been up us ahead."}],
[["adverb_high", 64], "Maze", {s:"Mike watered the flower he will sell yesterday, but he forgot to water the bush.", a:"x-x-x Prussia are legend day it's older resources, them year entry has least see loans."}],
[["adverb_low", 64], "Maze", {s:"Mike watered the flower he will sell tomorrow, but he forgot to water the bush.", a:"x-x-x Prussia are legend day it's older resources, them year entry has least see loans."}],
[["adverb_high", 63], "Maze", {s:"Mary called the applicant she will interview yesterday, but there was no answer.", a:"x-x-x least been handmade may it's somebody learning, any year any got weeks."}],
[["adverb_low", 63], "Maze", {s:"Mary called the applicant she will interview tomorrow, but there was no answer.", a:"x-x-x least been handmade may it's somebody learning, any year any got weeks."}],
[["adverb_high", 62], "Maze", {s:"James will fix the car he drove tomorrow, but he will need some help.", a:"x-x-x want net may I'd way god's resources, them year if I've say come."}],
[["adverb_low", 62], "Maze", {s:"James will fix the car he drove yesterday, but he will need some help.", a:"x-x-x want net may I'd way god's resources, them year if I've say come."}],
[["adverb_high", 61], "Maze", {s:"Linda will wear the sweater she washed tomorrow, but she won't wear her skirt.", a:"x-x-x all goals made vendors man roster address, them year fact glad been risky."}],
[["adverb_low", 61], "Maze", {s:"Linda will wear the sweater she washed yesterday, but she won't wear her skirt.", a:"x-x-x all goals made vendors man roster address, them year fact glad been risky."}],
[["adverb_low", 60], "Maze", {s:"Amy will visit the man she worked with last month, but she is nervous about it.", a:"x-x-x all we've made I've day you'll any care eyes had than were speaks need know."}],
[["adverb_high", 60], "Maze", {s:"Amy will visit the man she worked with next month, but she is nervous about it.", a:"x-x-x all we've made I've day you'll any care eyes had than were speaks need know."}],

```
[["adverb_low", 59], "Maze", {s:"Paul will marry the woman he just met last month,
    but the wedding will be small.", a:"x-x-x want one's need ready year it's gold
    need knew how than learned two him came."}],
[["adverb_high", 59], "Maze", {s:"Paul will marry the woman he just met next month,
    but the wedding will be small.", a:"x-x-x want one's need ready year it's gold
    need knew how than learned two him came."}],
[["adverb_low", 58], "Maze", {s:"Dan wrote the speech he will deliver next month, but
     he hasn't practiced it yet.", a:"x-x-x goal been truly two it's Clinton use dead
     him day entry interpret made wait."}],
[["adverb_high", 58], "Maze", {s:"Dan wrote the speech he will deliver last month,
    but he hasn't practiced it yet.", a:"x-x-x goal been truly two it's Clinton use
    dead him day entry interpret made wait."}],
[["adverb_low", 57], "Maze", {s:"Jeff planned the party he will hold next month, but
    he hasn't sent invitations.", a:"x-x-x climate been given new it's price sure
    lead them way entry stand persecuted."}],
[["adverb_high", 57], "Maze", {s:"Jeff planned the party he will hold last month, but
     he hasn't sent invitations.", a:"x-x-x climate been given new it's price sure
    lead them way entry stand persecuted."}],
[["adverb_high", 65], "Maze", {s:"Susan bought the wine she will drink last week, but
     she didn't buy any cheese.", a:"x-x-x nation been alive down it's whom need able
     one year course held need motion."}],
[["adverb_low", 65], "Maze", {s:"Susan bought the wine she will drink next week, but
    she didn't buy any cheese.", a:"x-x-x nation been alive down it's whom need able
    one year course held need motion."}],
[["adverb_low", 56], "Maze", {s:"Lisa will change the plans she made last week, but
    she won't cancel any of them.", a:"x-x-x much least made we've love our love feel
     one than easy slept made no know."}],
[["adverb_high", 56], "Maze", {s:"Lisa will change the plans she made next week, but
    she won't cancel any of them.", a:"x-x-x much least made we've love our love feel
     one than easy slept made no know."}],
[["adverb_low", 55], "Maze", {s:"Tom will plant the tree he bought last week, but he
    isn't sure where to put it.", a:"x-x-x year we'll been Japan two camera use feel
    her than fact use same no men with."}],
[["adverb_high", 55], "Maze", {s:"Tom will plant the tree he bought next week, but he
     isn't sure where to put it.", a:"x-x-x year we'll been Japan two camera use feel
     her than fact use same no men with."}],
[["adverb_low", 54], "Maze", {s:"Joseph brewed the beer he will serve next week, but
    it is not very tasty.", a:"x-x-x gusts been links day it's weird care less them
    way know did take flaws."}],
[["adverb_high", 54], "Maze", {s:"Joseph brewed the beer he will serve last week, but
     it is not very tasty.", a:"x-x-x gusts been links day it's weird care less them
    way know did take flaws."}],
[["adverb_low", 53], "Maze", {s:"Jane prepared the lecture she will give next week,
    but still needs to review it.", a:"x-x-x governor been justify man it's doing got
     feel him place fact was entire made."}],
```

