LEXICAL ACCESS ACROSS CONTEXTS AND WORD-CLASS

EFFECTS OF SEMANTIC CONTEXT AND WORD-CLASS ON SUCCESSFUL LEXICAL ACCESS

By JULIE BANNON, B.Sc. (Hons), M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

McMaster University DOCTOR OF PHILOSOPHY (2023)

Hamilton, Ontario (Psychology, Neuroscience and Behaviour)

TITLE: Effects of Semantic Context and Word-Class on Successful Lexical Access

AUTHOR: Julie Bannon, B.Sc. (University of Toronto), M.Sc. (McMaster University)

SUPERVISORS: Dr. Tamar Gollan, Dr. Victor Ferreira, Dr. Bruce Milliken, Dr. Judith Shedden, Dr. Elisabet Service

NUMBER OF PAGES: xiii, 128

LAY ABSTRACT

Language plays a key role in our everyday lives, including in social interactions, academic success, and overall daily functioning. The process of producing and understanding language is deceptively easy for the average person, but there are significant outstanding questions about how linguistic processes operate. The retrieval of individual words in particular has been the subject of decades of investigation. The goal of the present thesis is to investigate how we retrieve words when we speak, or the process of *lexical access*, by eliciting production of words across various contexts. The studies reported here demonstrate the effects of semantic context on lexical access, as well as how this process differs for words that convey syntactic versus meaningful content (i.e., words that differ in *lexical class*). Our findings build on theories of lexical access by demonstrating unique effects of the roles of semantic contexts and lexical class on word retrieval.

ABSTRACT

Language production is ubiquitous in everyday life. A critical component of language production is the retrieval of individual words. In this thesis, we investigated the process of lexical access across six experiments that required participants to produce words in different contexts. First, we examined whether semantic relationships between proper names lead to competition during lexical access. Participants were asked to name celebrity pictures after either reading a famous or non-famous prime name or classifying a prime name as belonging to a famous or non-famous person. Results revealed that successful name retrievals decreased with increasing trial number. Within individual trials, tip-of-the-tongue states increased only after the classification of famous prime names. These findings indicate that the effects of competition from related proper names vary based on the particular semantic context in which they are retrieved. Next, we examined how the broader semantic context of sentences affects access to object names. It is widely accepted that highly constraining contexts can facilitate lexical access through predictive processing. We examined whether prediction during language processing still confers a benefit in situations where predictions were either almost correct or completely incorrect. In three experiments that investigated both language production and comprehension, we found a clear cost to incorrect predictions which we hypothesize may be used as an error signal in language learning to fine tune the language system. Finally, we investigated function word production using a task that required individuals to read aloud short paragraphs that contained errors on function words under distracting versus silent conditions. We found that background speech did not affect the likelihood that speakers would spontaneously correct the errors, but did increase non-target function word substitution errors. Overall, these studies support a framework in which lexical access is influenced by both word-class and semantic context at the point of retrieval.

ACKNOWLEDGEMENTS

I am infinitely grateful for all the wonderful mentors, scientists, and friends who supported and encouraged me throughout this journey. To Dr. Tamar Gollan and Dr. Victor Ferreira, I cannot thank you enough for taking a chance supervising a student on the other side of the continent in the middle of a pandemic. You picked me up during a bump in the road and provided so much support as I rebuilt my graduate research. Thank you for welcoming me into your lab and for guiding me through my research. You have both been so positive throughout and restored my faith in the academic process.

I would also like to thank my committee members Dr. Bruce Milliken and Dr. Judith Shedden. You both helped me navigate the challenges of graduate school and were always there to give advice and help point me in the right direction. To Dr. Lili Service, who has been on my committee since my master's thesis, thank you for sticking with me until the end and providing valuable feedback and insight throughout the way. To Dr. Karin Humphreys, thank you for being endlessly supportive of me and my research. Even when things were tough, you always believed in me and provided guidance, even from the sidelines. I would also like to give a special thank you to all the members of the Language Production Lab. You all welcomed me into the lab and provided me with so much valuable feedback and support. It was wonderful to be accepted into your lab family. Finally, thank you to the numerous undergraduates who participated in these studies, without whom this research would not have been possible.

This section would be remiss without a heartfelt thank you to my friends and family. To my parents, thank you for always believing in me and encouraging me to pursue my academic goals. To my friends Marissa, Zane, and Max, after decades of friendship you continue to cheer me on and be there for me through all the ups and downs. Thank you for always planning fun

nights out and helping me take breaks from work (and paying for the occasional meal when graduate funding was tight!). I also want to give a special thanks to Raheleh. You have been a wonderful mentor and friend since I joined the PAL lab, and I'm so grateful for all the research talks, writing retreats, dinner parties, and life chats. And of course, to Stefania, my CogSci companion, thank you for the venting sessions, the beers, and the support.

Finally, to the best human, John Fast, thank you for being by my side throughout my PhD. You have always been there to celebrate my successes and to encourage me to keep going when the challenges seemed insurmountable. Thank you for the numerous camping trips, hobby nights, pizza Fridays, hallouma-versaries, record nights (RIP turntable), and so many other things that helped keep me sane. I can't imagine going through this process without you by my side. I'm so happy I get to see you pursue your own academic goals, and I'm so excited for the future adventures we'll share together. And I of course need to give a shout out to Mr. Norman Chips and his collie-gue Zarya, thank you for being the best work-from-home companions and always being there for a snuggle when I needed a writing break!

TABLE OF CONTENTS

LIST OF FIGURES

Chapter 2

Chapter 3

Chapter 4

LIST OF TABLES

Chapter 2

Chapter 3

Chapter 4

DECLARATION OF ACADEMIC ACHIEVEMENT

The General Introduction (Chapter 1) and General Discussion (Chapter 5) were drafted by Julie Bannon, with edits from Tamar Gollan and Victor Ferreira.

Chapter 2 has been published in *Memory and Cognition*. The experiments were conceptualized and designed by Julie Bannon, Tamar Gollan, and Victor Ferreira. Julie Bannon programmed the experiments and collected the data. Data were analyzed by Julie Bannon and Alena Stasenko. The manuscript was drafted by Julie Bannon, with main edits from Tamar Gollan and Victor Ferreira, and additional edits from Alena Stasenko.

Chapter 3 is under review at the *Journal of Experimental Psychology: Learning, Memory, and Cognition*. The experiments were conceptualized and designed by Julie Bannon, Tamar Gollan, and Victor Ferreira. Julie Bannon programmed the experiments, and collected and analyzed the data. The manuscript was drafted by Julie Bannon, with main edits from Tamar Gollan and Victor Ferreira.

Chapter 4 was conceptualized and designed by Julie Bannon, Tamar Gollan, and Victor Ferreira. Julie Bannon programmed the experiments, and collected and analyzed the data. The chapter was drafted by Julie Bannon, with main edits from Tamar Gollan and Victor Ferreira.

CHAPTER 1: GENERAL INTRODUCTION

Introduction

Language production is a complex task that requires rapid integration of syntactic and semantic information and processing to produce meaningful speech. The cognitive complexity of producing speech stands in stark contrast to the subjective ease of everyday language production. Speakers are typically able to produce speech at a rate of 2-3 words per second, only rarely producing errors. Our ability to retrieve and produce words in rapid succession is remarkable, especially given that the mental lexicon contains tens of thousands of words that vary based on a range of factors such as meaning and part-of-speech. Successful lexical selection requires both consideration of the most appropriate term in a given context, and selection of the correct term from a range of alternatives. For example, we select words to convey a particular idea, to describe a particular situation, or to identify a particular object or person. When difficulties arise during the course of lexical selection, it can result in a failure to convey an intended message, such as telling your friend to meet you at the mall, when you really wanted to meet at the park. Such failures in word retrieval are not limited to any particular type of word or linguistic context. The joint influences of lexical features and linguistic context therefore present an important area of investigation to further our understanding of lexical selection during language production.

When it comes to producing an intended word, the first step is to select a concept from the mental lexicon that can then be mapped onto a lemma containing the grammatical features of the to-be-produced word (Levelt et al., 1999). The mental lexicon is often conceived of as a network of lexical concepts that are connected based on semantic relationships (Collins & Loftus, 1975). During the initial step of lexical access, concepts within the mental lexicon receive activation that can spread through the network to close semantic competitors, and the

concept that receives the highest level of activation is selected and then mapped onto the corresponding lemma. Critically, spreading activation within the mental lexicon implies that multiple related candidates are activated during lexical selection. A key debate in the psycholinguistic literature concerns how the activation of close semantic competitors influences the process of lexical access. In single word production, evidence from picture-word interference (PWI) studies suggests the presence of a semantically related word results in slower speech onset times in the production of a target word (e.g., Aristei & Rahman, 2013; Hantsch et al., 2005; Rose et al., 2019, but see Gauvin et al., 2018; Mahon et al., 2007). Similarly, semantic interference can accumulate over successive productions of semantically related words (Damian & Als, 2005; Oppenheim et al., 2010; Schnur, 2014). These findings have led researchers to propose that lexical selection is a competitive process, whereby the presence of semantically related words can impede access to a target word.

Despite significant evidence that lexical concepts compete for selection during lexical access, it is less clear whether the effects of competition are consistent across different classes of words. Difficulties in word retrieval vary based on other linguistic factors, such as lexical frequency or word class. Certain types of words are particularly susceptible to word retrieval failures, such as proper names which are typically lower in frequency than common nouns, and also differ in several other ways (Brédart, 2017; Brédart & Valentine, 1998; Burke et al, 1991). For example, proper names have been the subject of considerable investigation because they do not convey meaningful information about their referent (Cohen, 1990). Evidence is mixed however about the effects of semantic competition in the retrieval of proper names (Abrams & Davis, 2017). While some have found that related names facilitate retrieval of an intended proper name (Oberle & James, 2013; Vitkovitch et al, 2006), others have found that the presence of a

related proper name can cause interference during retrieval (Brédart & Dardenne, 2015; Deffler, et al, 2016). Further, the effects of competition on word retrieval also appear to be stronger for proper names than object names (Marful et al., 2014), suggesting that competition between concepts in the mental lexicon is not uniform but instead differs among different types of words.

The above evidence suggests that activation of semantically related words can constrain word retrieval. Importantly, much of this research focuses on single word retrieval in a minimalistic context in which words are accessed in the presence of a single competitor. However, a recent study by Shao and Rommers (2020) found that semantic interference between related words can be influenced by sentential context. Using a paradigm that combined highly constraining questions and PWI, Shao and Rommers asked participants to name a picture in response to a question that had an easily predictable answer (e.g., "What did the monkey eat yesterday?"). The researchers found that semantic interference from a superimposed distractor word was significantly reduced when participants named the picture in response to a highly constraining question. This finding suggests that effects of semantic competition are not all-or-nothing, but are instead influenced by the broader semantic context. Further, this implies that the process of lexical access is affected by both specific lexical features as well as the broader context in which a word is produced.

How does semantic context affect the course of lexical access? One common theory is that individuals can use sentence context to generate predictions about upcoming words. Words are typically produced more quickly in contexts that are highly constraining where there are fewer reasonable alternatives that can fit, thereby making it easier for listeners to predict. For example, given a sentence like *'George taught his son to drive a* _____', the number of potential words that could be reasonably expected to complete the sentence is smaller, making it easier to

predict a word like *'car'* compared to a sentence like *'Today I saw a red _____'*. The benefits of prediction during sentence processing are well documented in the literature. In language comprehension, listeners look faster towards easily predicted targets (Altmann & Kamide, 1999), and read words that are highly predictable faster, than targets that are not as predictable (Hintz et al., 2016; Vaino et al., 2009). Similarly, during language production, speakers are faster to produce words that complete a highly constraining sentence (Griffin & Bock, 1998; Staub et al., 2015). The widespread evidence that generating accurate predictions confers a processing benefit has led to multiple theories of language processing that argue prediction is a central mechanism in both language comprehension and production (e.g., Christiansen & Chater, 2016; Dell & Chang, 2014; Pickering & Garrod, 2013).

Despite ample evidence that highly predictable words are processed faster than unpredictable words, a study by Luke and Christianson (2016) found that highly predictable words are relatively rare in natural language, and in many cases the target word has a more expected competitor. This suggests that, although generating predictions may facilitate processing, everyday language does not offer many opportunities to take advantage of this. Further, what are the potential costs of failed predictions, given that many predictable words have a more expected competitor? Since highly constraining sentences reduce the range of plausible sentence completions, it is likely that alternative endings that are consistent with the sentence context will be semantically related to the predicted term. For instance, in the sentence fragment *'George taught his son to drive a ____', 'car'* may be the most likely ending, but alternatives such as *'truck'* are not only consistent with the meaning of the sentence, they are also semantically related to the predicted term. This means that when individuals make erroneous predictions, it is likely that their prediction will still be almost correct. Evidence on the

processing costs of incorrect or almost correct predictions is mixed, with some studies suggesting minimal processing costs to failed predictions (Frisson et al., 2017), and others suggesting that related words are in fact facilitated during prediction (Ness & Meltzer-Asscher, 2020). The debate surrounding the cost of failed predictions highlights important considerations about the utility of generating predictions during language processing. Specifically, if opportunities to predict are rare, and there are potential costs to inaccurate predictions, what is the utility of predicting at all?

A final consideration in this discussion of lexical access concerns how differences in part of speech affect ease of retrieval. Words typically fall into two main categories: content words that provide semantic information and function words that provide grammatical structure. Function words tend to be higher in frequency, more easily predictable, and less prone to speech errors than content words (Dell, 1990; Luke & Christianson, 2016; Segalowitz & Lane, 2000). One potential reason for these differences is that function words are a closed class of words that have less flexibility in terms of lexical selection. Contrasting with this is the lexical flexibility of selecting content words, which often have multiple synonyms that can be used to convey meaning. It is possible that access to function words is more cognitively automatic than access to content words because of their higher frequency and status as a closed class of words. Indeed, there is evidence to suggest that function words are produced more quickly than content words (Bell et al., 2009), but there is still considerable debate regarding whether this is due to differences in lexical access or simply because of the high frequency nature of function words.

Overview of Thesis

Together, the evidence discussed here demonstrates that lexical selection is influenced by multiple linguistic factors that can either facilitate or inhibit word retrieval. However, there are

still considerable gaps in our understanding of how these processes operate. For example, why do some contexts show that proper names are competitive rather than facilitatory? What is the utility of generating predictions during language processing? How is the relative ease of producing function compared to content words affected by effortful processing? The goal of the present dissertation is to investigate these questions using behavioural paradigms that elicit language production to determine how both word class and linguistic context affect lexical access.

First, we examine competition in proper name retrieval across two studies that required participants to name celebrity pictures after either reading a prime name aloud or classifying a prime name as belonging to a famous or non-famous person. This allowed us to measure how both immediate priming of a related name and accumulating interference across successive retrievals of proper names affects one's ability to successfully retrieve a target name. Next, we considered the utility of prediction during language processing by investigating whether predicting incorrectly or almost correctly still confers a processing benefit relative to not predicting at all. Finally, we examined the processing of function words using a read aloud task to determine whether the production of function words is effortful or cognitively automatic. Together these studies provide insight about how lexical access occurs across distinct categories of words – namely proper names, common objects, and function words – and also provide information about the effects of semantic context on word retrieval.

CHAPTER 2: COMPETITION ACCUMULATES IN SUCCESSIVE RETRIEVALS OF PROPER NAMES

Introduction

Our ability to produce words is among our most impressive. We produce words at a rate of 2-3 per second, and only rarely make speech errors (Levelt, 1989). This ability reflects an architecture where retrieval processes, as underpinned by mechanisms of facilitation (of intended words) and interference (of unwanted words), operate extremely efficiently. Contrasting with this is the frustrating experience where we fail to remember a person's name. Whether it is the name of an old acquaintance or the star of the latest movie, failure to remember a name can be frustrating and sometimes embarrassing, although it is quite common in everyday communicative exchanges. In fact, evidence suggests that proper names are especially susceptible to retrieval difficulties (e.g., Brédart, 2017; Brédart & Valentine, 1998; Burke et al, 1991; Cohen, 1990), and often manifest as tip-of-the-tongue (TOT) states, whereby a person has the experience of knowing a word but is unable to retrieve it. During a TOT, speakers have access to the semantic information, and can often retrieve partial phonological information of the target word (Brown, 1991). Here, we exploit this greater difficulty in retrieving proper names, including their high rate of TOTs, to better understand the mechanisms of facilitation and interference that lead to our usually accomplished ability to retrieve proper names.

TOTs are more likely to occur on attempted retrieval of low frequency words (Harley & Bown, 1998), which may partially explain why they are more prevalent for proper names, which tend to be lower in frequency than other types of words (Burke et al, 1991; Conley et al, 1999). However, the increased difficulty in the retrieval of proper names has also been attributed to several of their other properties, such as their typical combination of 2-3 individual names

(Hanley & Chapman, 2008) or their lack of synonyms (Brédart, 1993; see Abrams & Davis, 2017, for review). Another unique feature of proper names that might contribute to their retrieval difficulty is the fact that they are arbitrary labels. Whereas common objects may provide multiple semantic details that aid in retrieval, proper names do not provide any unique information about the referent (Cohen, 1990). For example, the object label *'table'* refers to a common type of object where members of the object category share overlapping features and functionality. On the other hand, a proper name such as *'Todd'* may refer to multiple different people who do not share overlapping characteristics that may act as cues during name retrieval. In fact, proper names are easier to retrieve when they are descriptive terms that identify an attribute of the person (e.g., 'Grumpy'), than when they do not share overlapping features with the individual (e.g., 'Mary Poppins'; Brédart & Valentine, 1998; Fogler & James, 2007), suggesting that the lack of descriptive properties common to many proper names contributes to their greater retrieval difficulty.

Although proper names do not share semantic overlap with their referents, they do share similarities with other terms — specifically, other proper names. Names can be related to each other in several ways, typically based on the relationship between the people they refer to. For example, names may be related to each other based on family membership or celebrity status. Of particular interest is whether related names function as competitors, or whether they can mutually facilitate retrieval (Abrams & Davis, 2017). The evidence on this question is mixed and could suggest that both are true in different conditions. Oberle and James (2013) investigated this question by asking participants to name a picture of a celebrity following a semantically and phonologically related proper name prime. They found that exposure to a related name reduced the likelihood of a TOT when naming the target celebrity. Based on this finding, Oberle and

James suggested that related names improve the chance of retrieving a target name by increasing the strength of the connections between lexical and phonological information. Similar results have suggested that exposure to a semantically related name reduces errors during face naming (e.g., Vitkovitch et al, 2006), and that proper name retrieval is facilitated by phonologically related proper or common names (e.g., Burke et al., 2004; Cross & Burke, 2004). However, although phonological overlap may facilitate retrieval, White and colleagues (2013) showed that this effect is diminished when a name is semantically related to the target. For related work on retrieval of common nouns, see Meyer and Bock (1992).

Despite evidence suggesting that activation of related names decreases naming errors and increases successful retrieval of proper names, other research indicates that related names interfere with retrieval. For example, when individuals misname a familiar person, they are more likely to substitute a related name such as that of another family member (e.g., Brédart & Dardenne, 2015; Deffler, et al, 2016; Dupont, 2018; Griffin & Wangerman, 2013). Visual and conceptual similarity between individuals also appears to increase competition between proper names, such that individuals are more susceptible to the "Moses illusion" — a phenomenon in which individuals fail to notice naming errors in sentences such as *'How many animals did Moses take on the ark?'* — when the correct name is replaced with the name of a visually or semantically similar individual (Davis & Abrams, 2016; Erickson & Mattson, 1981). Evidence further suggests that competition between proper names may be greater than competition between object names, where successive naming of targets is slower when targets consist of pictures of related celebrities than related common objects (Marful et al., 2014). Taken together, this evidence suggests proper names can compete for selection, perhaps to a greater extent than

object names. However, it contrasts with evidence that suggests related names can facilitate retrieval, reducing instances of word finding errors.

The present research aims to further examine the role of competition between proper names to narrow down the contexts in which related proper names reduce or increase retrieval failures. To address this question, we investigated whether proper names compete for selection across two experiments using a novel priming paradigm in which participants were asked to read a prime name aloud (Experiment 1), or judge if a prime name is famous or not (Experiment 2), prior to attempting to name a celebrity picture. Prime names were either celebrities related in age, gender, and occupation or non-celebrity names, and our measure of interest was whether participants would experience more TOTs and fewer correct retrievals (GOTs, i.e., successful retrieval of the proper name) after being primed by the name of another celebrity. If participants experience more TOTs after seeing a related celebrity prime, this could suggest proper names compete for selection based on broad categorical relationships, particularly if observed with a concomitant decrease in GOTs (Gollan & Brown, 2006). Alternatively, if TOTs are equally likely after either prime type, this would suggest that the relationship between names does not affect the likelihood of retrieval success. In addition, we examined whether TOTs increased and GOTs decreased over the course of the experiment to determine if competition from continuous naming results in more retrieval failures. Thus, our study allowed us to investigate the potential for a global effect of competition across the experiment, in addition to a local effect at the trial level.

Experiment 1

Method

Participants

The sample consisted of 106 participants (*age range*: 18 – 22, *M* = 18.32, *SD* = 0.80) recruited from the McMaster University undergraduate participant pool who participated for course credit. All participants reported native or native-like competency in English. We aimed for a sample size of at least 100 participants in this study. Although previous studies investigating competition in proper name retrieval (e.g., Marful et al., 2014; Oberle & James, 2013; Vitovitch et al., 2006) have often used smaller samples sizes, we wanted to be sure we would have a sufficiently large dataset given the potential for missing data or inappropriate responding in online studies.

Materials

The TOT task included 93 celebrity images taken from the "Then and Now" series by artist Ard Gelinck (@ardgelinck). This series of photos depicts celebrities alongside a younger version of themselves (https://mymodernmet.com/then-and-now-celebrity-photos-ard-gelinck/). The dual depictions of celebrities at younger and older ages serve as better stimuli because it improves the likelihood of identification by individuals who may recognize them from earlier or later stages of their career.

To create the experimental conditions, each celebrity image was paired with the name of a phonologically unrelated non-celebrity name to serve as a non-famous prime. The celebrity images were sorted into groups of triplets based on similar demographic factors such as gender, age, and occupation. However, given the limits of our stimulus set, it was not always possible to create triplets with exact matches across all demographic factors because the stimulus set did not

have a sufficient number of celebrities to evenly create triplets that were matched for occupation and age. We therefore prioritized matching targets based on gender, then occupation (noting that several of the targets fall into more than one category), and finally age. Final stimuli were created by using two famous images from each triplet as targets, one of which was paired with their non-famous prime, and the other paired with the name of the remaining image in the triplet to serve as the famous-name prime. Images were counterbalanced across lists, such that each celebrity served as a prime and a target across the experiment. This resulted in a total of six lists of 62 targets, half of which were primed with a non-famous distractor name, and half primed with a famous distractor name. Trials were fully randomized for each participant to allow us to investigate whether TOTs increased throughout the testing block without being confounded by target order. Table 1 shows characteristics of famous names and distractors in each of the six counterbalanced lists. A list of the triplets and their unrelated non-celebrity primes is shown in the Online Supplement.

Table 1.

Characteristics of Target and Distractors for Each List

Name Condition	List	Length (letters) M (SD)	Syllables M (SD)	Age M (SD)	Peak ngram M (SD)	Current ngram M (SD)
Target	1	11.56 (3.32)	3.84 (1.31)	56.60 (13.86)	6.73 (36.24)	6.28 (36.17)
	2	11.61 (2.78)	3.84 (1.20)	57.40 (18.77)	2.84 (12.64)	2.42 (12.17)
	3	11.56 (3.19)	3.68 (1.07)	57.61 (18.61)	8.27 (38.09)	7.86 (37.91)
	4	11.61 (2.78)	3.84 (1.20)	57.40 (18.77)	2.84 (12.64)	2.42 (12.17)
	5	11.56 (3.19)	3.68 (1.07)	57.61 (18.61)	8.27 (38.09)	7.86 (37.91)

	6	11.56 (3.32)	3.84 (1.31)	56.60 (13.86)	6.73 (36.24)	6.28 (36.17)
Famous Primes	1	11.65 (2.67)	3.68 (0.94)	58.42 (22.65)	4.37 (17.73)	4.00 (17.15)
	2	11.55 (3.70)	3.68 (1.19)	56.81 (13.76)	12.17 (51.03)	11.71 (50.96)
	3	11.61 (2.96)	4.00 (1.41)	56.39 (14.18)	1.30 (2.39)	0.85 (1.30)
	4	11.55 (3.70)	3.68 (1.19)	56.81 (13.76)	12.17 (51.03)	11.71 (50.96)
	5	11.61 (2.96)	4.00 (1.41)	56.39 (14.18)	1.30 (2.39)	0.85 (1.3)
	6	11.65 (2.67)	3.68 (0.95)	58.42 (22.65)	4.37 (17.73)	4.00 (17.15)
Non-famous Primes	1	10.97 (1.45)	3.65 (0.84)			
	2	11.77 (1.89)	4.00 (0.93)			
	3	11.65 (2.78)	3.77 (1.15)		N/A	
	4	11.39 (2.42)	3.77 (1.15)			
	5	10.97 (1.45)	3.65 (0.84)			
	6	11.77 (1.89)	4.00 (0.93)			

Note. For celebrities targets who have passed away, their age was calculated based on how old they would be in present day. Ngram refers to frequency of Google searches for a particular string of text (in the case, the celebrity names). The ngram values were multiplied by a constant of 100,000.

In addition to the main experimental task, participants completed the Shipley Vocabulary Test and the Media Savvy Test (Cross & Burke, 2004; Oberle & James, 2013). The Media Savvy Test requires individuals to classify movie and television titles as real or fake and provides a general measure of familiarity with popular media. We adapted the test from Oberle and James (2013) by updating 10 real titles to reflect more recent movies and television shows. In addition,

18 fake titles were changed to reflect the fact that some of the titles that previously were fake have since been made into movies, and others were existing (albeit obscure) movies or television shows. The final test consisted of 70 titles, 35 of which were fake. The Shipley Vocabulary Test (Shipley, 1940) consisted of 40 words increasing in difficulty with trial number, and participants were required to identify a synonym for each target word out of four possible words. We included these measures of media knowledge and vocabulary to ensure similar levels of word knowledge across participants, as this has been shown to predict incidences of TOTs on objects (Dahlgren, 1998).

Procedure

In the main experimental task, participants were shown the distractor name, with the image of the target celebrity underneath. The distractor name appeared 250 ms before the target image to increase the salience of the distractor. Participants were instructed to read the distractor name out loud, and then to name the celebrity depicted in the image. For each celebrity image, there were four possible response options: (1) GOT (correctly named the celebrity), (2) TOT (experienced a TOT state), (3) DK (don't know the celebrity depicted), or (4) FAMILIAR (the celebrity looked familiar, but they did not know the name). GOT and TOT were included as responses as these were our measures of primary interest. We additionally included DK as a response so participants could indicate if they did not know the celebrity depicted, and FAMILIAR so that participants could indicate if they recognized the celebrity image but did not know the name. These four response options are consistent with previous research evaluating TOT states, as well as approaches that have used familiarity judgments to distinguish between a feeling of knowing and a TOT state (e.g., Huijbers et al., 2017). Following their response, participants were asked to type out the name of the celebrity if they knew it. This also provided

participants a chance to resolve their TOT state. In the final stage of each trial, participants were shown the correct name of the target celebrity displayed above the image and were asked to indicate if this was the name they were thinking of, and, if not, if the name was familiar, the image was familiar, or if they did not know the celebrity. See Figure 1 for a depiction of the sequence of events in an example trial.

Participants first completed a practice trial in which each of the four response types was explained. Following the main experimental task, participants completed the Shipley Vocabulary Test and the Media Savvy Test. For each of the Shipley and Media Savvy Tests, participant responses were time limited such that they were given a maximum of 20 s to respond to each item. The experiment was programmed in PsychoPy (Peirce et al, 2019) and administered online through Pavlovia (https://pavlovia.org/). The entire experiment took approximately 45 minutes to complete.

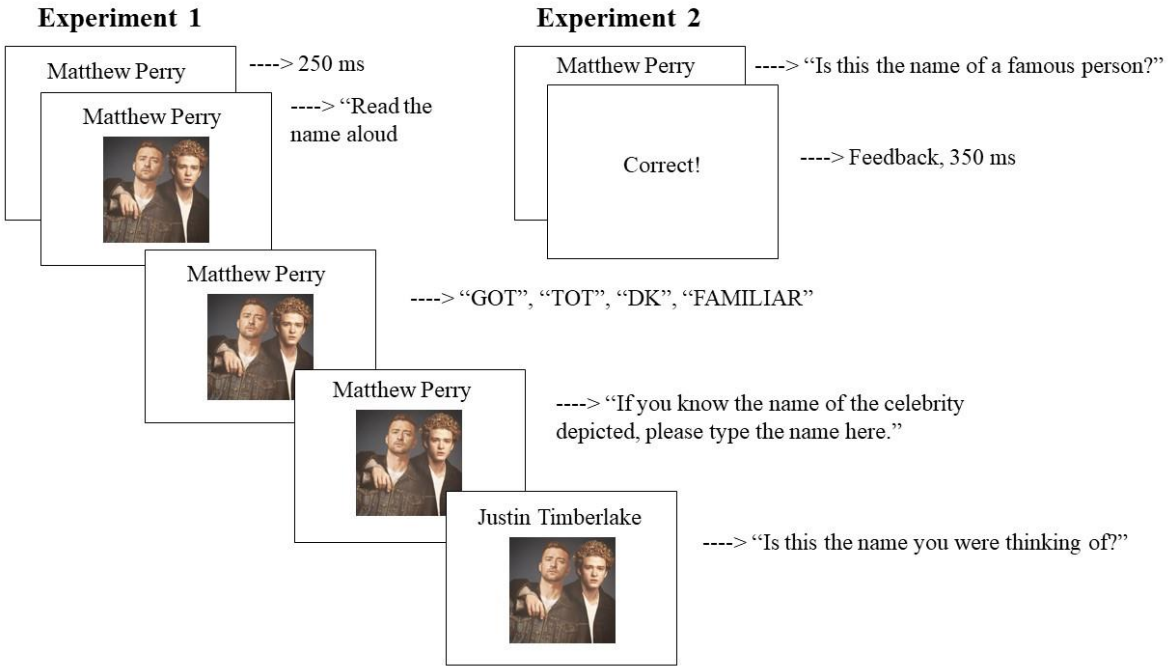

Figure 1. Example Trials in Experiment 1 (left) and Experiment 2 (right)

Data Preparation

To ensure that only valid responses were included in the analysis, responses were filtered such that GOTs and TOTs were only counted if participants indicated that the correct name was the name they were thinking of initially. For example, a GOT was removed if, after indicating that they knew the correct name, a participant failed to accurately identify the celebrity, suggesting that they did not actually successfully retrieve the name of the target (and in fact might not even know the name). Similar to previous research, TOTs were removed if the participant indicated that they were experiencing a TOT, but then reported that the correct celebrity name was not the name they were attempting to retrieve (e.g., Brown & Nix, 1996; Gollan & Brown, 2006; Resnik et al., 2014; Salthouse & Mandell, 2013; Shafto et al., 2007). We additionally filtered out trials where no response was given. This resulted in an average of 53.92 trials (*SD* = 10.28) out of a possible 62 trials remaining per participant. The percentage of each response type that was removed can be found in Table 2. We further excluded participants whose total number of trials after filtering was fewer than 2.5 SDs below the average number of remaining trials. This was done to ensure that participants who had a large number of invalid trials were not included in the final sample. This resulted in a final sample of 102 participants included in the analysis.[1] Descriptive statistics for age, vocabulary, and media savvy survey scores can be seen in Table 3.

Table 2.

Percentage of Responses Removed from the Analysis

	GOT	TOT	DON'T KNOW	FAMILIAR
Experiment 1	7%	29%	17%	9%
Experiment 2	19%	38%	31%	24%

[1] We note that the pattern of results reported here does not change if we conduct the analysis using the full, unfiltered dataset.

Table 3.

Descriptive Statistics for Participants in Experiments 1 and 2.

	Experiment 1		Experiment 2	
Measure	*M*	*SD*	*M*	*SD*
Age	18.32	0.80	19.10	2.77
Shipley Vocabulary Test	72%	13%	72%	10%
Media Savvy Survey	67%	11%	68%	9%

Results

Considering only valid responses, GOTs were the most common response (45%), followed by DKs (31%), FAMILIARs (14%), and TOTs (10%). Because our measures of interest were GOTs and TOTs, we do not report further on DK and FAMILIAR responses. A summary of the results for DKs and FAMILIARs can be found in the Online Supplement. All the raw data and analysis files are publicly available on the Open Science Framework (https://osf.io/zr9mp/?view_only=5fb22710ce2c40e7a2faa2cce60729b5).

Statistical analyses were carried out using R, Version 4.2.0. A logistic mixed effects regression analysis was carried out using the lme4 (Bates et al, 2015) and lmerTest (Kuznetsova et al, 2017) packages on GOTs and TOTs. Each model included a binary response (e.g., GOT or not GOT) as the dependent variable. Prime condition (contrast coded as .5 for famous vs. -.5 for non-famous) and trial number (which was centered and scaled), as well as their interaction were included as fixed effects. Random intercepts for participant and item were included, along with by-participant and by-item random slopes for condition, trial number, and their interaction. To avoid overfitting the model, we then determined which random slopes accounted for less than 1% of variance in the full model. This revealed that by-item and by-participant random slopes for condition were not required in the model examining TOTs, and that by-participant random slopes for condition and the interaction between trial and condition, and all by-item slopes, were not

required in the model examining GOTs. These slopes were removed in the final analyses reported here. See Table 4 for a detailed summary of the results.

Table 4.

Summary of Results for Logistic Mixed Effects Analysis for Experiment 1.

Effect	*B*	*SE* (*B*)	*Z*	*p*
TOTs				
(Intercept)	-2.88	0.16	-18.52	<.001
Trial	0.09	0.06	1.46	.145
Condition	0.01	0.09	0.07	.948
Trial x Condition	0.03	0.10	0.27	.786
GOTs				
(Intercept)	-0.44	0.32	-1.38	.169
Trial	-0.11	0.04	-2.64	**.008**
Condition	-0.02	0.08	-0.19	.847
Trial x Condition	-0.06	0.08	-0.70	.487

As can be seen in Figure 2, there was no effect of prime condition on GOTs or TOTs (*p*s > .05). However, participants experienced fewer successful retrievals with increasing trial number, main effect of trial number on GOTs (B = -0.11, SE = 0.04, Z = -2.64, p = .008). Conversely, the proportion of TOT responses did not significantly change with increasing trial number (p > .05). Figure 3 shows the relationship between number of responses and trial number, collapsed across condition.

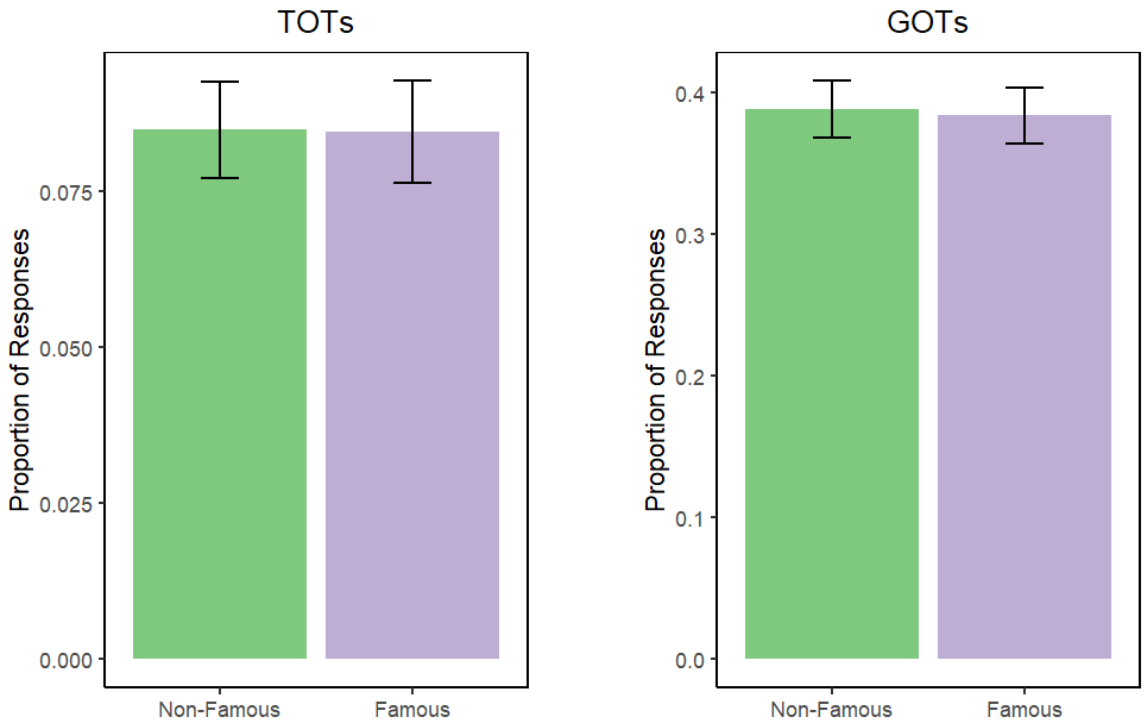

Figure 2. Proportion of TOTs and GOTs by Prime Condition in Experiment 1.

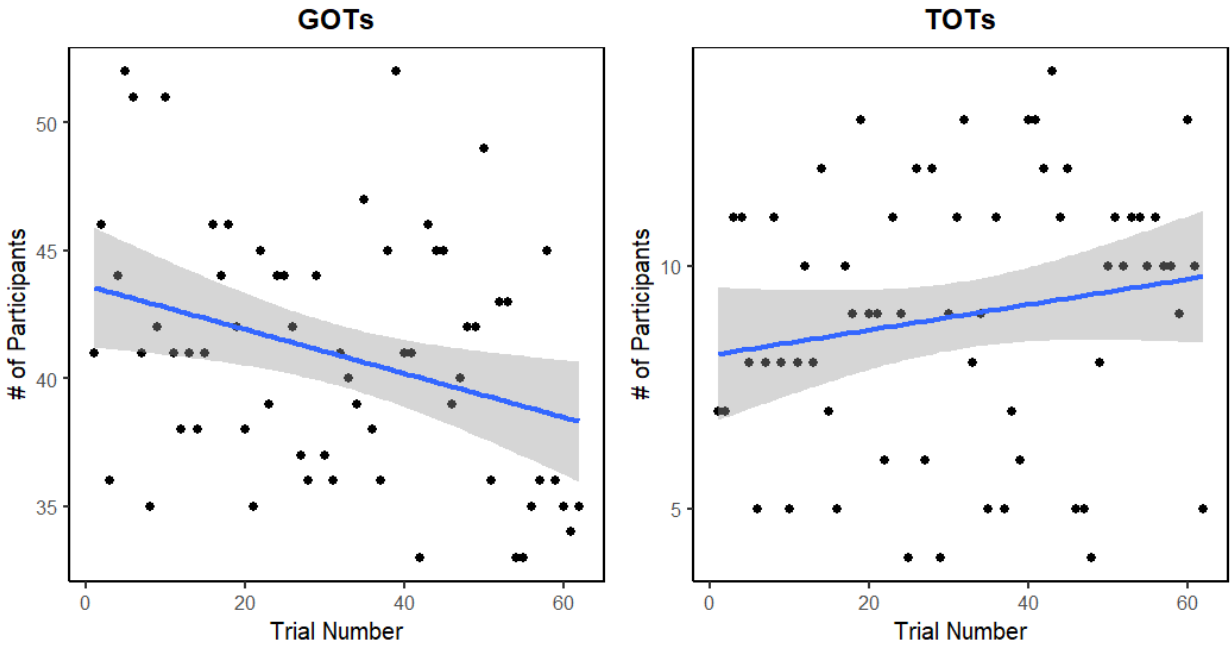

Figure 3. Number of Participant Responses per Trial for each Response Type in Experiment 1.