[["adverb_high", 53], "Maze", {s:"Jane prepared the lecture she will give last week, but still needs to review it.", a:"x-x-x governor been justify man it's doing got feel him place fact was entire made."}],
[["adverb_low", 52], "Maze", {s:"Sue insulted the candidate she will debate tomorrow, but she wishes she hadn't.", a:"x-x-x shootout all honestly us it's forgot pressure, year year man's need marry."}],
[["adverb_high", 52], "Maze", {s:"Sue insulted the candidate she will debate yesterday, but she wishes she hadn't.", a:"x-x-x shootout all honestly us it's forgot pressure, year year man's need marry."}],
[["adverb_low", 51], "Maze", {s:"David caught the fish he will cook tomorrow, but it is not his favorite kind.", a:"x-x-x nation made shown need it's spoke approach, them year into did take happens went."}],
[["adverb_high", 51], "Maze", {s:"David caught the fish he will cook yesterday, but it is not his favorite kind.", a:"x-x-x nation made shown need it's spoke approach, them year into did take happens went."}],
[["adverb_low", 50], "Maze", {s:"Robert will meet the friend he phoned yesterday, but he doesn't want to.", a:"x-x-x want she's was thank day Osama pressure, had than matter been year."}],
[["adverb_high", 50], "Maze", {s:"Robert will meet the friend he phoned tomorrow, but he doesn't want to.", a:"x-x-x want she's was thank day Osama pressure, had than matter been year."}],
[["adverb_low", 49], "Maze", {s:"Anne will serve the apples she picked yesterday, but she won't serve the plums.", a:"x-x-x want drug was Munich life chair pressure, him year fact lived been erupt."}],
[["adverb_high", 49], "Maze", {s:"Anne will serve the apples she picked tomorrow, but she won't serve the plums.", a:"x-x-x want drug was Munich life chair pressure, him year fact lived been erupt."}],
[["and_no_comma", 48], "Maze", {s:"The witness identified the man and his wife ran away from the police station.", a:"x-x-x Zealand depression year I've need down she's arms I've what year thanks allowed."}],
[["and_comma", 48], "Maze", {s:"The witness identified the man, and his wife ran away from the police station.", a:"x-x-x Zealand depression year I've, need down she's arms I've what year thanks allowed."}],
[["and_no_comma", 47], "Maze", {s:"Jenny talked to the reporter and the photographer took pictures of the scene.", a:"x-x-x taxes may than dropping any time approaches post weekend time but luck."}],
[["and_comma", 47], "Maze", {s:"Jenny talked to the reporter, and the photographer took pictures of the scene.", a:"x-x-x taxes may than dropping, any time approaches post weekend time but luck."}],
[["and_no_comma", 46], "Maze", {s:"The robber shot the jeweler and the salesman reported the crime to the police.", a:"x-x-x slows break are widths our than handmade episode been we'll day it's words."}],
[["and_comma", 46], "Maze", {s:"The robber shot the jeweler, and the salesman reported the crime to the police.", a:"x-x-x slows break are widths, our than handmade episode been we'll day it's words."}],