One possible mechanism that could underlie the decrease in successful retrievals is proactive interference, a phenomenon whereby previously learned information interferes with retrieval of new information (Anderson & McNeely, 1996; Underwood, 1957; Watkins & Watkins, 1975). It is possible that proactive interference from retrieval of celebrity names resulted in the decrease in successful name retrievals across trials.[2] To investigate this possibility, we conducted an exploratory analysis that examined whether the number of previously related or unrelated names influenced the chance of successfully retrieving a target name. We first classified all targets based on four general categories: female actors, male actors, female singers, and male singers. These categories were chosen as they represent the largest overlap in related features within our stimulus set. We then calculated the number of previous trials that matched the target's category (e.g., how many previous trials were also a female singer), as well as the number of previous trials that did not match the target's category. Previous matches were calculated using only the target's primary occupation to ensure maximum overlap between targets. For example, 'Will Smith' could be classified as both a singer and an actor but is more widely known for his acting career, and so previous matches were calculated based on the previous number of male actors that participants encountered prior to naming Will Smith.

A generalized linear model was constructed with GOT as the dependent variable and number of previous matches and number of previous non-matches as the predictor variables (centered and scaled). Participant and target were included as random factors, and number of previous matches was included as a by-participant slope (all other random slopes were determined to account for less than 1% of the variance, and were thus removed). This analysis revealed that participants were less likely to successfully retrieve a target name when they had encountered a greater number of previously related targets, B = -0.24, SE = 0.09, Z = -2.81, p =

[2] We thank an anonymous reviewer for pointing us in this direction.

.005. The number of previous unrelated targets did not have a significant effect on successful retrievals, $B = 0.08$, $SE = 0.07$, $Z = 1.19$, $p = .234$. Figure 4 shows the relationship between the proportion of successful retrievals and the number of previous related and unrelated targets. The analysis of previous matches on TOTs is not further considered as this response type did not show an effect of trial number, but is included in the Online Supplement.

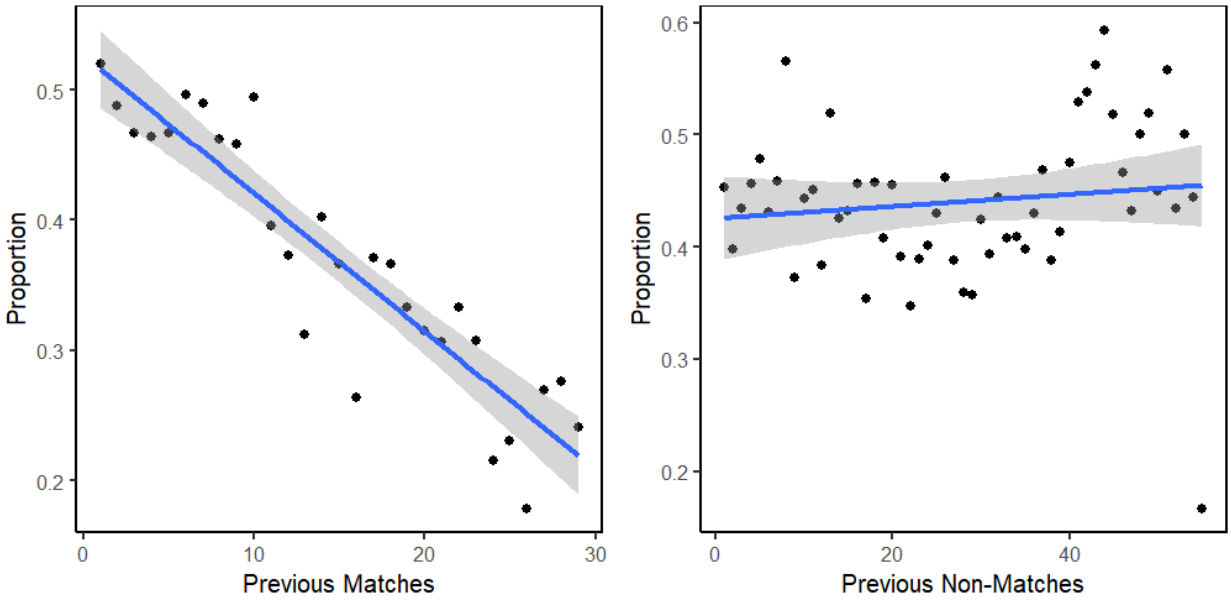

Figure 4. Proportion of GOTs per number of previous related targets (left) and unrelated targets (right) in Experiment 1.

Discussion

Experiment 1 investigated whether related names cause interference during the retrieval of proper names. We did not find any effect of prime condition or trial number on the occurrence of TOTs. However, analysis of GOTs revealed that successful retrievals decreased across trials, suggesting that successive retrieval of proper names leads to retrieval interference. Our exploratory analysis of whether GOTs depended on the number of previous related or unrelated targets indicated that individuals were less likely to successfully retrieve a target name when they had previously encountered a higher number of related celebrities. Thus, it appears that the effect

of trial number on successful retrievals may be a result of proactive interference from previous targets.

One potential limitation of this experiment is that the prime names were presented only a short interval before the target (250 ms), and participants were asked to read these names aloud while simultaneously viewing the target picture. It is therefore possible that the prime names were only processed at a shallow level, meaning participants may have had insufficient opportunity to recall the individual referred to by the prime name prior to accessing the name of the target. Further, this design did not allow us to control for the possibility that participants did not recognize prime names as belonging to celebrities. To account for these possibilities, we conducted a second experiment that required participants to identify the prime names as famous or non-famous prior to seeing the target. If competition from prime names depends on depth of processing, we should see an increase in TOTs following famous primes when participants are required to classify the prime name prior to naming the target celebrity. Alternatively, if competition between proper names only results from accumulating competition during successive retrievals, we should see the same effect of trial number on successive retrievals, but no effect of prime condition in Experiment 2.

Experiment 2

Method

Participants

The sample consisted of 234 participants (*age range* = 17 – 49, *M* = 19.08, *SD* = 2.77) from the McMaster University undergraduate participant pool who participated for course credit. One additional participant was removed because they did not speak English as their dominant language. We collected a larger sample of participants in Experiment 2 than Experiment 1

because although our target was to collect at least 100 participants for each, we were aware that there was a greater likelihood of removing invalid trials in Experiment 2. Specifically, participants were asked to classify the prime as famous or not famous which allowed us to detect trials where participants failed to recognize the celebrity status of the names. Since we aimed to remove trials where participants did not correctly classify the prime name (allowing us to ensure that participants recognized the prime name as famous), a larger sample size in Experiment 2 would provide a sufficiently large data set to investigate our primary research question.

Materials

Materials were identical to those used in Experiment 1.

Procedure

The procedure was the same as in Experiment 1, with the exception that in the main experimental task, participants were first shown the distractor name individually on the screen and were asked to classify the name as belonging to a famous person (famous condition) or non-famous person (non-famous condition). Following this classification, they were told if their response was correct. The feedback remained on the screen for 350 ms, after which the target celebrity image was displayed underneath the distractor name. Following the main experimental task, participants again completed the Shipley Vocabulary Test and Media Savvy Survey. The experiment was programmed in PsychoPy (Pierce et al, 2019) and administered online through Pavlovia (https://pavlovia.org/). The entire experiment took approximately 45 minutes to complete.

Data Preparation

Data were filtered the same way as in Experiment 1. In addition, we excluded trials where participants incorrectly classified the prime name to ensure that famous primes were in

fact recognized as belonging to celebrities. This resulted in an average of 45.96 trials (*SD* = 9.14) out of a possible 62 trials remaining per participant. The percentage of excluded trials can be found in Table 2.[3] We again excluded participants whose total number of trials after filtering was fewer than 2.5 SDs below the average number of remaining trials, resulting in a final sample of 229 participants included in the analysis. As can be seen in Table 3, participants in Experiment 2 were similar in terms of age, vocabulary, and media knowledge as those in Experiment 1.

Results

As in Experiment 1, GOTs were the most common response (45%), followed by DKs (29%), FAMILIARs (14%), and TOTs (12%). We do not report further on DK and FAMILIAR responses, but a summary of the results for DKs and FAMILIARs for Experiment 2 can be found in the Online Supplement.

Statistical analyses were the same as those performed in Experiment 1. We constructed two logistic mixed effects models with a binary response (e.g., GOT or not GOT) as the dependent variable, and prime condition (coded as .5 for famous vs. -.5 for non-famous) and trial number (which was centered and scaled), as well as their interaction were included as fixed effects. Random intercepts for participant and item were included, along with by-participant and by-item random slopes for condition, trial number, and their interaction. We again determined which random slopes accounted for less than 1% of variance in the full model. This resulted in the exclusion of by-participant random slopes for condition, and by-item random slopes for trial number in the model examining TOTs, as well as the exclusion of by-item and by-participant

[3] Note that, if we do not exclude trials based on the whether the prime name was classified correctly, the average number of remaining trials per participant is comparable to that of Experiment 1, *M* = 55.29, *SD* = 6.39). Additionally, if we do not consider whether participants correctly classified the prime name in Experiment 2, the percentage of trials that were excluded per response is comparable to Experiment 1 (i.e., 6%, 27%, 12%, and 8% for GOTs, TOTs, DKs, and FAMILIARs, respectively). As in Experiment 1, the pattern of results reported does not change if we perform the analysis with the full, unfiltered dataset.

random slopes for condition and trial number in the model examining GOTs. See Table 5 for a detailed summary of the results.

Table 5.

Summary of Results for Logistic Mixed Effects Analyses for Experiment 2.

Effect	*B*	*SE* (*B*)	*Z*	*p*
TOTs				
(Intercept)	-2.58	0.13	-20.09	<.001
Trial	-0.00	0.03	-0.09	.928
Condition	0.15	0.07	2.18	**.030**
Trial x Condition	-0.00	0.07	-0.03	.975
GOTs				
(Intercept)	-0.59	0.27	-2.19	.030
Trial	-0.14	0.03	-4.66	**<.001**
Condition	-0.05	0.06	-0.82	.415
Trial x Condition	0.09	0.08	1.18	.240

Beginning with TOTs, participants reported significantly more TOTs following famous primes than following control primes, a main effect of prime condition on TOTs (B = 0.15, SE = 0.07, Z = 2.18, p = .030). No other effect on TOTs reached significance (ps > .05). Proportion of response types by prime condition can be seen in Figure 5. There was also no effect of prime condition on GOTs (p > .05). However, over the course of the experiment the number of GOTs decreased progressively, a main effect of trial number for GOTs (B = -0.14, SE = 0.03, Z = -4.66, p <.001), replicating the effect of accumulating interference seen across trials in Experiment 1. See Figure 6 for visualization of responses by trial number.

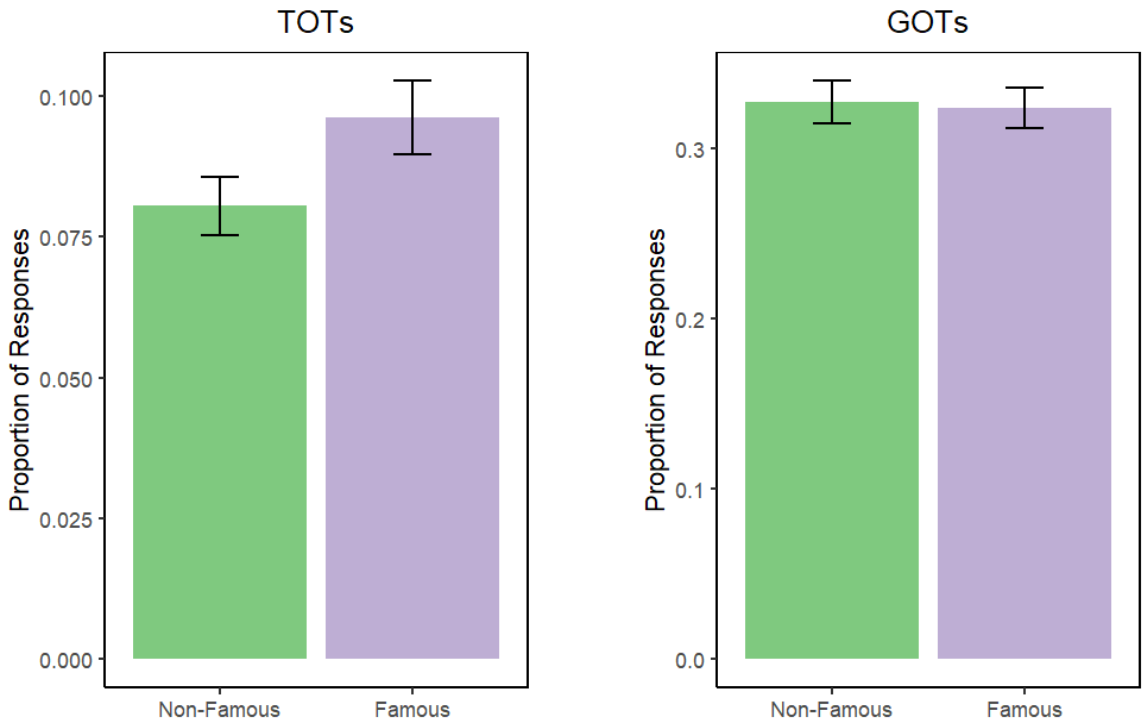

Figure 5. Proportion of TOTs and GOTs by Prime Condition in Experiment 2.

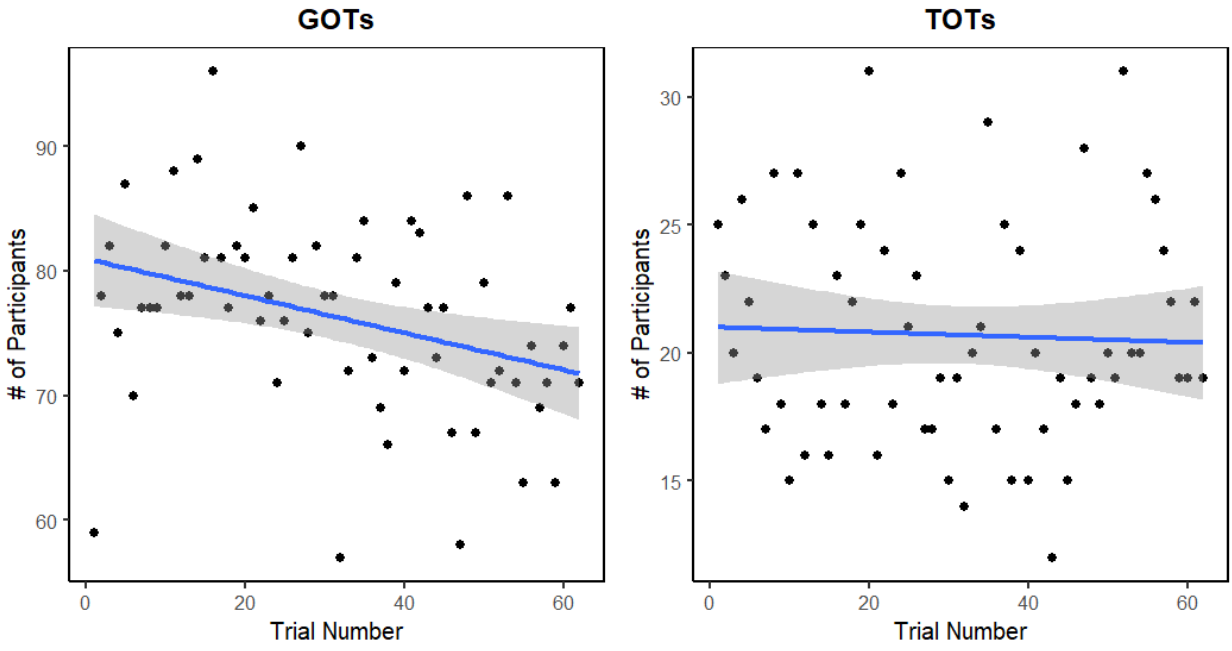

Figure 6. Number of Participant Responses per Trial for each Response Type in Experiment 2.

As in Experiment 1, we conducted a follow-up exploratory analysis to investigate the potential effects of proactive interference on successful retrievals. Data were restricted to targets that fell into the categories of female actor, male actor, female singer, and male singer. A generalized linear model was constructed with GOT as the dependent variable, and number of previous matches and number of previous non-matches as the predictor variables (centered and scaled). Participant and target were included as random effects, and number of previous matches and previous non-matches was included as a by-participant slopes (by-target random slopes were excluded because they accounted for less than 1% of the variance, and were thus removed). This analysis mirrored the results of Experiment 1, showing that correct retrievals decreased with increasing number of previously related targets, $B = -0.19$, $SE = 0.06$, $Z = -3.09$, $p = .002$, but there was no significant relationship between successful retrievals and number of previous unrelated targets, $B = 0.00$, $SE = 0.05$, $Z = 0.04$, $p = .969$. Figure 7 shows the relationship between the proportion of successful retrievals and the number of previous related and unrelated targets. Given the non-significant effect of trial number on TOTs, we do not report further on this analysis here, but a summary of the results for TOTs can be found in the Online Supplement.

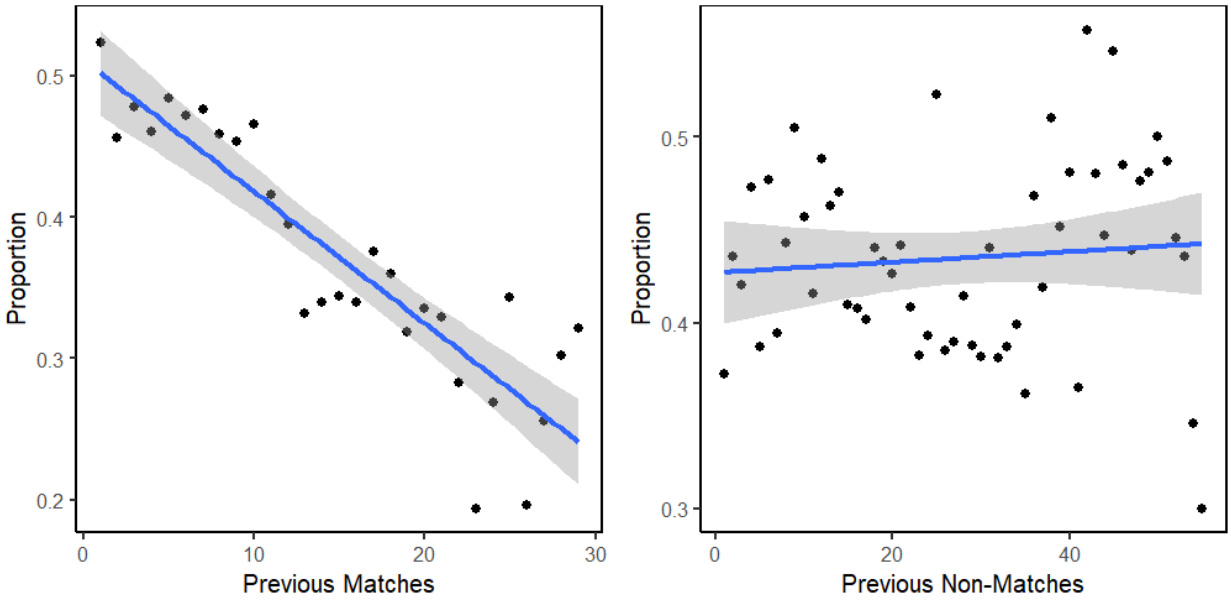

Figure 7. Proportion of GOTs per Number of Previous Related Targets (panel 1) and Unrelated Targets (panel 2) in Experiment 2.

Discussion

Experiment 2 investigated competition between proper names by drawing increased attention to primes using a fame-judgment task prior to naming a target celebrity. Participants experienced more TOTs following famous primes than non-famous primes, suggesting that when related names are processed deeply in advance of viewing a target they interfere with retrieval. However, prime names had no effect on GOTs, which complicates the interpretation of the priming manipulation. By contrast, successful retrievals did decrease with increasing trial number, providing clearer evidence of competition for selection between proper name representations. Exploratory analysis revealed that individuals are less likely to successfully retrieve a proper name as the number of previous related targets increases. Thus, it appears that the decrease in successful retrievals across trials can be explained by proactive interference from the number of previously related targets. Overall, the results of this second experiment indicate that related names can cause interference during name retrieval, but the strength of interference

is dependent on context. Whereas successful retrievals reliably decrease due to accumulating interference between pictures of related famous people across trials, TOTs only increase following semantic classification of written famous name primes on individual trials, suggesting a weaker form of trial-to-trial effects. Because word primes did not affect GOTs, the increase in TOTs does not provide clear evidence of competition for selection (discussed in detail below).

General Discussion

Across two experiments, we investigated whether proper names compete for selection in a local context within trials and in a global context across trials. In Experiment 1, we found successful retrievals (GOTs) decreased with increasing trial number, but no effect of written prime names on GOTs or TOTs, suggesting that accumulating competition between attempted retrieval of the names of depicted famous faces increases retrieval difficulties. Experiment 2 replicated the effect of trial number on GOTs, again showing robust effects of trial number on successful retrievals. In addition, in Experiment 2, participants were more likely to experience a TOT for celebrity names after judging a written famous name prime as famous than after judging a written non-famous name as non-famous. Thus, written primes that immediately preceded targets increased the likelihood of TOTs when participants are forced by the task to process the written primes at a semantic level. Overall, the results indicate that interference can accumulate across trials when speakers attempt to name multiple loosely related famous names in succession. Additionally, exploratory analyses suggested that the decrease in successful retrievals across trials in Experiment 1 and 2 can be attributed to proactive interference from previous related names. By contrast, proactive interference did not significantly affect TOTs across trials in either experiment.

Given the discrepancy in the effects of written primes on TOTs between Experiment 1 and 2, but the consistent effects of trial number in both experiments, it appears that within trial interference from written primes is weaker than cumulative interference across the course of the experiment (which could reflect interference from both written primes and related pictures). Written prime effects were dependent on depth of processing of the related name. In Experiment 1, participants were only required to passively read the prime names, and we did not find any effect of the primes on incidences of TOTs. In Experiment 2, participants were required to actively process primes by classifying the prime name based on whether it belonged to a famous or non-famous person, which resulted in more TOTs following famous primes. We interpret these results based on a theory that lexical access relies on a two-step process involving retrieving a concept from the mental lexicon, followed by encoding the specific word form (Levelt, 1989). A common assumption is that successful word retrieval (GOTs) reflects completed lexical access, whereas TOTs reflect partial lexical access wherein an individual has successfully retrieved the lexical concept, but failed to encode the word form (Brown, 1991; Gollan & Brown, 2006; Meyer & Bock, 1992; but see Huebert et al., 2023). Although the present experiments were not designed to assess partial lexical access in TOT states, one possible explanation is that semantically related names may provide only a slight boost in retrieval. On this account, written primes may have been sufficient to allow participants to access the lexical concept when they otherwise might not have (Gollan & Brown, 2006), but not sufficient to cue retrieval of phonology, resulting in partial lexical access and therefore more TOTs but not more GOTs. By contrast, it is more difficult to explain why related written names would interfere with retrieval but only increase TOTs while not decreasing GOTs. Such an account would need to

make a further assumption, for example that written primes could only very temporarily interfere with retrieval (eliciting a brief TOT but not reducing GOTs).

The difficulty in determining whether related names facilitate or interfere with retrieval is well documented in the literature, with multiple studies arguing for either interference (e.g., Brédart & Dardenne, 2015; Davis & Abrams, 2016; Deffler, et al, 2016; Dupont, 2018; Griffin & Wangerman, 2013; Marful et al., 2014), or facilitation (e.g., Burke et al., 2004; Cross & Burke, 2004; Oberle & James, 2013; Vitkovitch et al, 2006). Abrams and Davis (2017) suggest that proper names are not uniformly competitive, and effects of facilitation and competition may depend on the type of overlap between items. For example, competition can arise due to increased visual similarity between competitors (e.g., Davis & Abrams, 2016), whereas shared phonological and semantic information can facilitate retrieval (e.g., Oberle & James, 2013; Vitkovitch et al, 2006). In the present experiments, primes appeared as written names, while all targets were presented as pictures. Whereas written primes might have led to greater focus on semantic representations in our study, visual presentation of pictures of many famous people across trials may have led to a greater interference with retrieval, resulting in a dissociation between GOTs and TOTs in the present studies. Previous research has also demonstrated that GOTs and TOTs can be differently affected by experimental manipulations (e.g., Cross & Burke, 2004; Gollan et al., 2014). For example, Cross and Burke (2004) found that the production of related names prior to naming a target picture did not affect TOTs, but did result in fewer incorrect responses. Thus, it appears that lexical access is sensitive to the way stimuli are presented, with different experimental manipulations affecting different levels of the process.

Although the effect of prime names on TOTs could arguably reflect facilitation, boosting a DK response to a TOT, or interference, the decrease in GOTs across trials demonstrates robust

interference from previous related names. Previous research has suggested that proper names are especially prone to weak connections between their conceptual features and lexical concepts (e.g., Burke et al., 2004; Mortensen et al., 2006), which may explain why lexical access is particularly susceptible to interference across trials when individuals continuously name multiple related targets. Further, it might seem that the accumulating interference across trial number is unique to retrieval of proper names given null effects of trial number in a recent study of TOTs in which pictures of objects elicited naming responses. Stasenko and Gollan (2019) examined the frequency of TOTs on common objects among monolingual and bilingual speakers following exposure to video clips. In that study, exposure to a short movie clip (about ten minutes) in another language increased TOTs for objects, but GOTs were not affected by exposure to the movie clips. Neither GOTs nor TOTs were affected by increasing trial number. Notably, that study involved a higher number of trials than the present studies, suggesting that the effect found in our studies is not likely due to fatigue, but rather a result of greater interference caused by the unique properties of proper names. However, the objects in Stasenko and Gollan were mostly unrelated object names. As with proper names, retrieval of object names is susceptible to interference from closely related terms (Rose et al., 2018), and phonological cues provide a better retrieval aid than semantically related words (Meyer & Bock, 1992), leaving open the possibility that retrieval of multiple semantically related targets results in proactive interference within a single list, regardless of whether individuals are retrieving object or proper names.

Our exploratory analyses suggest that the decrease in GOTs across trials may be driven by proactive interference, as participants were less likely to successfully retrieve a target name when they had encountered a greater number of related targets prior. Studies investigating proactive interference typically demonstrate that memory from previously encoded targets

interferes with recall of more recently encoded, related targets. For example, participants may be asked to encode and then recall consecutive lists of related items. During recall, related items from earlier lists may be inserted, indicating that earlier information interferes with recall of newly encoded items (e.g., Kliegl et al., 2015; Underwoord, 1957; Weinstein et al., 2011; Wixted & Rohrer, 1993). Here we show that interference from previously related items may disrupt successful retrieval of related names during successive recall of targets, even in the absence of a memory test that requires free recall of a list of items. Although we did not directly manipulate the number of previous related names, the relationship between the number of previous retrieval attempts of related names had a highly robust and disruptive effect on lexical access of subsequent names.

Finally, although we ruled out cognitive fatigue as the mechanism driving the decrease in successful retrievals in favor of proactive interference, one could argue that because proper names are more difficult to retrieve, successive retrievals are more likely to result in fatigue. However, while successful retrievals decreased as the number of previous related targets increased, we also observed an opposite effect for FAMILIAR responses in Experiment 2 whereby FAMILIAR responses were more likely as participants encountered more related targets ($B = 0.08$, $SE = 0.03$, $Z = 2.34$, $p = .019$). Although FAMILIAR responses were not a primary variable of interest, they were included in this study to allow participants to differentiate between feelings of familiarity and true TOT states. The finding that familiarity increased while successful retrieval decreased suggests that participants were still able to recognize the targets but had greater difficulty accessing the lexical concept associated with the target name.

Together, the results of these two experiments suggest that repeated attempts to name pictures of famous people elicits interference effects that are stable across small variations in the

task, while the effects of short-term priming from related written names may require specific processing of related primes. Although it was not possible to determine whether the TOT effects reflected facilitation or interference from related names, the consistent finding that GOTs decreased across successive retrievals of proper names arguably constitutes more powerful evidence of longer-term and more profound competition for selection between related names. In this respect, the present study contributes to a growing number of experimental results revealing that proper names can compete for selection when considering long term retrieval of multiple proper names. Speculating, if interference effects can also accumulate over a lifetime of proper name retrieval, the current results might explain why aging related difficulty with proper name retrieval is one of the most robust effects found in the literature on cognitive aging (Burke et al., 1991; Mortensen et al., 2006). Indeed, previous research has suggested that changes in cognitive abilities across the lifespan may reflect the increasing knowledge that individuals acquire as they age (e.g., Dahlgren, 1998; Ramscar et al., 2014). This interpretation is limited by the fact that we cannot control the number of related or unrelated names that individuals encounter across their lifetime. In follow-up work it will be interesting to determine if the effects of trial number can be replicated during successive attempted retrieval of semantically related object names, or if such effects are restricted to retrieval of proper names. If the latter is found, it could imply that part of what makes retrieval of proper names so difficult is the very large number of semantically similar competitors within this category, which is a unique feature of proper names not previously considered among explanations of why proper names are so TOT prone.

References

Abrams, L., & Davis, D. K. (2017). Competitors or teammates: how proper names influence each other. *Current Directions in Psychological Science*, *26*(1), 87-93. http://dx.doi.org/10.1177/0963721416677804

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), Memory: Handbook of perception and cognition (2nd ed., pp. 237–313). San Diego, CA: Academic Press.

Bates, D., Maechler, B., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*, 1—48. http://dx.doi.org/10.18637/jss.v067.i01

Brédart, S. (1993). Retrieval failures in face naming. *Memory*, *1*(4), 351-366. http://dx.doi.org/10.1080/09658219308258243

Brédart, S. (2017). The cognitive psychology and neuroscience of naming people. *Neuroscience & Biobehavioral Reviews*, *83*, 145-154. http://dx.doi.org/10.1016/j.neubiorev.2017.10.008

Brédart, S., & Dardenne, B. (2015). Similarities between the target and the intruder in naturally occurring repeated person naming errors. *Frontiers in psychology*, *6*, 1474. http://dx.doi.org/10.3389/fpsyg.2015.01474

Brédart, S., & Valentine, S. B. T. (1998). Descriptiveness and proper name retrieval. *Memory*, *6*(2), 199-206. https://doi.org/10.1080/741942072

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological bulletin*, *109*(2), 204. http://dx.doi.org/10.1037/0033-2909.109.2.204

Brown, A. S., & Nix, L. A. (1996). Age-related changes in the tip-of-the-tongue experience. *The American journal of psychology*, 79-91. https://doi.org/10.2307/1422928

Burke, D. M., Locantore, J. K., Austin, A. A., & Chae, B. (2004). Cherry pit primes Brad Pitt: Homophone priming effects on young and older adults' production of proper names. *Psychological Science*, *15*(3), 164-170. http://dx.doi.org/10.1111/j.0956-7976.2004.01503004.x

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults?. *Journal of memory and language*, *30*(5), 542-579. http://dx.doi.org/10.1016/0749-596X(91)90026-G

Cohen, G. (1990). Why is it difficult to put names to faces?. *British Journal of Psychology*, *81*(3), 287-297. http://dx.doi.org/10.1111/j.2044-8295.1990.tb02362.x

Conley, P., Burgess, C., & Hage, D. (1999). Large-scale databases of proper names. *Behavior Research Methods, Instruments, & Computers*, *31*(2), 215-219. http://dx.doi.org/10.3758/BF03207713

Cross, E. S., & Burke, D. M. (2004). Do alternative names block young and older adults' retrieval of proper names?. *Brain and language*, *89*(1), 174-181. http://dx.doi.org/10.1016/S0093-934X(03)00363-8

Dahlgren, D. J. (1998). Impact of knowledge and age on tip-of-the-tongue rates. *Experimental Aging Research*, *24*(2), 139-153. http://dx.doi.org/10.1080/036107398244283

Davis, D. K., & Abrams, L. (2016). Here's looking at you: Visual similarity exacerbates the Moses illusion for semantically similar celebrities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 75. http://dx.doi.org/10.1037/xlm0000144

Deffler, S. A., Fox, C., Ogle, C. M., & Rubin, D. C. (2016). All my children: The roles of semantic category and phonetic similarity in the misnaming of familiar individuals. *Memory & cognition*, *44*(7), 989-999. http://dx.doi.org/10.3758/s13421-016-0613-z

Dupont, M. (2018). The similarities between the target and the intruder in naturally occurring person naming errors: A comparison between repeated and single naming confusions. *Journal of psycholinguistic research*, *48*(1), 33-42. http://dx.doi.org/10.1007/s10936-018-9586-3

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551. https://doi.org/10.1016/S0022-5371(81)90165-1

Fogler, K. A., & James, L. E. (2007). Charlie Brown versus Snow White: The effects of descriptiveness on young and older adults' retrieval of proper names. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *62*(4), 201-207. http://dx.doi.org/10.1093/geronb/62.4.P201

Gollan, T. H., & Brown, A. S. (2006). From tip-of-the-tongue (TOT) data to theoretical implications in two steps: when more TOTs means better retrieval. *Journal of Experimental Psychology: General*, *135*(3), 462. http://dx.doi.org/10.1037/0096-3445.135.3.462

Gollan, T. H., Ferreira, V. S., Cera, C., & Flett, S. (2014). Translation-priming effects on tip-of-the-tongue states. *Language, Cognition and Neuroscience*, *29*(3), 274-288. http://doi.org/10.1080/01690965.2012.762457

Griffin, Z. M., & Wangerman, T. (2013). Parents accidentally substitute similar sounding sibling names more often than dissimilar names. *PLoS One*, *8*(12), e84444. http://dx.doi.org/10.1371/journal.pone.0084444

Hanley, J. R., & Chapman, E. (2008). Partial knowledge in a tip-of-the-tongue state about two- and three-word proper names. *Psychonomic Bulletin & Review*, *15*(1), 156-160. http://dx.doi.org/10.3758/PBR.15.1.156

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, *89*(1), 151-174. http://dx.doi.org/10.1111/j.2044-8295.1998.tb02677.x

Huebert, A. M., McNeely-White, K. L., & Cleary, A. M. (2023). On the relationship between tip-of-the-tongue states and partial recollective experience: Illusory partial recollective access during tip-of-the-tongue states. *Journal of Experimental Psychology: General*, *152*(2), 542. https://doi.org/10.1037/xge0001292

Huijbers, W., Papp, K. V., LaPoint, M., Wigman, S. E., Dagley, A., Hedden, T., ... & Sperling, R. A. (2017). Age-related increases in tip-of-the-tongue are distinct from decreases in remembering names: a functional MRI study. *Cerebral Cortex*, *27*(9), 4339-4349. https://doi.org/10.1093/cercor/bhw234

Kliegl, O., Pastötter, B., & Bäuml, K. H. T. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1778. http://dx.doi.org/10.1037/xlm0000096

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1—26. http://dx.doi.org/10.18637/jss.v082.i13

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press. http://dx.doi.org/10.7551/mitpress/6393.001.0001

Marful, A., Paolieri, D., & Bajo, M. T. (2014). Is naming faces different from naming objects? Semantic interference in a face-and object-naming task. *Memory & cognition*, *42*(3), 525-537. http://dx.doi.org/10.3758/s13421-013-0376-8

Meyer, A. S., & Bock, K. (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation?. *Memory & cognition*, *20*(6), 715-726. http://dx.doi.org/10.3758/BF03202721

Mortensen, L., Meyer, A. S., & Humphreys, G. W. (2006). Age-related effects on speech production: A review. *Language and Cognitive Processes*, *21*(1-3), 238-290. http://dx.doi.org/10.1080/01690960444000278

Oberle, S., & James, L. E. (2013). Semantically-and phonologically-related primes improve name retrieval in young and older adults. *Language and cognitive processes*, *28*(9), 1378-1393. http://dx.doi.org/10.1080/01690965.2012.685481

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203. http://dx.doi.org/10.3758/s13428-018-01193-y

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5-42. https://doi.org/10.1111/tops.12078

Resnik, K., Bradbury, D., Barnes, G. R., & Leff, A. P. (2014). Between thought and expression, a magnetoencephalography study of the "tip-of-the-tongue" phenomenon. *Journal of Cognitive Neuroscience*, *26*(10), 2210-2223. https://doi.org/10.1162/jocn_a_00611

Rose, S. B., Aristei, S., Melinger, A., & Abdel Rahman, R. (2019). The closer they are, the more they interfere: Semantic similarity of word distractors increases competition in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 753. https://doi.org/10.1037/xlm0000592

Salthouse, T. A., & Mandell, A. R. (2013). Do age-related increases in tip-of-the-tongue experiences signify episodic memory impairments?. *Psychological science*, *24*(12), 2489-2497. https://doi.org/10.1177/0956797613495881

Shafto, M. A., Burke, D. M., Stamatakis, E. A., Tam, P. P., & Tyler, L. K. (2007). On the tip-of-the-tongue: neural correlates of increased word-finding failures in normal aging. *Journal of cognitive neuroscience*, *19*(12), 2060-2070. https://doi.org/10.1162/jocn.2007.19.12.2060

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, *9*(2), 371-377. http://dx.doi.org/10.1080/00223980.1940.9917704

Stasenko, A., & Gollan, T. H. (2019). Tip of the tongue after any language: Reintroducing the notion of blocked retrieval. *Cognition*, *193*, 104027. http://dx.doi.org/10.1016/j.cognition.2019.104027

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, *64*(1), 49. https://doi.org/10.1037/h0044616

Vitkovitch, M., Potton, A., Bakogianni, C., & Kinch, L. (2006). Will Julia Roberts harm Nicole Kidman? Semantic priming effects during face naming. *Quarterly Journal of Experimental Psychology*, *59*(6), 1134-1152. http://dx.doi.org/10.1080/02724980543000178

Weinstein, Y., McDermott, K. B., & Szpunar, K. K. (2011). Testing protects against proactive interference in face–name learning. *Psychonomic bulletin & review*, *18*(3), 518-523. https://doi.org/10.3758/s13423-011-0085-x

White, K. K., Abrams, L., & Frame, E. A. (2013). Semantic category moderates phonological priming of proper name retrieval during tip-of-the-tongue states. *Language and Cognitive Processes*, *28*(4), 561-576. http://dx.doi.org/10.1080/01690965.2012.658408

Wixted, J. T., & Rohrer, D. (1993). Proactive interference and the dynamics of free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(5), 1024. https://doi.org/10.1037/0278-7393.19.5.1024

CHAPTER 3: EFFECTS OF CLOZE PROBABILITY AND SEMANTIC SIMILARITY ON FAILED PREDICTIONS

Introduction

The use of prediction during language comprehension has become widely accepted as a beneficial mechanism that can facilitate processing (e.g., Christiansen & Chater, 2016; F. Ferreira, & Chantavarin, 2018). Specifically, sentence processing is made easier in predictable contexts, where upcoming information can be readily anticipated (e.g., Altmann & Kamide, 1999; DeLong et al., 2005; Heilbron et al., 2022; Van Berkum et al., 2005). Theories that propose benefits of prediction in language processing often support the notion that predicting will necessarily be better than not predicting. While it seems clear that predicting correctly confers processing benefits relative to when predictions cannot be generated (e.g., Griffin & Bock, 1998; Hintz et al., 2016; Vainio et al., 2009), there is still significant debate regarding whether predicting almost correctly, or entirely incorrectly, is still better than not predicting at all. In the current research, we address this debate by directly comparing the costs of failed predictions in highly constraining contexts to the same words presented in low constraint contexts to determine whether unexpected words in highly constraining contexts are produced more slowly than in contexts where predictions cannot be generated. Unexpected words were either related or unrelated to the predicted term, thus allowing us to ask whether predicting almost correctly is better than predicting completely incorrectly, and if predicting almost correctly is better than not predicting at all.