[["and_no_comma", 45], "Maze", {s:"The journalist criticized Nick and Sam called the newspaper to complain.", a:"x-x-x threatened violations it'll are arts course than somewhat way fastest."}],
[["and_comma", 45], "Maze", {s:"The journalist criticized Nick, and Sam called the newspaper to complain.", a:"x-x-x threatened violations it'll, are arts course than somewhat way fastest."}],
[["and_no_comma", 44], "Maze", {s:"The actress yelled at the cameraman and the director hurried out of the room.", a:"x-x-x retired Congo year now sclerosis us time haven't burnout year make more mind."}],
[["and_comma", 44], "Maze", {s:"The actress yelled at the cameraman, and the director hurried out of the room.", a:"x-x-x retired Congo year now sclerosis, us time haven't burnout year make more mind."}],
[["adverb_high", 66], "Maze", {s:"Chris cleaned the bookcase he will sell last week, but it is still very dusty.", a:"x-x-x salmon been colonize day it's feet care free any year know does help rant."}],
[["adverb_low", 66], "Maze", {s:"Chris cleaned the bookcase he will sell next week, but it is still very dusty.", a:"x-x-x salmon been colonize day it's feet care free any year know does help rant."}],
[["and_no_comma", 43], "Maze", {s:"Sam hired the plumber and the carpenter ordered the materials for the house.", a:"x-x-x towns been Assange them than healthier Angeles are cultural day out while."}],
[["and_comma", 43], "Maze", {s:"Sam hired the plumber, and the carpenter ordered the materials for the house.", a:"x-x-x towns been Assange, them than healthier Angeles are cultural day out while."}],
[["and_no_comma", 42], "Maze", {s:"The police arrested the burglar and his brother phoned a lawyer for help.", a:"x-x-x comes kingdom been lighten year make aren't timers made smile time night."}],
[["and_comma", 42], "Maze", {s:"The police arrested the burglar, and his brother phoned a lawyer for help.", a:"x-x-x comes kingdom been lighten, year make aren't timers made smile time night."}],
[["and_no_comma", 40], "Maze", {s:"The customer complained about the waiter and the chef gave him a free dessert.", a:"x-x-x careful geographic know year wisely need day reads month been than I've spells."}],
[["and_comma", 40], "Maze", {s:"The customer complained about the waiter, and the chef gave him a free dessert.", a:"x-x-x careful geographic know year wisely, need day reads month been than I've spells."}],
[["and_no_comma", 39], "Maze", {s:"The woman could not find Bill and his girlfriend became nervous and upset.", a:"x-x-x comes year year told dead how way collected choice rarely your Wales."}],
[["and_comma", 39], "Maze", {s:"The woman could not find Bill, and his girlfriend became nervous and upset.", a:"x-x-x comes year year told dead, how way collected choice rarely your Wales."}],
[["and_no_comma", 38], "Maze", {s:"Bobby yelled at the teacher and the principal asked his parents for a meeting.", a:"x-x-x Paulo been year meaning us need purposes easy even you've your than exactly."}],

[["and_comma", 38], "Maze", {s:"Bobby yelled at the teacher, and the principal asked his parents for a meeting.", a:"x-x-x Paulo been year meaning, us need purposes easy even you've your than exactly."}],
[["and_no_comma", 37], "Maze", {s:"The teacher praised the girl and her family was proud of her good grades.", a:"x-x-x anyway cowboys it's I'd year use that's it's click into two it's dated."}],
[["and_comma", 37], "Maze", {s:"The teacher praised the girl, and her family was proud of her good grades.", a:"x-x-x anyway cowboys it's I'd, year use that's it's click into two it's dated."}],
[["and_comma", 36], "Maze", {s:"The crowd cheered for the model, and the designer took a bow after the show.", a:"x-x-x loan Canucks been way alone, is than nowhere games been ate part year left."}],
[["and_no_comma", 36], "Maze", {s:"The crowd cheered for the model and the designer took a bow after the show.", a:"x-x-x loan Canucks been way alone is than nowhere games been ate part year left."}],
[["and_comma", 35], "Maze", {s:"The juggler entertained the children, and their parents drank wine at the party.", a:"x-x-x bronzed conditioned or received, year day you've toes lies day now less."}],
[["and_no_comma", 35], "Maze", {s:"The juggler entertained the children and their parents drank wine at the party.", a:"x-x-x bronzed conditioned or received year day you've toes lies day now less."}],
[["and_comma", 34], "Maze", {s:"The woman dressed her baby, and her son got his clothes from the dresser.", a:"x-x-x comes Muslims year view, are than low team than Florida way it's Assange."}],
[["and_no_comma", 34], "Maze", {s:"The woman dressed her baby and her son got his clothes from the dresser.", a:"x-x-x comes Muslims year view are than low team than Florida way it's Assange."}],
[["and_comma", 32], "Maze", {s:"The producer replaced the actor, and the actress quit the movie after the fight.", a:"x-x-x retired decades year urban, what need waters yard been key day year ready."}],
[["and_no_comma", 32], "Maze", {s:"The producer replaced the actor and the actress quit the movie after the fight.", a:"x-x-x retired decades year urban what need waters yard been key day year ready."}],
[["and_comma", 31], "Maze", {s:"The audience applauded the guitarist, and the band cheered for him very loudly.", a:"x-x-x birthday Iranians make swallowed, day make doubt firm's been than say remake."}],
[["and_no_comma", 31], "Maze", {s:"The audience applauded the guitarist and the band cheered for him very loudly.", a:"x-x-x birthday Iranians make swallowed day make doubt firm's been than say remake."}],
[["and_comma", 30], "Maze", {s:"Jim listened to the pianist, and the singer watched the organist at the concert.", a:"x-x-x holidays may year onstage, any go he'll Angeles year issuers year than harder."}],
[["and_no_comma", 30], "Maze", {s:"Jim listened to the pianist and the singer watched the organist at the concert.", a:"x-x-x holidays may year onstage any go he'll Angeles year issuers year than harder."}],