The processing benefits of accurately predicting upcoming words are well documented in the literature on both language production and language comprehension. For example, studies have found faster reading times and faster speech onset latencies for words appearing in highly

constraining contexts (e.g., Griffin & Bock, 1998; Hintz et al., 2016; Vainio et al., 2009), along with ERP components suggesting facilitated processing for highly predictable terms (e.g., Boudewyn et al., 2015; DeLong et al., 2005; Heilbron et al., 2022). Most of this research relies on data elicited using a cloze task, which requires individuals to fill in a missing word in a given sentence (Taylor, 1953). Words are said to have a higher cloze probability when the likelihood of people producing a single term to fill in the sentence is closer to 100%. Evidence reliably shows that individuals are faster to produce high-cloze words that appear in highly constraining contexts (e.g., *"The children all sang happy birthday before they cut the cake"*) compared to words in low constraint contexts where there are multiple low-cloze alternatives that could fit the sentence context (e.g., *"Shirley was nervous that her guests wouldn't like her cake"*). Such findings are typically considered to reflect the prediction of upcoming words through the pre-activation of lexical features prior to encountering the specific input (DeLong et al., 2014; DeLong et al. 2005; Kutas & Federmeier, 2000). Importantly, evidence indicates that unexpected words from the same semantic category as a predicted term are facilitated relative to unexpected terms from a different semantic category, despite similar cloze probabilities of both (Federmeier et al., 2000). This evidence is taken to suggest that facilitated processing in the cloze task reflects prediction of specific lexical features, rather than the process of integrating words in a particular context.

An important question in this literature is how predictions are generated during the cloze task, and specifically whether individuals make predictions about specific words, or a broader set of potential responses. One theory by Staub and colleagues (2015) suggests that predictions are generated through an activation-based race model, in which several potential responses are activated, and a prediction is generated based on the first response to pass a particular activation

threshold. On this account, multiple potential responses are activated independently of each other. Similar explanations have suggested comprehenders generate predictions by computing the probability of encountering different continuations, and can flexibly shift these probabilities when the bottom-up input confirms or disconfirms the predicted continuation (Delong et al., 2005; Wlotko & Federmeier, 2012). This latter account is well supported by evidence of reduced N400s – an ERP component considered to reflect the brain's neural response to meaningful stimuli – when individuals encounter high-cloze words in a sentence. The N400 effect in language processing reliably shows a reduced N400 amplitude to high-cloze words, and is shown to be inversely correlated with the probability of encountering a specific term (DeLong et al., 2005). Thus, word predictability is often considered to be graded, such that the benefits of predictive processing decrease as cloze-probability approaches zero and accurate predictions are unlikely to be generated (Brothers & Kuperberg, 2021).

Although evidence suggests that processing during the cloze task relies on predictive processing in which prediction occurs in a graded fashion, important questions remain regarding the utility of prediction during language processing. Prediction has been posited as an essential mechanism in language processing, with some theories going as far as to say that prediction is necessary for language comprehension (e.g., Christiansen & Chater, 2016; Kuperberg & Jaeger, 2016; Pickering & Garrod, 2013; but see Huettig & Mani, 2016). However, an investigation by Luke and Christianson (2016) on the predictability of words appearing in 55 different text passages demonstrated that only a small number of words appearing in natural contexts are highly predictable. If prediction is only beneficial in certain potentially rare contexts, then this calls into question the usefulness of prediction as a general mechanism that facilitates language processing. It is therefore possible that prediction is not simply used to facilitate lexical access,

but is also beneficial for other components of linguistic processing such as learning. One influential framework proposed by Dell and Chang (2014) argues that prediction is a central component of language processing because prediction naturally leads to prediction error, which in turn leads to implicit learning. The P-chain model outlined by Dell and Chang is rooted in the idea that processing necessarily involves prediction, but there will be a natural deviation between what is predicted and what is encountered since we rarely encounter exactly what we predict. These prediction errors result in a processing cost that leads the language system to tune its knowledge to reduce errors in the future (leading to what is commonly observed as priming). This account emphasizes the information value of the cost of failed predictions, rather than the benefits of successful predictions, as a critical aspect of linguistic processing. Specifically, disconfirmed prediction may lead to longer-term benefits as the language system adapts to reduce potential prediction errors in the future.

Empirical evidence supporting the claim that prediction failure leads to learning suggests both children and adults adapt their output when they are exposed to syntactic structures that appear in a surprising context (Fazekas et al., 2020). Similarly, individuals show better recognition of unexpected than expected words when their memory is probed following a reading task (Haeuser & Kray, 2021; Hodapp & Rabovsky, 2021). The idea that failed predictions may be beneficial for language processing suggests that, even in the absence of opportunity to generate accurate predictions, predictive processing may still be a useful mechanism. However, for incorrect predictions to benefit language processing, there must be some processing cost that can provide a signal for learning. Although some research has indicated processing costs when words are unexpected within context (e.g., Federmeier et al., 2007; Van Berkum, et al., 2005; Van Petten & Luka, 2012), others have found that there is no

cost for disconfirmed predictions. For example, Luke and Christianson (2016) provided evidence from eye-movements that indicated speakers do not take longer to process unexpected words, suggesting that there is not a significant processing cost when predictions are disconfirmed. Further research by Frisson and colleagues (2017) found that unexpected words were not read more slowly than words appearing in a neutral context, and instead suggest that unexpected but related words may actually be processed faster.

These findings indicate that, although unexpected terms may not incur a processing cost, there is a benefit when an unexpected term is semantically related to the predicted term (see also, Federmeier & Kutas, 1999; Ness & Meltzer-Asscher, 2020; Staub et al., 2015, for evidence suggesting semantically related terms receive facilitation). Notably, both Luke and Christianson (2016) and Frisson et al. (2017) used eye-tracking measures to argue that failed predictions do not result in a processing cost. Although eye-tracking measures are frequently used as an indirect measure of language comprehension, studies that rely purely on eye movements alone may not fully capture the processing costs associated with failed predictions because the measure does not fully capture the difficulty of lexical access. For example, Morgan and colleagues (2020) argue that faster reading times, which are often considered to reflect easier comprehension, are not sufficiently indicative of facilitated processing. Instead, we must also consider other measures of sentence processing.

Overall, the available research is mixed regarding the costs and benefits of prediction. Although it is clear that predicting correctly confers a processing benefit, we are not likely to regularly encounter many highly predictable words in natural contexts. On the other hand, failed predictions may be useful for implicit learning, allowing the language system to adapt to new contexts in the future, but this relies on the notion that failed predictions result in a processing

cost. This debate is central to the question of why we predict at all – is it to speed comprehension or to further tune the language system with flexible input? If we generate predictions to facilitate comprehension, then the costs of failed predictions should be minimal compared to when predictions are not generated, or, at a minimum, predicting almost correctly should result in similar processing times as when predictions are not generated. Alternatively, if prediction is useful for language learning, there should be a clear cost to failed predictions that can be used by the language system to tune its response to future input.

The goal of the present research is to further investigate the potential costs of prediction error to determine the extent to which failed predictions affect processing relative to when predictions cannot be generated. Across three experiments, we manipulate cloze probability, semantic relatedness, and language modality (language production vs. language comprehension) to determine if failed predictions are more costly when the target is related or unrelated to a predicted term, compared to when no prediction is generated in a low constraint sentence. This allows us to consider the costs of failed predictions during both language production and comprehension using the same stimuli and similar methodology, which is especially important given that theories of prediction often encompass both language production and comprehension (e.g., Pickering & Garrod, 2013). In Experiments 1 and 2, we measure whether related and unrelated terms were named more quickly in the low constraint sentences than in the high constraint sentences. Picture naming is a common method used in psycholinguistics to measure language production processes because it requires individuals to fully engage the process of lexical access. In Experiment 3, participants were required to perform a semantic classification of a written word instead of naming a picture. Presenting a written word instead of a picture allowed us to measure comprehension of a given term rather than engaging in word production.

If they are named or classified more quickly in the low constraint sentences, this would suggest there is a cost to prediction failure, such that processing would be faster if individuals had not made predictions at all. Alternatively, if naming onset latencies or classification times are similar in both the high constraint and low constraint conditions, this would indicate that prediction failure does not result in additional processing costs beyond that of lexical access without prediction. Finally, if results are similar in both production and comprehension, this would suggest our conclusions apply to language processing more generally and not just to production or comprehension alone. Thus, the current experiments allow us to determine both whether there is a cost to failed predictions relative to when predictions are not generated, as well as whether the potential costs of failed predictions vary across different contexts, using a task that requires individuals to engage fully in the process of lexical access in both language production and comprehension.

Experiment 1

Method

Participants

One hundred and ten participants (age range = 17-27, M = 18.76, SD = 1.64, 91 female, 17 male, and 2 who did not specify gender) from the McMaster University undergraduate participant pool participated for course credit. Participants had native or native-like proficiency in English and had normal or corrected-to-normal vision. Four additional participants were removed from the study because they failed to follow instructions (n = 1) or because they restarted the experiment partway through (n = 3). Although similar studies often recruit samples of approximately 40 participants (e.g., Frisson et al., 2017; Ness & Meltzer-Asscher, 2018; Staub

et al., 2015), we aimed for a sample of at least 100 participants in each study to account for the fact that the study was conducted online which may result in noisier data.

Materials

Sixty-nine semantically related word pairs were selected based on the University of South Florida Free Association Norms (Nelson et al., 2004). Targets words were also matched for frequency based on the SUBTLEX-US corpus (Brysbeart & New, 2009; average log frequency: 3.10, *SD* = 0.61). A highly constraining sentence was designed for each word in the pair such that the word would be the predicted ending for the sentence. We also constructed a single low constraint sentence for each word pair such that predictions could not be generated from the sentence context. Sentences were either uniquely created for this experiment, or were taken from previously published work (Block & Baldwin, 2010; Gollan et al., 2011; Griffin & Bock, 1998). This resulted in a set of three sentences for each word pair – two high constraint sentences that would end with each word, and one low constraint sentence that could reasonably accommodate both terms as the sentence final word. This yielded a total of 207 sentences (three per word pair). See Table 1 for examples of word pairs and their corresponding sentences.

Table 6.

Original Normed Sentences for Three Related Word Pairs.

Word Pair	Sentence Condition	Sentence
Brain/skull	High Constraint 1	The frontal lobe is an important part of the **brain.**
	High Constraint 2	The pirate's flag displayed cross bones and a **skull.**
	Low Constraint	After the class field trip, the girl told her mom all about the **(brain/skull).**
Apple/orange	High Constraint 1	In the story, Snow White ate a poisoned **apple.**
	High Constraint 2	Every morning, John had a cup of juice squeezed from a fresh **orange.**

	Low Constraint	The artist painted a picture of a single **(apple/orange).**
Cake/pie	High Constraint 1	The children all sang happy birthday before they cut the **cake.**
	High Constraint 2	Thanksgiving dessert consisted of ice cream and a pumpkin **pie.**
	Low Constraint	Shirley was nervous that her guests wouldn't like her **(cake/pie).**

Norming. Sentences were normed to verify the cloze probability of the sentence final words for the high and low constraint sentence for each word pair. A sample of 34 undergraduate participants (age range = 18-26, M = 18.72, SD = 1.60, 30 female, 4 male) who did not participate in the main study were shown sentence fragments with the sentence final word removed. Participants were asked to complete the sentence with the first word that came to mind. Cloze probability was calculated by counting the number of participants that completed the sentence with the same term. High constraint sentences were only selected as stimuli if more that 67% of participants used the same sentence final term, and low constraint sentences were selected if the same term was written by participants less than 20% of the time. These cut-offs are based on Block and Baldwin's (2010) cloze norming study in which they considered sentences to be high constraint if at least 67% of participants completed the sentence with the same word. Final sentence and word pairs were only included if the high constraint sentence for each word in the pair achieved the high constraint threshold. This resulted in 48 word pairs that were used as stimuli in the experiment. See Appendix A for details about each word pair.

Experimental Conditions. Our research question considered whether participants would respond faster to the sentence final word of a highly constraining sentence if it was replaced with a term that was related to the high cloze ending (high constraint (HC): small prediction error condition), a term that was unrelated to the high cloze ending (high constraint (HC): large

prediction error condition), or if the term appeared following a low constraint sentence (low constraint (LC): no prediction condition). Sentences were counterbalanced across six lists (16 sentences in each condition) such that each sentence and each word appeared in every condition. To create the HC: small prediction error condition, the high cloze sentence final word was replaced with the related word from the pair. For example, given the high constraint sentence *'The frontal lobe is an important part of the'*, where the high cloze ending is *'brain'*, the final word was replaced with the related term *'skull'*. In this example, since the high cloze ending *'brain'* is replaced with *'skull'*, the term *'brain'* is not used. We used the removed high cloze endings from the HC: small prediction error condition as sentence final endings in the HC: large prediction error condition. For example, given the high constraint sentence *'In the story, Snow White ate a poisoned'*, where the high cloze ending is *'apple'*, the high cloze ending was replaced with leftover terms from the HC: small prediction error condition (in this example, *'brain'*). Finally, the LC: no prediction condition was created by using one term in the word pair as the final word for the low constraint sentence.[4] See Table 2 for an example of counterbalancing across the six lists using the example sentences in Table 1.

Table 7.

Example of counterbalancing using the stimuli displayed in Table 1.

List 1	List 2

[4] We note that the premise of our experiments assumes that individuals do not generate predictions of individual words in low constraint sentences. Based on the proportional pre-activation account of prediction (Brothers & Kuperberg, 2021), individuals pre-activate words in proportion to their estimated likelihood. In low constraint contexts, the number of potential words that can be used to fit a sentence will be quite large, meaning the likelihood of encountering any single word will be low. It is therefore unlikely that the pre-activation of any specific set of lexical features will be sufficiently strong to predict any particular word. Conversely, the likelihood of encountering a specific word in a high constraint context is quite high which allows for greater pre-activation of specific words. We assume that this difference in pre-activation in low and high constraint contexts will result in strong predictions of specific high-cloze words in the high constraint contexts compared with limited pre-activation of a larger set of lexical features in the LC: no prediction condition that are not sufficiently strong to generate specific predictions.

Condition	Sentence	Target	Sentence	Target
HC: small prediction error	The frontal lobe is an important part of the	skull	The pirate's flag displayed a cross bones and	brain
HC: large prediction error	In the story, Snow White ate a poisoned	brain	Every morning, John had a cup of juice squeezed from a fresh	skull
LC: no prediction	Shirley was nervous that her guests wouldn't like her	cake	Shirley was nervous that the guests wouldn't like her	pie
	List 3		**List 4**	
Condition	Sentence	Target	Sentence	Target
HC: small prediction error	The children all sang happy birthday before they cut the	pie	Thanksgiving dessert consisted of ice cream and pumpkin	cake
HC: large prediction error	The frontal lobe is an important part of the	cake	The pirate's flag displayed a cross bones and	pie
LC: no prediction	The artist painted a picture of a single	apple	The artist painted a picture of a single	orange
	List 5		**List 6**	
Condition	Sentence	Target	Sentence	Target
HC: small prediction error	In the story, Snow White ate a poisoned	orange	Every morning, John had a cup of juice squeezed from a fresh	apple
HC: large prediction error	The children all sang happy birthday before they cut the	apple	Thanksgiving dessert consisted of ice cream and pumpkin	oranges
LC: no prediction	After the class field trip, the girl told her mom all about the	brain	After the class field trip, the girl told her mom all about the	skull

Sentences for word pair *'brain/skull'*; Sentences for word pair *'apple/orange'*; Sentences for word pair *'cake/pie'*

Procedure

All experiments presented here were programmed in PsychoPy (Peirce et al., 2019), and administered online through Pavlovia.org. Sentences were displayed one word at a time on the screen for 500 ms and participants were instructed to read each word aloud as it appeared. The final word in the sentence was replaced by a picture that participants were instructed to name. The same pictures were used across the experiment for each to-be-named word, with the exception that if a sentence required the term to be pluralized, the picture was altered to show

multiple instances of the object. Because data were collected remotely, participants pressed the space key when they produced the word. Participants were then asked to type out the name of the picture to verify that they named the picture correctly. Target sentences were randomly ordered for each participant. The entire experiment took approximately 20 minutes to complete. This study was approved by McMaster University's Research Ethics Board (Project ID: 5109).

Data Preparation

Responses were checked to confirm that pictures were named correctly. To ensure that participants accessed the target word at the time of response, only exact matches were included. For example, if participants produced an over-specified response (e.g., *'black circle'* for *'circle'*), or did not provide a typed response to the image, the trial was excluded. We additionally excluded trials where participants provided an incorrect name for the target image. Spelling errors and plurals were included, along with clear synonyms. This resulted in the removal of 7.4% of trials. (7.3% of all HC: large prediction error trials, 8.5% of all HC: small prediction error trials, and 6.4% of all low constraint trials). The average number of errors, based on the number of incorrect responses provided by participants in each condition, for all experiments can be found in Table 3. Note that we did not include over-specified responses in the count of errors since these are not incorrect per se. Similarly, we did not include trials in which no label for the target image was provided in the error count as we cannot be sure which term the participant accessed when they saw the image.

After removing trials where pictures were not correctly named, we additionally removed trials where reaction times were longer than 4000 ms or shorter than 250 ms (2% of remaining trials). Remaining trials that were greater than three standard deviations from the overall mean were removed (2.3% of remaining trials). This resulted in a final sample of 4694 trials (1599 in

the LC: no prediction condition, 1545 in the HC: small prediction error condition, 1550 in the HC: large prediction error condition). The full data set, analysis files, and experimental materials for all experiments reported here are available through the Open Science Framework (https://osf.io/wsf8c/?view_only=02c5ceee0b6646b0a79737261febc58e).

Table 8.

Average Number of Errors by Condition (n =16 items/condition).

	Condition		
	HC: small prediction error *M* (*SD*)	HC: large prediction error *M* (*SD*)	LC: no prediction *M* (*SD*)
Experiment 1	1.25 (1.45)	1.05 (1.20)	0.81 (1.00)
Experiment 2	1.57 (1.68)	1.31 (1.75)	0.93 (1.06)
Experiment 3	0.60 (1.06)	1.24 (1.50)	0.57 (1.00)

Results

Statistical analyses were carried out using R, Version 4.2.0. A linear mixed effects regression analysis was carried out using the lme4 (Bates et al, 2015) and lmerTest (Kuznetsova et al, 2015) packages with reaction times as the dependent variable and condition as the predictor variable. Condition was contrast coded with the HC: small prediction error condition as the reference group. The first contrast compared the HC: small prediction error condition and the HC: large prediction error condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: .5, Low Constraint: 0), and the second contrast compared the HC: small prediction error condition with the LC: no prediction condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: 0, Low Constraint: .5). Participant and target were included as random effects, and contrast conditions were included as random slopes. We

constructed a first model that allowed for correlated random slopes, and a second model that included decorrelated random slopes. Model comparison revealed that the model with correlated slopes was a better fit for the data (χ^2 = 17.49, p = .008). We additionally checked for random slopes that account for less than 1% of the variance in the models using the rePCA function in R (Bates et al., 2015). This revealed that a model including only the first contrast as a by-participant and by-item random slopes was appropriate.