[["and_comma", 29], "Maze", {s:"The ranger gave matches to the camper, and his friend made a fire by the tent.", a:"x-x-x Biden month academy this not Vettel, year year months team year due year but reign."}],
[["and_no_comma", 29], "Maze", {s:"The ranger gave matches to the camper and his friend made a fire by the tent.", a:"x-x-x Biden month academy this not Vettel year year months team year due year but reign."}],
[["and_comma", 28], "Maze", {s:"The little girl fed her pet cat, and her dog wanted a can of food, too.", a:"x-x-x least leave fee say Iraq ends, day over July games year than time ago make."}],
[["and_no_comma", 28], "Maze", {s:"The little girl fed her pet cat and her dog wanted a can of food, too.", a:"x-x-x least leave fee say Iraq ends day over July games year than time ago make."}],
[["and_comma", 27], "Maze", {s:"The tourist photographed the swimmer, and the runner got ready for the race.", a:"x-x-x delete schooling than Yorker, day day she'll day June know than knows."}],
[["and_no_comma", 27], "Maze", {s:"The tourist photographed the swimmer and the runner got ready for the race.", a:"x-x-x delete schooling than Yorker day day she'll day June know than knows."}],
[["and_comma", 26], "Maze", {s:"The swimmer disappointed her coach, and her mother tried to console her.", a:"x-x-x revise efficient use moved, need more longer rate year hardest last."}],
[["and_no_comma", 26], "Maze", {s:"The swimmer disappointed her coach and her mother tried to console her.", a:"x-x-x revise efficient use moved need more longer rate year hardest last."}],
[["and_comma", 25], "Maze", {s:"The nurse examined the mother, and the child played quietly in the corner.", a:"x-x-x cares advances been seems, our time I'd rights emerged way if awards."}],
[["and_no_comma", 25], "Maze", {s:"The nurse examined the mother and the child played quietly in the corner.", a:"x-x-x cares advances been seems our time I'd rights emerged way if awards."}],
[["relative_high", 24], "Maze", {s:"The niece of the butler who scolded herself for losing the key was very upset.", a:"x-x-x Modi more are jeans its lineups happens us click been led work day blow."}],
[["relative_low", 24], "Maze", {s:"The niece of the butler who scolded himself for losing the key was very upset.", a:"x-x-x Modi more are jeans its lineups happens us click been led work day blow."}],
[["relative_high", 23], "Maze", {s:"The aunt of the waiter who trained herself to cook wanted to own a restaurant.", a:"x-x-x bless more are mails way academy choice us keeps least may side may appearance."}],
[["relative_low", 23], "Maze", {s:"The aunt of the waiter who trained himself to cook wanted to own a restaurant.", a:"x-x-x bless more are mails way academy choice us keeps least may side may appearance."}],
[["relative_high", 22], "Maze", {s:"The sister of the boy who taught herself advanced mathematics was very smart.", a:"x-x-x shown more back cost see sector charge charges apologize way world seat."}],

```
[["relative_low", 22], "Maze", {s:"The sister of the boy who taught himself advanced
    mathematics was very smart.", a:"x-x-x shown more back cost see sector charge
    charges apologize way world seat."}],
[["and_comma", 33], "Maze", {s:"Diane hugged her boyfriend, and her friend felt
    uncomfortable watching them.", a:"x-x-x sexes been arrested, day time course tax
    collecting percent need."}],
[["and_no_comma", 33], "Maze", {s:"Diane hugged her boyfriend and her friend felt
    uncomfortable watching them.", a:"x-x-x sexes been arrested day time course tax
    collecting percent need."}],
[["relative_high", 21], "Maze", {s:"The daughter of the king who devoted herself to
    the kingdom was never emotional.", a:"x-x-x recently time back issue way comics
    unless work than noticed work I'm whenever."}],
[["relative_low", 21], "Maze", {s:"The daughter of the king who devoted himself to
    the kingdom was never emotional.", a:"x-x-x recently time back issue way comics
    unless work than noticed work I'm whenever."}],
[["relative_high", 20], "Maze", {s:"The niece of the fisherman who got herself a
    sailboat learned to sail.", a:"x-x-x Modi more are deficits way team unless see
    Incheon accept has exams."}],
[["relative_low", 20], "Maze", {s:"The niece of the fisherman who got himself a
    sailboat learned to sail.", a:"x-x-x Modi more are deficits way team unless see
    Incheon accept has exams."}],
[["relative_high", 19], "Maze", {s:"The grandma of the policeman who educated herself
     at night became a teacher.", a:"x-x-x Arabia day off premiums way Indiana moving
     us must issues it's airport."}],
[["relative_low", 19], "Maze", {s:"The grandma of the policeman who educated himself
    at night became a teacher.", a:"x-x-x Arabia day off premiums way Indiana moving
    us must issues it's airport."}],
[["relative_high", 18], "Maze", {s:"The uncle of the girl who prepared himself for
    the race was a great athlete.", a:"x-x-x tied day back needs up classic forget us
     than we've need it's end Arabia."}],
[["relative_low", 18], "Maze", {s:"The uncle of the girl who prepared herself for the
     race was a great athlete.", a:"x-x-x tied day back needs up classic forget us
    than we've need it's end Arabia."}],
[["relative_high", 17], "Maze", {s:"The nephew of the queen who praised himself all
    the time was very rude.", a:"x-x-x else's day back uses see cowboys posted work
    back I'm year day dawn."}],
[["relative_low", 17], "Maze", {s:"The nephew of the queen who praised herself all
    the time was very rude.", a:"x-x-x else's day back uses see cowboys posted work
    back I'm year day dawn."}],
[["relative_high", 16], "Maze", {s:"The grandfather of the policewoman who treated
    himself so badly was troubled.", a:"x-x-x Argentina than back granulated not
    Muslim unless year cats year leisure."}],
[["relative_low", 16], "Maze", {s:"The grandfather of the policewoman who treated
    herself so badly was troubled.", a:"x-x-x Argentina than back granulated not
    Muslim unless year cats year leisure."}],
```

```
[["relative_high", 15], "Maze", {s:"The nephew of the maid who cut himself with the
    knife called the doctor.", a:"x-x-x else's day back Biden see York unless way but
     Vegas office it's fully."}],
[["relative_low", 15], "Maze", {s:"The nephew of the maid who cut herself with the
    knife called the doctor.", a:"x-x-x else's day back Biden see York unless way but
     Vegas office it's fully."}],
[["relative_high", 13], "Maze", {s:"The son of the princess who scratched himself in
    public was awfully embarrassed.", a:"x-x-x goes time back purposes back society's
     subject work though day hoodie excellence."}],
[["relative_low", 13], "Maze", {s:"The son of the princess who scratched herself in
    public was awfully embarrassed.", a:"x-x-x goes time back purposes back society's
     subject work though day hoodie excellence."}],
[["relative_low", 12], "Maze", {s:"The sister of the prince who injured himself
    falling off a roof was still sad.", a:"x-x-x shown more back funds way Zealand
    expect artists our how brief our day uses."}],
[["relative_high", 12], "Maze", {s:"The sister of the prince who injured herself
    falling off a roof was still sad.", a:"x-x-x shown more back funds way Zealand
    expect artists our how brief our day uses."}],
[["relative_low", 11], "Maze", {s:"The daughter of the actor who hated himself for
    failing always seemed unhappy.", a:"x-x-x recently time back abuse way cave
    unless know Madrid night enter travels."}],
[["relative_high", 11], "Maze", {s:"The daughter of the actor who hated herself for
    failing always seemed unhappy.", a:"x-x-x recently time back abuse way cave
    unless know Madrid night enter travels."}],
[["relative_low", 9], "Maze", {s:"The daughter of the man who complimented himself in
     public was beautiful.", a:"x-x-x recently time back use way personalised choice
    work though new products."}],
[["relative_high", 9], "Maze", {s:"The daughter of the man who complimented herself
    in public was beautiful.", a:"x-x-x recently time back use way personalised
    choice work though new products."}],
[["relative_low", 8], "Maze", {s:"The aunt of the schoolboy who hurt himself was
    concerned about the injury.", a:"x-x-x bless more are launchers its unit leading
    them decisions him know liked."}],
[["relative_high", 8], "Maze", {s:"The aunt of the schoolboy who hurt herself was
    concerned about the injury.", a:"x-x-x bless more are launchers its unit leading
    them decisions him know liked."}],
[["relative_low", 7], "Maze", {s:"The grandma of the fireman who criticized himself
    far too often was anxious.", a:"x-x-x Arabia day off crammed been adaptation
    leading law team law good publish."}],
[["relative_high", 7], "Maze", {s:"The grandma of the fireman who criticized herself
    far too often was anxious.", a:"x-x-x Arabia day off crammed been adaptation
    leading law team law good publish."}],
[["relative_low", 6], "Maze", {s:"The grandfather of the woman who killed herself
    last summer had been to prison.", a:"x-x-x Argentina than back needs way policy
    income love longer say who know apart."}],
```