As can be seen in Figure 1, participants were fastest to name sentence-final pictures in the LC: no prediction condition (*M* = 943 ms, *SD* = 314 ms), followed by the HC: small prediction error condition (*M* = 987 ms, *SD* = 365 ms), and finally the HC: large prediction error condition (*M* = 1028 ms, *SD* = 360 ms). Therefore, although related words were produced faster than unrelated words, they still did not benefit from appearing in a high constraint sentence frame, thus indicating a clear cost of failed predictions. This conclusion is supported by the statistical comparisons that revealed a significant effect of relatedness (HC: small prediction error vs. HC: large prediction error: B = 65.07, *SE* = 13.23, *t*(89.51) = 4.92, *p* = <.001), as well as a significant effect of constraint (HC: small prediction error vs. Low Constraint: B = -56.35, *SE* = 11.94, *t*(91.56) = -4.72, *p* = <.001).

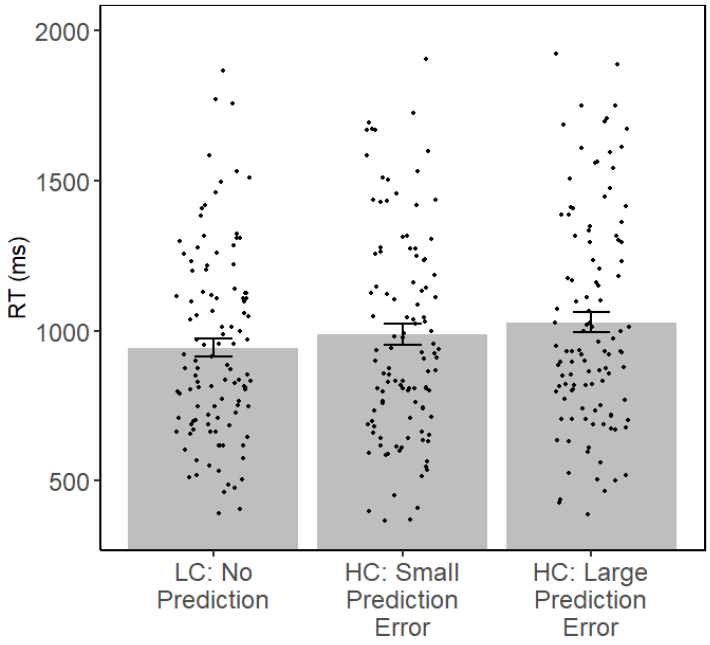

Figure 8. Average reaction times by condition for Experiment 1. Error bars represent the standard error for each condition.

Discussion

The results of Experiment 1 revealed that when a predicted ending is replaced with a related term, it is accessed faster than when it is replaced with an unrelated term, but that related term is not produced more quickly than when it appears after a low-constraint sentence. Therefore, these results demonstrate a clear processing cost for disconfirmed predictions, indicating that although there is a small benefit to predicting almost correctly, it is not better than if nothing had been predicted at all.

One potential limitation of this conclusion is that expected endings were never present across the experiment. It is therefore possible that participants did not commit as strongly to their predictions if they realized they would never be confirmed. As outlined in the introduction, research suggests that the experimental context can influence predictive processing, such that individuals do not generate strong predictions in contexts where the validity of their predictions is low (Brothers et al., 2017). To determine whether the experimental context influenced the

results of Experiment 1, we conducted a second experiment that followed the same procedure, with the exception that additional high constraint sentences that were completed with their expected ending were included as filler trials throughout the experiment. This was done to encourage predictive processing, and allow us to determine whether the costs of failed predictions differ when top-down processing emphasizes prediction.

Experiment 2

Method

Participants

One hundred and fourteen participants (age range = 17-41, *M* = 18.99, *SD* = 3.62, 91 female, 20 male, and 3 who did not specify gender) from the McMaster University undergraduate participant pool, who did not participate in the first experiment or the norming study, participated for course credit. Participants had native or native-like proficiency in English, and had normal or corrected-to-normal vision.

Materials

The experimental materials and conditions in this experiment were identical to those in Experiment 1. We additionally included 32 high cloze filler trials that were taken from the unused stimuli in each list. For example, given the sentences used in List 1 displayed in Table 2, the high constraint sentences *'Thanksgiving dessert consisted of ice cream and a pumpkin'* and *'Every morning, John had a cup of juice squeezed from a fresh'* and their respective sentence completions *'pie'* and *'orange'* do not appear anywhere in the experimental list. Therefore, they can be used as high cloze fillers without repeating stimuli across the experiment. The 32 high cloze fillers were added to each list in a pseudo-random order, with the stipulation that a high cloze filler appeared at least every third trial.

Procedure

The procedure was the same as Experiment 1.

Data Preparation

Data were cleaned following the same guidelines as in Experiment 1. We first removed trials where participants did not provide an exactly accurate picture name (9.0% of all experimental trials; 9.4% of all HC: large prediction error trials, 11.0% of all HC: small prediction error trials, and 6.6% of all low constraint trials). The average number of errors for each experimental condition can be found in Table 3. An additional 1.9% of remaining trials were removed after trimming trials that were greater than 4000 ms or less than 250 ms, and then removing trials that were greater than 3 SDs from the overall mean. This resulted in a final sample of 4785 trials (1653 in the LC: no prediction condition, 1561 in the HC: small prediction error condition, 1571 in the HC: large prediction error condition).

Results

Statistical models were constructed the same way as in Experiment 1, using reaction time as the dependent variable and condition as the predictor variable. Contrast codes for the conditions were also done the same way as in Experiment 1, with the first contrast comparing the HC: small prediction error condition and the HC: large prediction error condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: .5, Low Constraint: 0), and the second contrast comparing the HC: small prediction error condition with the LC: no prediction condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: 0, Low Constraint: .5). Participant and target were included as random effects, and contrast conditions were included as random slopes. We constructed a first model that allowed for correlated random slopes, and a second model that included decorrelated random slopes. Model comparison

revealed that the model with correlated slopes was a better fit for the data (χ^2 = 30.56, p < .001). As in Experiment 1, we checked which slopes accounted for less than 1% of the variance in the models. This analysis determined that the by-participant random slopes for the first contrast were not necessary and was thus removed.

Figure 2 shows the average reaction times for each experimental condition, as well as the average reaction time for the high constraint filler trials for comparison. Although the high constraint fillers were not included in the main analyses, we note that reaction times were fastest for these trials (*M* = 847 ms, *SD* = 274 ms), confirming that accurate prediction confers a processing benefit. Considering the conditions of interest, reaction times were fastest in the LC: no prediction condition (*M* = 925 ms, *SD* = 321 ms), followed by the HC: small prediction error condition (*M* = 949 ms, *SD* = 374 ms), and finally the HC: large prediction error condition (*M* = 993 ms, *SD* = 386 ms). The average response times in each condition are consistent with those of Experiment 1, confirming that there is a cost to failed predictions. This conclusion is again supported by statistical analysis, revealing a significant effect of HC: small prediction error vs. HC: large prediction error, B = 69.63, *SE* = 21.14, *t*(81.51) = 3.29, *p* = .001, as well as a significant effect of HC: small prediction error vs. Low Constraint, B = -70.53, *SE* = 20.07, *t*(78.11) = -3.51, *p* <.001. Although the results of Experiment 2 follow the same pattern as those in Experiment 1, it should be noted that the difference in reaction times between the HC: small prediction error and LC: no prediction conditions is much smaller in Experiment 2 (HC: small prediction error – Low Constraint: 24 ms in Experiment 2 vs. 44 ms in Experiment 1), whereas the difference between the HC: small prediction error and HC: large prediction error conditions was similar across experiments (HC: large prediction error – HC: small prediction error: 44 ms in Experiment 2 vs. 41 ms in Experiment 1). Therefore, although the costs of mis-prediction are

still apparent despite the inclusion of high cloze fillers to encourage prediction across the experiment, the average reaction times suggest that the cost of producing a related term in place of a predicted term is lower when the global context encourages predictive processing.

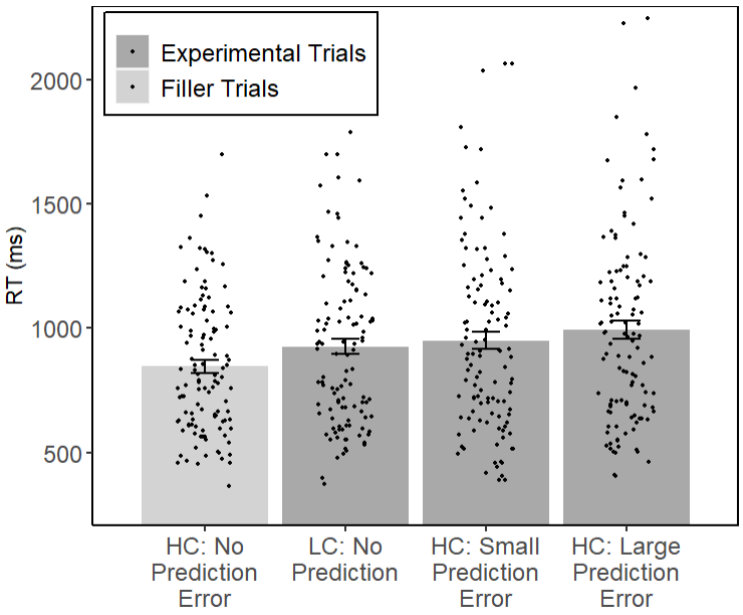

Figure 9. Average reaction times by condition for Experiment 2. Error bars represent the standard error for each condition.

Discussion

The results of the first two experiments revealed that related terms are accessed faster than unrelated terms when they appear in place of an expected ending, but neither are produced faster than when the same words appear in a low constraint sentence. Critically, when prediction was encouraged in Experiment 2, predicting almost correctly still did not result in faster processing than when the same words appeared in a low constraint context. This suggests that, although accurate predictions confer a benefit, incorrect predictions do not provide a benefit relative to not predicting, regardless of whether the sentence final word is related or unrelated to the predicted term.

These first two experiments focused on whether word production benefits from prediction in cases where predictions are not confirmed. However, a significant portion of research on prediction has focused on how predictive processing benefits language comprehension. Therefore, we conducted a final experiment to examine whether the same words that were produced in Experiments 1 and 2 are processed faster when they appear in highly constraining sentences when they are related to the predicted term, relative to when they are unrelated or when they appear in a low constraint sentence during language comprehension.

Experiment 3

Method

Participants

One hundred and three participants (age range = 17-52, *M* = 18.66, *SD* = 3.47, 90 female, 12 male, and 1 who did not specify gender) from the McMaster University undergraduate participant pool, who did not participate in the previous two experiments or the norming study, participated for course credit. Participants reported native or native-like proficiency in English, and had normal or corrected-to-normal vision. Six other participants were removed from the analysis because they did not complete the entire experiment (*n* = 3) or because they restarted the experiment partway through (*n* = 3).

Materials

The materials were identical to those used in Experiment 2, with the exception that instead of the sentence-final word being replaced with a picture, the final word was displayed in capital letters and in red font.

Procedure

The procedure was the same as the previous two experiments, except that instead of naming a picture, participants were asked to classify the sentence-final word as being an object that is typically found in one's house or not. Participants were instructed that they would read sentences one word at a time off the screen, and when they encountered a word in all caps in red font color, they should press 'k' if the word referred to an object that is typically found in one's house, or 'd' if it is an object that is not typically found in one's house. Of the target words, 56.25% referred to objects that are typically found in one's house, 10.42% were objects that are not typically found in one's house, and 33.33% could reasonably be classified as either. See Appendix A for details about how each target was classified.

Prior to the start of the experimental trials, participants completed 12 practice trials where they classified individual words as objects that are typically found in one's house, and 6 practice trials where they read a low constraint sentence in which the sentence final word was presented in all caps and red font color and they were asked to classify the final word as an item that is typically found in one's house. Participants received feedback during all practice trials to provide guidance on how to classify objects. For example, for words such as *'tiger'*, they were told to interpret these as the actual object, not a related object such as a tiger appearing in a painting or a stuffed animal. For objects such as *'teeth'*, they were instructed to include the object in any capacity, such as the teeth in one's mouth which are inside the house. Following the practice, participants completed 80 experimental trials presented in a pseudo-random order, 32 of which were high constraint fillers appearing at least every third trial, and feedback was not provided. The experiment was programmed in PsychoPy (Pierce et al., 2019), and administered online through Pavlovia.org.

Data Preparation

Data were cleaned following the same guidelines as in Experiment 1 and 2, with the exception that instead of removing trials where the images were labelled incorrectly, we removed trials in which participants incorrectly classified the target word. This resulted in removing 5.0% of all experimental trials due to misclassification errors (7.8% of all HC: large prediction error trials, 3.8% of all HC: small prediction error trials, and 3.6% of all low constraint trials). Of the errors removed due to misclassification of target, 32.8% were objects typically found in one's house, and 67.2% were objects not typically found in one's house. The average number of errors for each experimental condition can be found in Table 3. An additional 3.1% of remaining trials with reaction times that were greater than 4000 ms or less than 250 ms, or greater than 3 SDs from the mean after the data were trimmed. This resulted in a final sample of 4551 trials (1550 in the LC: no prediction condition, 1537 in the HC: small prediction error condition, 1464 in the HC: large prediction error condition).

Results

Statistical models were constructed the same way as in Experiments 1 and 2, using reaction time as the dependent variable and condition as the predictor variable. Contrast codes for the conditions were also done the same way as in Experiments 1 and 2, with the first contrast comparing the HC: small prediction error condition and the HC: large prediction error condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: .5, Low Constraint: 0), and the second contrast comparing the HC: small prediction error condition with the LC: no prediction condition (contrast codes: HC: small prediction error: -.5, HC: large prediction error: 0, Low Constraint: .5). Participant and target were included as random effects, and contrast conditions were included as random slopes. We constructed a first model that allowed for

correlated random slopes, and a second model that included decorrelated random slopes. Model comparison revealed that the model with correlated slopes was a marginally better fit for the data (χ^2 = 10.65, p = .010). To avoid overfitting the model, we then determined which random slopes accounted for less than 1% of variance in the full model. This revealed that by-participant random slopes were not required.

Figure 3 shows the average reaction times for each condition. As in Experiment 2, reaction times were fastest for the high constraint filler trials (*M* = 784 ms, *SD* = 167 ms). The filler trials were not analyzed further. Considering the conditions of interest, reaction times were fastest in the LC: no prediction condition (*M* = 832 ms, *SD* = 177 ms), followed by the HC: small prediction error condition (*M* = 862 ms, *SD* = 195 ms), and finally the HC: large prediction error condition (*M* = 901 ms, *SD* = 188 ms). The average response times in each condition are consistent with those of Experiment 1 and 2, confirming that there is a cost to failed predictions in comprehension as well as production. This conclusion is again supported by statistical analysis, revealing a significant effect of HC: small prediction error vs. HC: large prediction error, B = 65.07, *SE* = 13.23, *t*(89.51) = 4.92, *p* <.001, as well as a significant effect of HC: small prediction error vs. Low Constraint, B = -56.35, *SE* = 11.94, *t*(91.56) = -4.72, *p* <.001.

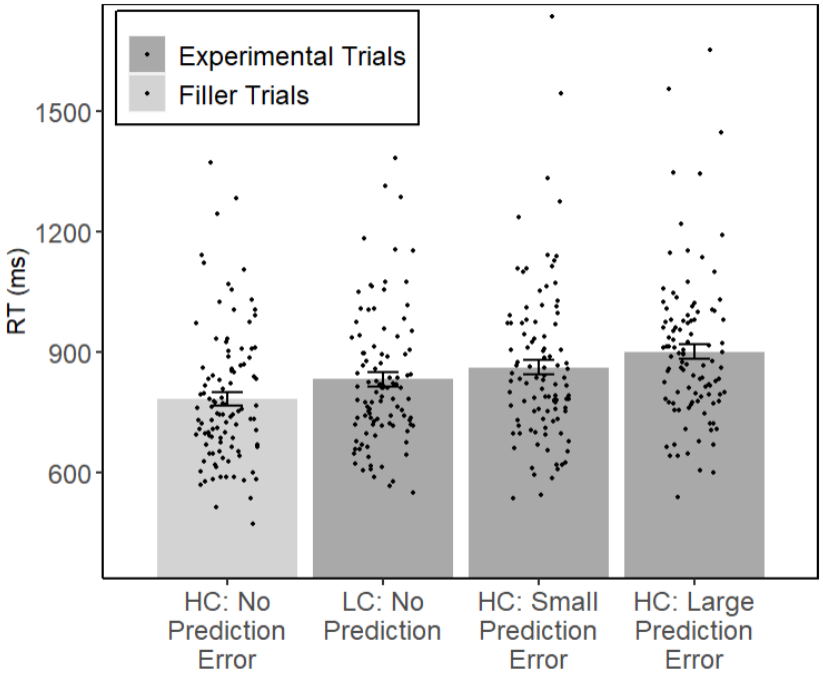

Figure 10. Average reaction times by condition for Experiment 3. Error bars represent the standard error for each condition.

Discussion

Experiment 3 examined whether the costs of failed predictions show the same pattern of results in a comprehension task relative to the production methods used in Experiment 1 and 2. Consistent with our previous results, participants were faster to classify sentence final terms in highly constraining contexts when the term was related to the predicted ending compared to unrelated terms. Critically, however, participants classified terms that appeared at the end of a low constraint sentence more quickly than both unexpected but related and unexpected but unrelated terms that appeared following a high constraint sentence. These results extend our previous findings to a comprehension task, and indicate that there is a clear cost to failed prediction during both language production and comprehension.

General Discussion

Across three experiments, we investigated whether predicting almost correctly in a highly constraining context would be better than predicting completely incorrectly, or better than not predicting at all in a low constraint context. Experiment 1 revealed that, although predicting almost correctly resulted in faster responses than predicting completely incorrectly, neither of these was better than not predicting at all in a low constraint sentence. Experiment 2 extended these results by presenting the same sentences in a context that encouraged prediction through the inclusion of high constraining filler sentences paired with highly predictable endings. The results of Experiment 2 mirrored those of Experiment 1, once again revealing that individuals were always fastest to produce words in a low constraint context where predictions cannot be generated, relative to producing words in a highly constraining sentence where the final word is either related or unrelated to the predicted term. Finally, Experiment 3 investigated whether the above findings could be replicated in a comprehension task where participants were asked to perform a semantic classification of the sentence final word rather than producing the word. Once again participants were faster to classify sentence final words in highly constraining sentences when they were related to the predicted term relative to when they were unrelated to it, but were still faster to classify the words when they appeared in a low constraint sentence where predictions are unlikely to be generated. Overall, across three experiments, we show that, although predicting almost correctly is better than predicting completely incorrectly, incorrect predictions never result in faster processing times relative to when words are not predicted at all. This suggests that although generating predictions during both language production and comprehension is useful when predictions are confirmed, generating predictions results in costs

when predictions are not confirmed, regardless of whether the unpredicted term is related or unrelated to the predicted term.