[["relative_high", 6], "Maze", {s:"The grandfather of the woman who killed himself last summer had been to prison.", a:"x-x-x Argentina than back needs way policy income love longer say who know apart."}],
[["relative_low", 5], "Maze", {s:"The brother of the schoolgirl who burned herself was usually very careful.", a:"x-x-x percent day back touchscreen its fleet expect day energy state scheme."}],
[["relative_high", 5], "Maze", {s:"The brother of the schoolgirl who burned himself was usually very careful.", a:"x-x-x percent day back touchscreen its fleet expect day energy state scheme."}],
[["relative_low", 4], "Maze", {s:"The son of the lady who politely introduced herself was popular at the party.", a:"x-x-x goes time back costs up Estonia somewhere unless year schools work than god."}],
[["relative_high", 4], "Maze", {s:"The son of the lady who politely introduced himself was popular at the party.", a:"x-x-x goes time back costs up Estonia somewhere unless year schools work than god."}],
[["relative_low", 1], "Maze", {s:"The son of the actress who shot herself on the set was under investigation.", a:"x-x-x goes time back courses way clear expect need but less back week ultimately."}],
[["relative_high", 1], "Maze", {s:"The son of the actress who shot himself on the set was under investigation.", a:"x-x-x goes time back courses way clear expect need but less back week ultimately."}],
[["filler", 134], "Maze", {s:"The children of the rich man were spoiled, but they were charming and handsome.", a:"x-x-x actually than back costs way been broadly, day way up mentions need abortion."}],
[["filler", 133], "Maze", {s:"Yesterday the wife of the politician discussed health care with old people.", a:"x-x-x want ready time why significance languages north least him yet states."}],
[["filler", 132], "Maze", {s:"The boyfriend of the model was killed in an accident while skiing last week.", a:"x-x-x instance than are value off you'll us now governor need capita say less."}],
[["filler", 131], "Maze", {s:"The cute girl who was on the cover of the magazine became a famous doctor.", a:"x-x-x loan York than off see than size back than entered county year estate worse."}],
[["filler", 130], "Maze", {s:"The writer of the novels thought himself to be a genius, but he wasn't.", a:"x-x-x hasn't year over Potter country example when it's it's parks, work way across."}],
[["filler", 129], "Maze", {s:"Todd wanted to be a barber, but his shaky hands prevented him from becoming one.", a:"x-x-x least may new time tapes, us way liars upon Harrison been you sources over."}],
[["filler", 128], "Maze", {s:"Julie dated Adam and Andy, but she married Jeff in Las Vegas last month.", a:"x-x-x gulf finds are tons, me way effect argue year worn bye can't died."}],
[["filler", 127], "Maze", {s:"The student who was running out of money gave himself a haircut last week.", a:"x-x-x expect time up points year year four heart you'll than hackers those away."}],

[["filler", 125], "Maze", {s:"The gardener was happy with the flowers, but the owner of the house was not.", a:"x-x-x mindful year list more then faster, make know bet day year while off did."}],
[["filler", 126], "Maze", {s:"Judy cooked in the kitchen, and Richard barbecued outside in the yard.", a:"x-x-x Arabia been way measure, day request headbands played day than sight."}],
[["filler", 124], "Maze", {s:"Sara and her mother had steak, potatoes, and green beans for dinner last night.", a:"x-x-x way over tried them lame, predict, year knows hates year stated years find."}],
[["filler", 123], "Maze", {s:"Margo will open bakeries in Chicago and New York after getting a loan next year.", a:"x-x-x him level fluidity him losing way may least state against day actor four those."}],
[["filler", 122], "Maze", {s:"Cindy treated herself to a vacation in China and several other Asian countries.", a:"x-x-x ensure suggest us than Atlantic take stand way minutes can't actor happened."}],
[["filler", 121], "Maze", {s:"The nurse took care of the old lady who could not take care of herself.", a:"x-x-x cares post war year year I've feels good find your I'm found way seemed."}],
[["filler", 120], "Maze", {s:"The pilot will fly a new airplane to Europe and back beginning next week.", a:"x-x-x swear than grew been way sponsors year female up day pressure might care."}],
[["filler", 119], "Maze", {s:"Megan got angry with her boss and sued him for discrimination last month.", a:"x-x-x care enemy know year alive we Gaza been know laboratory life died."}],
[["filler", 118], "Maze", {s:"The father of the pretty girl cleaned up the house all by himself last week.", a:"x-x-x you'll than back needs media salmon new way he's day make couple end music."}],
[["adverb_high", 69], "Maze", {s:"Jim painted the picture he will display last month, but he isn't happy with it.", a:"x-x-x breast any you've take it's passing love wants them time fact check good see."}],
[["adverb_low", 69], "Maze", {s:"Jim painted the picture he will display next month, but he isn't happy with it.", a:"x-x-x breast any you've take it's passing love wants them time fact check good see."}],
[["filler", 117], "Maze", {s:"Bill will quit his job next month to take care of his kids and step-children.", a:"x-x-x want he'd made lost men club into two least year year I'd did description."}],
[["filler", 116], "Maze", {s:"The thief noticed the boy and the girl watching him from the upstairs window.", a:"x-x-x Biden advance are knew day day south material I've it's it's blaming tells."}],
[["filler", 115], "Maze", {s:"The boy who was wearing a blue cap went missing on his way home last week.", a:"x-x-x goes best new budget day areas rid white created us than need few game god."}],
[["filler", 114], "Maze", {s:"Hugh took too long to propose, and his girlfriend decided to marry someone else.", a:"x-x-x least been city day abused, him way connected weight been Iraq example turn."}],

```
[["relative_low", 10], "Maze", {s:"The sister of the salesman who made a fool of
    himself at work was very angry.", a:"x-x-x shown more back mileage see team now
    relax need choice work than them day route."}],
[["relative_high", 10], "Maze", {s:"The sister of the salesman who made a fool of
    herself at work was very angry.", a:"x-x-x shown more back mileage see team now
    relax need choice work than them day route."}],
[["filler", 113], "Maze", {s:"The old man amused himself by gossiping with his
    children and friends.", a:"x-x-x done ever ounces current are journeyed no it's
    happened good however."}],
[["filler", 111], "Maze", {s:"The infant was able to stand by himself for the first
    time last week.", a:"x-x-x else's time order us areas why current year it's it's
    being days yes."}],
[["practice", 108], "Maze", {s:"The semester will start next week, but the students
    and teachers are not ready.", a:"x-x-x anyways see least group music, new than
    example way dropped many get fire."}],
[["practice", 107], "Maze", {s:"The mother of the prisoner sent him packages that
    contained cookies and novels.", a:"x-x-x you'll than back advised sense even
    Lebanon get awareness arrives any Marcus."}],
[["practice", 105], "Maze", {s:"The reporter had dinner yesterday with the baseball
    player who Kevin admired.", a:"x-x-x would've year senate challenge would it's
    letting stand them frame hackers."}],
[["filler", 112], "Maze", {s:"Jane and John studied math and history yesterday, but
    they failed the exams.", a:"x-x-x much hours Zealand asks day months secretary,
    him way flight been adapt."}],
[["practice", 104], "Maze", {s:"The therapist set up a meeting with the upset woman
    and her husband yesterday.", a:"x-x-x Colombia post been way finally good year
    wars York year but happens marriage."}],
[["practice", 103], "Maze", {s:"Maya played with the blocks and the balls, but she
    soon got bored with them.", a:"x-x-x guess been year refer year year debut, year
    year due sure grows know well."}],
[["practice", 102], "Maze", {s:"The patient who the doctor treated became better
    after only a few treatments.", a:"x-x-x Angeles than up beach Angeles choice that
    's need year did end separately."}],
[["practice", 101], "Maze", {s:"The husband of the beautiful woman bought her roses
    and candy for her birthday.", a:"x-x-x expect day back pressure pay coast say
    urged been Saudi why it's somebody."}],
[["relative_low", 3], "Maze", {s:"The uncle of the waitress who hurt herself was
    shocked by the accident.", a:"x-x-x tied day back bothers into wide waiting us
    counts your year Angeles."}],
[["relative_high", 3], "Maze", {s:"The uncle of the waitress who hurt himself was
    shocked by the accident.", a:"x-x-x tied day back bothers into wide waiting us
    counts your year Angeles."}],
[["and_comma", 41], "Maze", {s:"The knight greeted the king, and the queen waved to
    her people at the feast.", a:"x-x-x Delhi peppers are July, two time we'll ethic
    been down music we not flu."}],
```

```
[["and_no_comma", 41], "Maze", {s:"The knight greeted the king and the queen waved to
     her people at the feast.", a:"x-x-x Delhi peppers are July two time we'll ethic
    been down music we not flu."}],
[["relative_high", 14], "Maze", {s:"The brother of the ballerina who found himself in
     a lot of trouble phoned home.", a:"x-x-x percent day back scrubbing use team
    unless why than ago may Charles tights world."}],
[["relative_low", 14], "Maze", {s:"The brother of the ballerina who found herself in
    a lot of trouble phoned home.", a:"x-x-x percent day back scrubbing use team
    unless why than ago may Charles tights world."}],
[["practice", 106], "Maze", {s:"The visitors at the zoo watched the zookeeper who the
     monkeys and apes teased.", a:"x-x-x Zealand than how wont victory see Feingold
    year than weights day lows melon."}],
[["relative_low", 2], "Maze", {s:"The brother of the bride who embarrassed herself at
     the wedding felt ashamed.", a:"x-x-x percent day back wont back interactive
    recent work than learned code tablet."}],
[["relative_high", 2], "Maze", {s:"The brother of the bride who embarrassed himself
    at the wedding felt ashamed.", a:"x-x-x percent day back wont back interactive
    recent work than learned code tablet."}],
];
```