These results speak to a key issue in theories that argue prediction plays an important role in language comprehension – namely, is there an overall benefit to predictive processing? Consistent with numerous other studies (e.g., Boudewyn et al., 2015; DeLong et al., 2005; Griffin & Bock, 1998; Hintz et al., 2016; Vainio et al., 2009), Experiments 2 and 3 revealed a processing benefit when predictions were confirmed, as indicated by faster response times in the highly constraining filler sentences relative to the LC: no prediction condition. Additionally, unexpected, related words were processed faster than unexpected, unrelated words, which is in line with other studies that showed words that are semantically related to an expected ending were processed more easily than those that are completely unrelated (Federmeier & Kutas, 1999; Ness & Meltzer-Asscher, 2020; Staub et al., 2015). However, given that highly predictable words are relatively rare in natural language (e.g., Luke & Christianson, 2016), the critical question is whether there is still a benefit to predicting in situations where predictions in general are disconfirmed. Although previous studies have found that incorrect predictions do not incur a processing cost (e.g., Frisson et al., 2017; Luke & Christianson, 2016), here we reliably show that incorrect predictions come at a cost that is greater than if predictions were not generated at all, even when the word they encounter is semantically related to the predicted term.

The finding that incorrect predictions reliably result in a processing cost during both production and comprehension suggests that prediction does not consistently speed language processing. If prediction only confers a processing benefit when predictions are exact – a phenomenon that is relatively rare in natural language (Luke & Christianson, 2016) – what is the utility of prediction during language processing? According to Dell and Chang's P-chain theory,

prediction is useful for language learning in the long-run, rather than for short-term processing benefits. They argue that prediction naturally leads to prediction error, which is used to tune the knowledge of the system to reduce future errors. Our results demonstrate a clear cost to failed predictions, even in cases where prediction errors are small, which is in line with the idea that prediction naturally leads to prediction error. Although our experiments were not designed to test the second part of the P-chain theory – prediction errors lead to implicit learning – a closer examination of the data from Experiment 1 suggests that reaction times decreased over the course of the experiment, indicating that participants were able to adapt to the unexpected input in both high and low constraint sentences. Specifically, the correlations between trial number and reaction time were significant in the HC: large prediction error condition, $r = -.57$, $p < .001$, and LC: no prediction condition, $r = -.50$, $p < .001$, and were marginally significant in the HC: small prediction error condition, $r = -.24$, $p < .10$; see Figure 4).[5] These correlations suggest that performance did improve over the course of the experiments, especially in the HC: large prediction error and LC: no prediction conditions. The correlation between trial number and reaction time was only marginally significant in the HC: small prediction error condition, but this may be because the closer semantic relationship between the predicted ending and the target provided a slight boost in activation that was not possible in the HC: large prediction error and LC: no prediction conditions. Further, it is difficult to determine whether the decrease in reaction times over the course of the experiment is due specifically to learning, or task adaptation more generally. Future research will be needed to specifically compare the effect of processing costs on implicit learning in language.

[5] Trials in Experiments 2 and 3 followed a fixed random order to ensure that the HC filler trials appeared at least every third trial. This makes it difficult to determine whether the relationship between trial order and reaction times is these studies is consistent with negative relationship found in Experiment 1. This would need to be confirmed with a study using a different random order for every participant so that each trial could occur in each position.

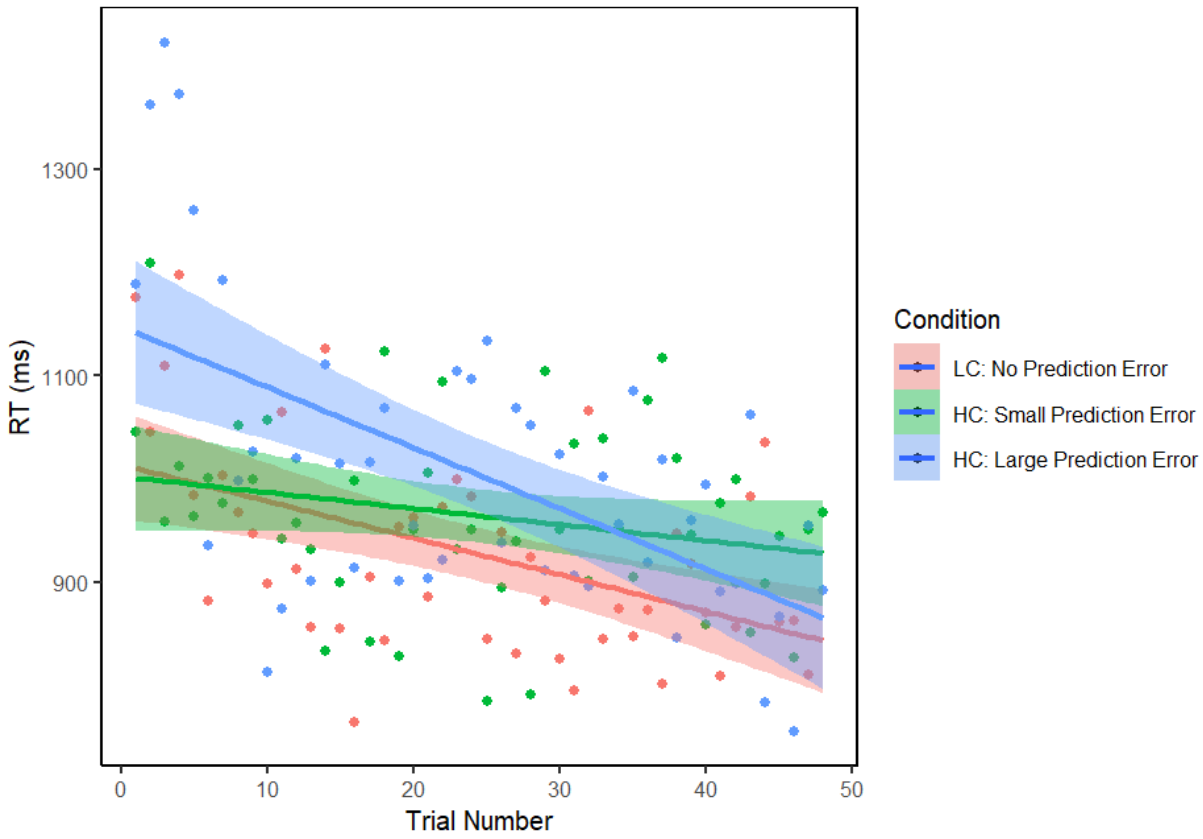

Figure 11. Reaction Times for Each Condition by Trial Number for Experiment 1.

The studies presented in this paper address the processing costs of incorrect predictions in both language production (Experiments 1 and 2) and comprehension (Experiment 3). The data showed the same patterns of results across both language modalities. Finding the same pattern in both modalities does not necessarily mean that prediction in both comprehension and production relies on the same mechanisms. However, several theories of prediction in language argue that prediction relies on production mechanisms (e.g., Dell & Chang, 2013; Martin et al., 2018; Pickering & Garrod, 2013; Pickering & Garrod, 2007; Rommers et al., 2020). The P-chain theory in particular makes the strong claim that prediction *is* production. Similarly, Martin and colleagues (2018) found that the N400 effect in prediction was reduced when language production was interrupted through articulatory suppression. Although the current experiments are not set up to determine whether prediction relies on language production, these theories could

possibly explain why we find similar effects of processing costs for failed predictions in both comprehension and production.

Our conclusion that that there is a cost to predicting incorrectly relative to when people do not generate predictions at all relies on the assumption that participants were not making predictions in the LC: no prediction condition. Although we selected low constraint sentences in which the sentence was not completed with the same word more than 20% of the time in the norming phase (and that none of the words generated by participants in the norming phase were those used in the experimental condition), it is possible than individuals could still generate predictions in these contexts. However, previous studies have suggested that prediction during language processing is graded, rather than all-or-nothing (e.g., Brothers & Kuperberg, 2021; Boudewyn et al., 2015; Szewczyk & Federmeier 2022; Luke & Christianson, 2016), indicating that individuals are unlikely to generate predictions of individuals words in low constraint contexts. Based on this research, it is possible that participants may generate low level predictions in the LC: no prediction condition in the current experiment, such as anticipating the part of speech or morphology, but it is unlikely that participants would be able to generate specific predictions about the upcoming word without encountering a highly constraining sentence. Additionally, if predictions are generated in the low constraint condition, they are very likely to be wrong. Given the clear cost of incorrect predictions in the high constraint contexts, if participants were generating predictions in the low constraint condition we would expect to see a similar processing cost. However, targets were processed faster in the low constraint condition than in the HC: small prediction error and HC: large prediction error conditions, suggesting that words are either not predicted in the low constraint context, or the processing cost is significantly smaller than in high constraint sentences. Thus, we can conclude that participants made strong

predictions in the high constraint conditions, whereas they were unlikely to predict specific words in the LC: no prediction condition.

One potential limitation of this conclusion is that, because of the nature of the experimental design, some of the target sentences might be considered implausible, which could result in slowed reaction times. However, as part of our stimulus norming procedures we collected plausibility ratings of each target sentence, allowing us to investigate the impact of sentence plausibility on reaction times. The correlations between sentence plausibility and reaction times within each condition were mostly nonsignificant, with some exceptions. In just Experiments 1 and 2, responses were faster for target words that were considered more plausible, as evidenced by significant negative correlations between plausibility and reaction times in the LC: no prediction condition, $r = -.44$, $p = .011$ and $r = -.26$, $p < .011$ in Experiments 1 and 2, respectively. This suggests that sentence plausibility primarily has a relationship with the production of words that cannot be predicted based off of the sentence context. Therefore, the results suggest that language processing is slowed when predictions are disconfirmed, rather than slowed because of the plausibility of the targets.

A final consideration concerns how the present data relate to current theories regarding the mechanisms of prediction in language processing. It is often argued that predictions are generated by computing the probability of encountering multiple potential alternatives (e.g., Delong et al., 2005; Wlotko & Federmeier, 2012), and words that are more likely within a context receive a greater level of pre-activation (e.g., Staub et al., 2015). This means that words that are less likely to appear within a given context must receive a greater level of activation once they are encountered to pass a certain activation threshold, thus slowing lexical access. Our results confirm that when one encounters a term that is a greater semantic distance from a

predicted term, lexical access is slower relative to when a target is more closely related to the predicted term. Thus, the present data support a framework in which words are pre-activated based on their likelihood of being encountered.

Overall, the results presented here demonstrate that when individuals make incorrect predictions, their responses are much slower than if they had not predicted at all. Further, even when predictions are almost correct, they are still not faster than if they had not generated a prediction. The results limit the degree to which prediction can be used to speed language processing. Instead, to the extent that people predict, it may be to tune their language knowledge, making smalls costs in terms of processing times worthwhile. In short, if we predict, we are likely to predict to learn, not to understand faster.

References

Altmann, Gerry TM, and Yuki Kamide. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* 73.3 (1999): 247-264. https://doi.org/10.1016/S0010-0277(99)00059-1

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*

Bates, D., Maechler, B., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, *42*(3), 665-670. https://doi.org/10.3758/brm.42.3.665

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607-624. https://doi.org/10.3758/s13415-015-0340-0

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. https://doi.org/10.1016/j.neuropsychologia.2019.107225

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of memory and language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of memory and language*, *93*, 203-216. https://doi.org/10.1016/j.jml.2016.10.002

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990. https://doi.org/10.3758/brm.41.4.977

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, *39*. https://doi.org/10.1017/s0140525x1500031x

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, *7*(11), 180877. https://doi.org/10.31234/osf.io/3phxu

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469-495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain research*, *1146*, 75-84. https://doi.org/10.1016/j.brainres.2006.06.101

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current directions in psychological science*, *27*(6), 443-448. https://doi.org/10.1177/0963721418794491

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200-214. https://doi.org/10.1016/j.jml.2017.04.007

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: the frequency-lag hypothesis. *Journal of Experimental Psychology: General*, *140*(2), 186. https://doi.org/10.1037/e520602012-200

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313-338. https://doi.org/10.1006/jmla.1997.2547

Haeuser, K. I., & Kray, J. (2021). Effects of prediction error on episodic memory retrieval: evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, 1-17. https://doi.org/10.1080/23273798.2021.1924387

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. *The Quaterly Journal of Experimental Psychology*, *69*(6), 1056-1063. http://dx.doi.org/10.1080/17470218.2015.1131309

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82,* 1–26. https://doi.org/10.18637/jss.v082.i13

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific reports*, *8*(1), 1-9. https://doi.org/10.1038/s41598-018-19499-4

Morgan, A. M., von der Malsburg, T., Ferreira, V. S., & Wittenberg, E. (2020). Shared syntax between comprehension and production: Multi-paradigm evidence that resumptive pronouns hinder comprehension. *Cognition*, *205*, 104417. https://doi.org/10.1016/j.cognition.2020.104417

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407. https://doi.org/10.3758/bf03195588

Ness, T., & Meltzer-Asscher, A. (2020). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, 104509. https://doi.org/10.1016/j.cognition.2020.104509

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347. https://doi.org/10.1017/s0140525x12001495

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension?. *Trends in cognitive sciences*, *11*(3), 105-110. https://doi.org/10.1016/j.tics.2006.12.002

Rommers, J., Dell, G. S., & Benjamin, A. S. (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition*, *198*, 104206. https://doi.org/10.1016/j.cognition.2020.104206

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1-17. https://doi.org/10.1016/j.jml.2015.02.004

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of memory and language*, *123*, 104311. https://doi.org/10.1016/j.jml.2021.104311

Vainio, S., Hyönä, J., & Pajunen, A. (2009). Lexical predictability exerts robust effects on fixation duration, but not on initial landing position during reading. *Experimental psychology*, *56*(1), 66. https://doi.org/10.1027/1618-3169.56.1.66

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading

times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443. https://doi.org/10.1037/0278-7393.31.3.443

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

APPENDIX A

Word Pair		Frequency		Syllables		Word Association		Semantic Classification	
Word #1	Word #2	Word #1	Word #2	Word #1	Word #2	Forward	Backward	Word #1	Word #2
apple	orange	3.08	3.06	2	2	0.17	0.08	Inside one's house	Inside one's house
finger	toe	3.27	2.81	2	1	0.09	0.03	Inside one's house	Inside one's house
circle	square	3.04	3.21	2	1	0.64	0.47	Either	Either
dress	pants	3.65	3.48	1	1	0.22	0.01	Inside one's house	Inside one's house
trees	plants	3.52	3.15	1	1	0.02	0.07	Outside one’s house	Inside one's house
truck	car	3.57	4.39	1	1	0.26	0.11	Either	Either
airplane	helicopter	2.75	2.91	2	4	0.02	0.07	Outside one's house	Outside one's house
dog	cat	3.99	3.53	1	1	0.67	0.51	Either	Either
knife	fork	3.38	2.65	1	1	0.33	0.37	Inside one's house	Inside one's house
wine	beer	3.49	3.59	1	1	0.19	0.09	Either	Either
nose	eye	3.55	3.76	1	1	0.03	0.07	Inside one's house	Inside one's house
hammer	axe	2.80	4.31	2	1	0.03	0.10	Either	Either
guitar	piano	2.90	3.10	2	3	0.08	0.02	Either	Either
plate	bowl	3.12	3.04	1	1	0.04	0.05	Inside one's house	Inside one's house
window	door	3.64	4.17	2	1	0.15	0.16	Inside one's house	Inside one's house
chair	table	3.40	3.73	1	2	0.31	0.76	Inside one's house	Inside one's house
glove	scarf	2.71	2.38	1	1	0.02	0.00	Inside one's house	Inside one's house

computer	printer	3.48	2.03	3	2	0.03	0.46	Inside one's house	Inside one's house
spider(s)	fly	2.71	3.64	2	1	0.02	0.00	Either	Either
lipstick	mascara	2.65	1.81	2	3	0.01	0.03	Inside one's house	Inside one's house
crackers	cheese	2.46	3.30	2	1	0.14	0.17	Inside one's house	Inside one's house
brain	skull	3.59	2.88	1	1	0.03	0.13	Either	Either
cake	pie	3.36	3.17	1	1	0.07	0.17	Inside one's house	Inside one's house
pancakes	waffles	2.69	2.12	2	2	0.13	0.25	Inside one's house	Inside one's house
hat	coat	3.52	3.33	1	1	0.10	0.16	Inside one's house	Inside one's house
shark	whale	2.88	2.76	1	1	0.01	0.04	Outside one's house	Outside one's house
candy	chocolate	3.26	3.18	2	3	0.05	0.20	Inside one's house	Inside one's house
watch	clock	4.23	3.48	1	1	0.12	0.09	Inside one's house	Inside one's house
butter	milk	3.02	3.34	2	1	0.06	0.02	Inside one's house	Inside one's house
eggs	bacon	3.29	2.78	1	2	0.18	0.52	Inside one's house	Inside one's house
crayons	marker	1.79	2.43	2	2	0.08	0.08	Either	Either
bed	pillow	3.98	2.76	1	2	0.02	0.13	Inside one's house	Inside one's house
banana(s)	strawberry(ies)	2.74	2.45	3	3	0.01	0.01	Inside one's house	Inside one's house
bread	sandwich	3.16	3.05	1	2	0.03	0.07	Inside one's house	Inside one's house
brick	wood	2.72	3.14	1	1	0.01	0.01	Either	Either
bridge	tunnel	3.37	2.96	1	2	0.03	0.03	Outside one's house	Outside one's house
skeleton	bone	2.42	3.12	3	1	0.63	0.05	Either	Either
fire	match	4.04	3.40	1	1	0.02	0.25	Either	Either
onion	garlic	2.34	2.49	2	2	0.08	0.08	Inside one's house	Inside one's house

devil	angel	3.32	3.60	2	2	0.10	0.47	Either	Either
jar	can	2.63	5.43	1	1	0.08	0.05	Inside one's house	Inside one's house
tire	wheel	2.80	3.14	1	1	0.13	0.01	Either	Either
ladder	stairs	2.67	3.08	2	1	0.04	0.03	Either	Either
moon	star	3.41	3.62	1	1	0.12	0.12	Outside one's house	Outside one's house
elephant(s)	giraffe	2.76	1.89	3	2	0.02	0.04	Outside one's house	Outside one's house
tomato	lettuce	2.48	2.24	3	2	0.06	0.29	Inside one's house	Inside one's house
envelope	stamp(s)	2.71	2.48	3	1	0.07	0.09	Inside one's house	Either
eraser	pencil	1.72	2.70	3	2	0.51	0.05	Inside one's house	Inside one's house

CHAPTER 4: FUNCTION WORD AUTOCORRECTIONS ARE UNAFFECTED BY DISTRACTION: SEPARATION OF AUTOCORRECTIONS AND OTHER SPEECH ERRORS UNDER NOISY CONDITIONS

Introduction

Everyday speech production requires speakers to rapidly integrate both linguistic content and linguistic structure to form grammatically correct and meaningful sentences. Linguistic content refers to the words that we say, including what they mean and how they sound. Conversely, linguistic structure refers to the organization of individual sounds, phrases, or sentences that enable us to produce grammatically correct speech. While both are essential to producing and understanding language, the possible separation between content and structure in linguistic processing has led to significant debate in the literature (e.g., Chen et al., 2020; Cleland & Pickering, 2003; Ferreira F. & Clifton, 1986; Peterson et al., 2001). Further, the extent to which specific representations of linguistic structure operate during language processing is not well understood.

The goal of the present research is to further examine how grammatical units such as function words – a closed class of words that primarily operate to provide structure and represent relationships between words in a sentence – are processed during language production, independent of the semantic content. Function words contribute to the linguistic structure of sentences and are used to link together words in a sentence rather than contribute to the semantic content. This difference in the role of content and function words has led researchers to consider whether their different grammatical roles result in differences in lexical access. Early theories of lexical access proposed that the mode of access for function and content words may be distinct (Garrett, 1975) based on evidence from speech errors which demonstrated that both phonological

errors and word exchanges were less likely to occur on function than content words. Although some have argued that these differences can be explained by the higher frequency and predictability of function words (e.g., Dell, 1990; Segalowitz & Lane, 2000), other evidence suggests that function words are produced more quickly, even after accounting for the effects of words frequency (Bell et al., 2009).