References

Bartek, B., Lewis, R. L., Vasishth, S., and Smith, M. R. (2011). In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5):1178.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28.

Cohen, J. (1992). Statistical power analysis. *Current directions in psychological science*, 1(3):98–101.

Dale, R. (2022). $nlp: How to spend a billion dollars. *Natural Language Engineering*, 28(1):125–136.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Enochson, K. and Culbertson, J. (2015). Collecting psycholinguistic response time data using amazon mechanical turk. *PLOS ONE*, 10(3):1–17.

Floridi, L. and Chiriatti, M. (2020). Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694.

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

Frazier, L. (1987). Syntactic processing: evidence from dutch. *Natural Language & Linguistic Theory*, 5(4):519–559.

Frazier, L. and Clifton, C. (1997). Construal: Overview, motivation, and some new evidence. *Journal of Psycholinguistic Research*, 26(3):277–295.

Gibson, E. and Pearlmutter, N. J. (1998). Constraints on sentence comprehension. *Trends in cognitive sciences*, 2(7):262–268.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., and Baroni, M. (2018).

Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spacy: Industrial-strength natural language processing in python.

Jones, L. V. (1952). Test of hypotheses: one-sided vs. two-sided alternatives. *Psychological Bulletin*, 49(1):43.

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., and Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Jurafsky, D. and Martin, J. H. (2018). Speech and language processing (draft). *preparation [cited 2020 June 1] Available from: https://web. stanford. edu/˜ jurafsky/slp3*.

Marcus, M., Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The penn treebank: Annotating predicate argument structure. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Marian, V., Bartolotti, J., Chabal, S., and Shook, A. (2012). Clearpond: Cross-linguistic easy-access resource for phonological and orthographic neighborhood densities.

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133.

McGuffie, K. and Newhouse, A. (2020). The radicalization risks of gpt-3 and advanced neural language models. *arXiv preprint arXiv:2009.06807*.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Mitchell, D. (1984). *An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading*. Lawrence Earlbaum Associates.

Neuliep, J. W. (1990). Editorial bias against replication research. *Journal of Social Behavior and Personality*, 5(4):85.

Palan, S. and Schitter, C. (2018). Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419.

Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., and Dean, J. (2021). Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*.

Peer, E., Rothschild, D., Gordon, A., Evernden, Z., and Damer, E. (2021). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods*, pages 1–20.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving language understanding by generative pre-training.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Salazar, J., Liang, D., Nguyen, T. Q., and Kirchhoff, K. (2019). Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Scheible, R., Thomczyk, F., Tippmann, P., Jaravine, V., and Boeker, M. (2020). Gottbert: a pure german language model. *arXiv preprint arXiv:2012.02110*.

Schweter, S. (2020). German gpt-2 model.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., and Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Smith, R. D. (2012). Distinct word length frequencies: distributions and symbol entropies. *arXiv preprint arXiv:1207.2334*.

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). Luminosoinsight/-wordfreq: v2.2.

Stevenson, A. (2010). *Oxford dictionary of English*. Oxford University Press, USA.

Tanenhaus, M. K. and Trueswell, J. C. (1995). Sentence comprehension.

Thunström, A. O. and Steingrimsson, S. (2022). Can gpt-3 write an academic paper on itself, with minimal human input?

Tincani, M. and Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science*, 42(1):59–75.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.

Wagenmakers, E.-J., Lee, M., Lodewyckx, T., and Iverson, G. J. (2008). Bayesian versus frequentist inference. In *Bayesian evaluation of informative hypotheses*, pages 181–207. Springer.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

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

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P.,

Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zehr, J. and Schwarz, F. (2018). Penncontroller for internet based experiments (ibex). *DOI: https://doi. org/10.17605/OSF. IO/MD832*.

Zipf, G. K. (1949). Human behavior and the principle of least effort.