A related debate to potential differences in lexical access is the question of whether access to function words is more cognitively automatic. Evidence from eye-movements during silent reading has indicated that individuals fixate less on function than content words (Roussel et al., 2016) and are less likely to notice repeated function than content words while reading (Staub et al., 2019), suggesting that function words are processed relatively quickly and automatically during reading. However, eye-movements do not tell us how function words are retrieved from the mental lexicon. To determine if access to function words is more automatic, we can instead look to evidence from speech errors in language production. Naturally occurring speech errors have received decades of attention in language production research, largely because of what they tell us about the process of lexical access. In particular, speech errors such as malapropisms can also reveal insights into the organization of language. Malapropisms are a specific type of error in which a speaker produces a phonologically related word in place of an intended word (Fay & Cutler, 1977). They are typically unrelated in meaning to the intended word and are thought to arise when the language system erroneously selects a phonologically related word from the mental lexicon in place of the intended word (Cutler & Fay, 1982; Dell & Kim, 2005; Fay & Cutler, 1977; Vitevitch, 1997). Although a formal comparison of the frequency of naturally occurring malapropisms on content and function words is not currently

available, malapropisms present an interesting opportunity to compare lexical access in function and content words.

To address differences in the production of malapropisms in content and function words, Gollan and colleagues (2020) employed a read aloud task that asked participants to read short passages in which a small number of words in each passage were substituted with malapropisms on either content or function words (e.g., 'Every shelf was lined *which* books, and I knew I could never finish the *haul* collection.'). Participants were told to read the paragraphs exactly as they were written, which required them to produce malapropisms that did not fit the context of the sentence. Notably, these malapropisms reflected substitutions of well-known words that are similar to malapropisms that occur during natural speech. The results showed that speakers spontaneously corrected the malapropisms and instead produced a word that fits the context, but these "autocorrections" were more likely to occur on function than on content words. Similar results have been obtained in mixed-language passages, showing that bilingual speakers will translate function words to match the dominant language of the paragraph more frequently than they will spontaneously translate content words (Gollan et al., 2014). These results suggest that individuals have more difficulty producing malapropisms on function words, and instead will correct them during the course of reading in order to produce grammatically correct sentences. Conversely, malapropisms are less likely to be corrected on content words, indicating that it is easier to produce semantically than grammatically incorrect features of sentences.

Why are function words particularly susceptible to autocorrections? One possible explanation is that differences in planning linguistic structure versus content lead to differences in planning and monitoring of function and content words. Specifically, speakers seem to be less efficient at monitoring structural elements of language during processing, perhaps due to greater

automaticity of syntactic planning. In a study by Schotter and colleagues (2019), Chinese-English bilinguals read aloud mixed language paragraphs in which a small subset of the words in each paragraph were written in the other language. Speakers produced more language intrusion errors – in which they automatically translated written switch words to avoid switching languages during speech production – on function than content words. This difference could not be explained by differences in gaze duration as gaze durations on words that elicited an intrusion error were similar across function and content words. Therefore, it is more likely that autocorrections on function words reflect increased failure in monitoring relative to content words rather than a lack of attention to function words. The monitoring system may be less likely to detect potential errors on function words while reading aloud because grammatical planning is more cognitively automatic and less error prone (Garrett, 1975) and so may not be monitored as closely. Thus, the increase in autocorrections for function words may reflect their status as closed-class words that support the grammatical structure of language, unlike content words that are open-class and contribute to the intended message.

The difference in susceptibility to autocorrections between content and function words presents a unique opportunity to investigate how linguistic structure is represented during speech production. Specifically, we can use the frequency of autocorrections on function words to investigate whether the production of function words is a cognitively automatic process. In the present experiment, participants were asked to read aloud short passages which contained 10 malapropisms on function words while listening to speech that was either intelligible (i.e., English speech) or unintelligible (i.e., Hebrew speech) to our participants, or without background noise. The three reading conditions allow us to differentiate between potential linguistic disruptions due to any background noise, or background noise that can be easily processed by the

participants. Based on Marsh et al.'s (2008) interference-by-process hypothesis, intelligible speech should be more disruptive than unintelligible speech because only intelligible speech can be processed semantically by participants (see Martin et al., 1988 for a related account of linguistic distraction during reading). Therefore, if the production of function words does in fact rely on effortful processing during language production, then we would expect a decrease in autocorrections while reading aloud in noise, compared to reading without distraction, and this effect should be even larger in the English reading condition since intelligible speech is more likely to disrupt linguistic processing. However, if function words are produced relatively automatically, autocorrections should occur equally as frequently across all three reading conditions.

For comparison, we also considered whether speech errors on non-target words would vary as a function of reading condition. Speech errors are a normal part of everyday language production, and similar errors can arise during naturalistic reading. Comparing autocorrections to errors on non-target words allows us to examine whether autocorrections share a common mechanism with spontaneous language errors. Specifically, we can examine whether error rates on non-target content and function words differ from patterns of autocorrect errors to determine whether grammatical components of sentences are differently affected by noise relative to semantic content.

Method

Participants

Sixty participants (age range = 17-30, M = 18.30, SD = 1.67, 50 female, 8 male, and 2 who did not specify gender) from the McMaster University undergraduate community participated for course credit. All had normal or corrected-to-normal vision and hearing, learned

English before the age of 12, and did not speak or understand Hebrew. Two additional participants were removed from the experiment because they either didn't complete the task properly or did not meet the language requirements.

Materials and Procedure

The read aloud materials consisted of 12 paragraphs that were either adapted from previous research (Gollan et al., 2020) or were developed for this study. Paragraphs were an average of 172 words long (*SD* = 15.05, *range* = 153 – 198). In each paragraph, 10 function words were replaced with an autocorrect target. An example paragraph can be found below. The first version shows the paragraph as the participants saw them, and the second version shows the paragraphs with the autocorrect targets emphasized.

> Like a lot of people, my family used to have grass in our front yard. It made our house look like a picturesque home it the suburbs. On the weekends my dad would mow the lawn any pull up weeds. Unfortunately we have been in a drought for most than a decade now, and it has become unrealistic to maintain a lawn in this conditions. So a few years ago, we dug up what little grass had managed to survive and started over for scratch. We decided to plant cacti and succulents that are native to they area to cut down on water use. At first it was strange to no longer have lush, green grass it front of the house, but over time the new landscaping has grown in us. The cacti and rocks turned our home onto a little desert oasis, and they are much more tolerant of drought that the grass was.

> Like a lot of people, my family used to have grass in our front yard. It made our house look like a picturesque home ***it*** the suburbs. On the weekends my dad would mow the lawn ***any*** pull up weeds. Unfortunately we have been in a drought for ***most*** than a decade now, and it has become unrealistic to maintain a lawn in ***this*** conditions. So a few years ago, we dug up what little grass had managed to survive and started over ***for*** scratch. We decided to plant cacti and succulents that are native to ***they*** area to cut down on water use. At first it was strange to no longer have lush, green grass ***it*** front of the house, but over time the new landscaping has grown ***in*** us. The cacti and rocks turned our home ***onto*** a little desert oasis, and they are much more tolerant of drought ***that*** the grass was.

The auditory stimuli consisted of short audio clips of an audiobook of Alice in Wonderland (Malik-Moraleda et al., 2022). The Hebrew audio clips were a translation of the English version of Alice and Wonderland, and were selected from the same chapters of the book. Each participant read the 12 paragraphs aloud in a fixed order. The reading conditions were

presented in three separate blocks, with four paragraphs in each reading condition. The order of the reading conditions was counterbalanced across participants.

During the experiment, participants were seated in front of a computer in a sound-dampened room. The audio was played out of two speakers, one in front and a second speaker behind the participant. Participants were instructed to read each paragraph out loud at a comfortable pace. They were further told that some of the words in the paragraphs may appear to be incorrect, but regardless they should read the words exactly as they are written. Paragraphs were displayed on the screen in 18 pt Times New Roman font, and participants wore a head-worn microphone to capture their voice as they read the paragraphs. The experiment began with two practice paragraphs each with six target words, followed by the 12 experimental paragraphs. The experiment was programmed in PsychoPy (Peirce et al., 2019), and took approximately 20 minutes to complete.

Results

Reading times were calculated for each paragraph to determine whether speech rate affected the production of errors during reading. The total reading for each paragraph in seconds was divided by the number of syllables in each paragraph to control for the different lengths of the paragraphs. Average speech rate was quite similar across the three conditions: Silent condition: 4.16 syllables/s; English condition: 4.13 syllables/s; and Hebrew condition: 4.18 syllables/s.

Errors were identified based on whether any error was produced on a target or non-target word, and further coded based on specific error types. Errors on target words were coded as autocorrections in cases where participants produced an alternate word in place of the written target. Partial errors, in which a participant started to produce an error but stopped before they

fully produced the word, were not included in main analyses. All other errors were coded based on whether participants produced an alternate word in place of what was written (substitutions), failed to produce a written word (omissions), produced an additional word that was not written in the paragraph (additions), repeated a word or series of words (repetitions), produced a phonological or morphological error, initiated a word and then immediately restarted (false starts), exchanged the order of words written in the paragraph (word exchanges), produced a filled pause, or mispronounced a written word. Finally, errors were coded based on whether participants corrected themselves after initiating an error (repairs). Note that we did not code repairs on repetitions, false starts, or filled pauses because these error types do not represent a misreading of the text.

Participants produced errors on 3.36% (SD = 2.04%) of words per paragraph. Paragraphs in which the average number of total errors produced was greater than three standard deviations from the mean were removed from the analysis. This resulted in the removal of 10 paragraphs across four participants. A summary of the number of errors of each type in each reading condition can be found in Table 1. See Appendix A for a full breakdown of the numbers of error subtypes.

Table 1.

Number and Percentage of each Error Type

	Reading Condition		
	English	Hebrew	Silent
Error Type	Total Errors (%Error)	Total Errors (%Error)	Total Errors (%Error)
Autocorrections	483 (20.09%)	475 (19.80%)	475 (19.63%)
Non-Autocorrection Target Errors	104 (4.34%)	102 (4.25%)	94 (4.00%)
Non-Target Function Words	343 (1.71%)	323 (1.61%)	306 (1.51%)
Non-Target Content Words	379 (2.10%)	391 (2.17%)	359 (1.98%)

We next analyzed the effect of reading condition on the production of errors. All statistical analyses were carried out using R, Version 4.2.0. Errors were analysed using logistic mixed effects analysis with the lme4 (Bates et al, 2015) and lmerTest (Kuznetsova et al, 2015) packages, with errors as the binary dependent variable. Speech rate was included as a covariate in all analyses to control for the possible influence of speech rate on error production. A summary of the results can be found in Table 2.

Autocorrections

The first analysis examined the likelihood of producing an autocorrection in each reading condition. Autocorrections were included as the binary dependent variable and reading condition and speech rate, and their interaction, as the fixed effects. Reading condition was contrast coded with the silent reading condition as the reference group. The first contrast compared the Silent reading condition with the English reading condition (contrast codes: Silent: -.5, English: .5, Hebrew: 0), and the second contrast compared the Silent reading condition with the Hebrew reading condition (contrast codes: Silent: -.5, Hebrew: .5, English: 0). Speech rate was centered and scaled. Participant and paragraph were included as random effects. Speech rate and reading condition were included as by-participant random slopes, and speech rate was included as a by-participant random slope. All other random slopes were determined to account for less than 1% of the variance in the model and were thus removed. As can be seen in Figure 1, there was no effect of reading condition on autocorrections (Silent vs. English Condition: B = 0.01, *SE* = 0.09, *Z* = 0.16, *p* = .870; Silent vs. Hebrew Condition: B = 0.05, *SE* = 0.09, *Z* = 0.59, *p* = .558).

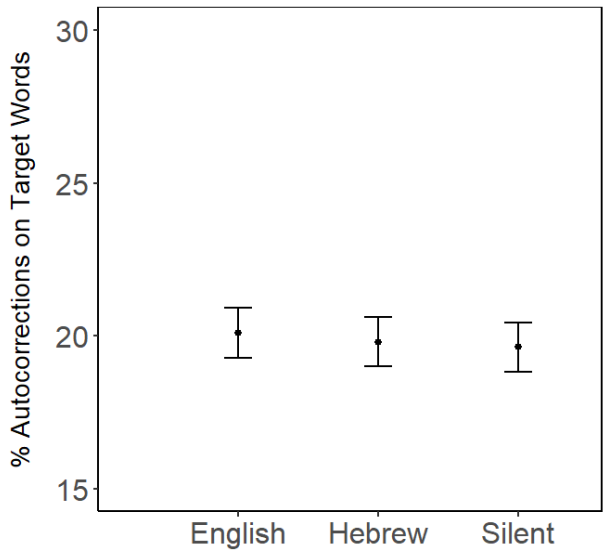

Figure 1. Percentage of Autocorrection in Each Reading Condition.

Non-target Errors

Although our first analysis examined the effect of reading condition and speech rate on autocorrections, we additionally considered whether the reading conditions had an effect on substitution errors on function words, which are the most conceptually similar to autocorrections in this experiment. Substitutions were also the most common type of error on non-target words, and for function words there were 126, 98, and 96 substitution errors in the English, Hebrew, and Silent conditions respectively. Reading condition, word position, speech rate, and their interactions were included as fixed effects in the model. Participant and paragraph were included as random effects. Reading condition and the interaction between word position and speech rate were included as by-participant random slopes, and word position and the interaction between word position and reading condition were included as by-paragraph random slopes. The results for substitution errors are shown in Figure 2. Function word substitutions were more likely to occur in the English than Silent reading conditions, B = 0.39, *SE* = 0.16, *Z* = 2.41, *p* = .016, and

marginally more likely to occur on words within five positions of a target, B = -0.38, *SE* = 0.21, *Z* = -1.78, *p* = .075.

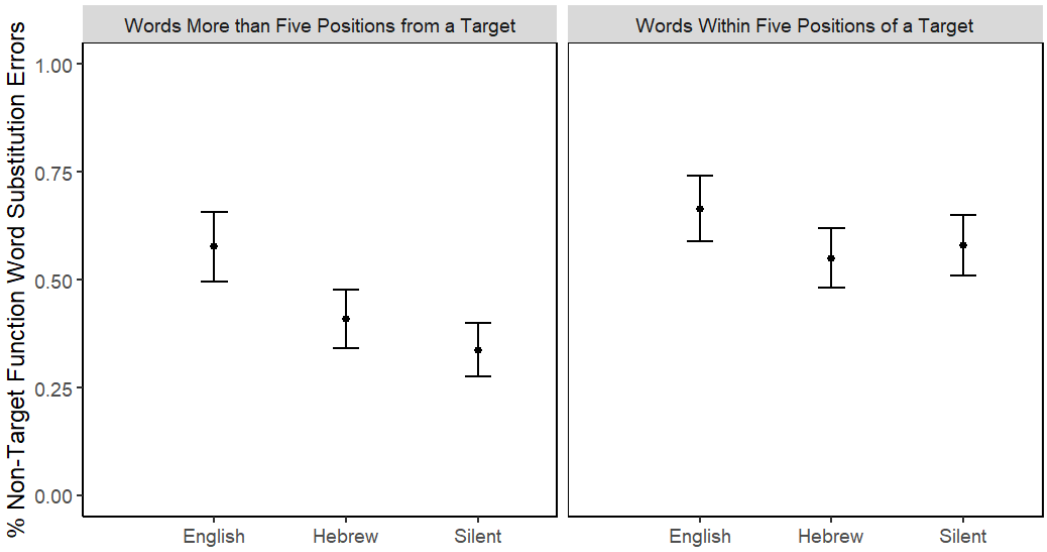

Figure 2. Percentage of Substitution Errors on Non-Target Words based on Reading Condition, Part of Speech, and Word Position.

Given the significant effect of reading condition on non-target function word substitutions, we next considered whether reading condition affected errors on all non-target errors. For this analysis, errors were collapsed across all error subtypes, and analyzed using a logistic mixed effects regression with errors as the binary dependent variable. The first analysis included reading condition, part of speech (content or function word), and speech rate as the fixed effects, and participant and paragraph as random effects. However, since errors were more common on target words, we also considered whether reading disruptions from target words would affect error production on non-target words that were in closer proximity to the target. We therefore constructed a second logistic mixed effects model that included word position as an additional fixed effect. Word position was included as a two-level factor comparing words that were within five positions of a target versus words that were greater than five positions from a target. Model comparison revealed that the analysis including word position provided a better fit

to the data (χ^2 = 86.20, p < .001), and thus we report the results based on the model that includes word position, reading condition, part of speech, and speech rate as fixed effects. Participant and paragraph were included as random effects, along with by-participant slopes for reading condition, word position, and part of speech, and by-paragraph slopes for word position and part of speech. All other random slopes were determined to account for less than 1% of the variance and were removed from the model.

The analysis revealed that non-target errors were more likely to occur on words that were within five positions of a target (B = -0.33, *SE* = 0.10, *Z* = -3.36, *p* = .001), and were more frequent on content than function words (B = 0.31, *SE* = 0.08, *Z* = 3.74, *p* = <.001). None of the interactions reached significance. See Figure 3 for the percentage of non-target errors in each condition.

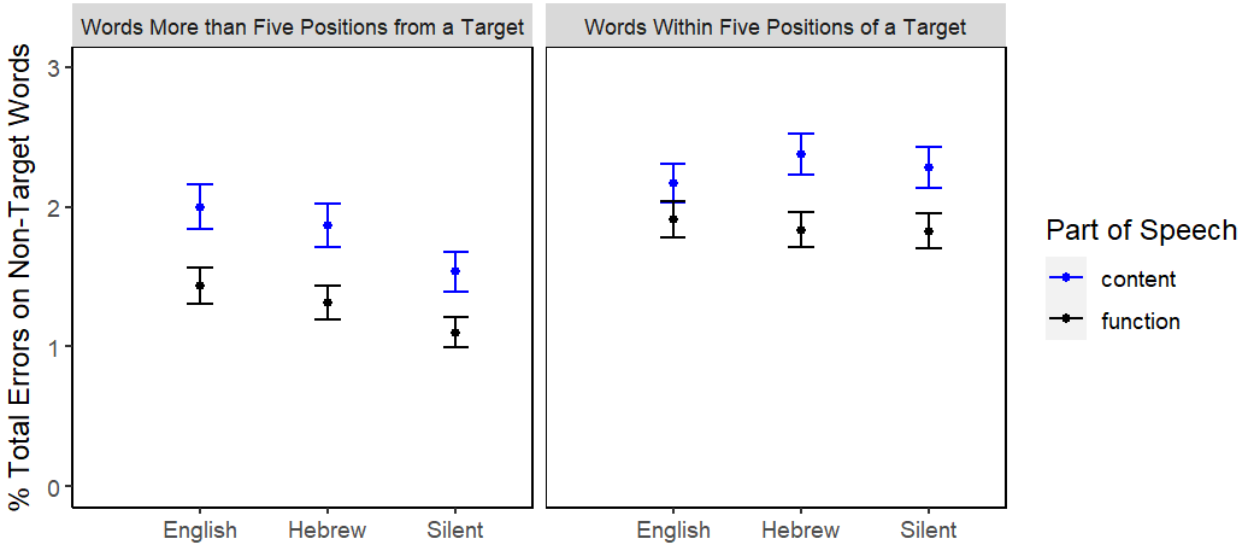

Figure 3. Percentage of Errors on Non-Target Words based on Reading Condition, Part of Speech, and Word Position.

Repairs

Finally, we examined whether reading condition affected the likelihood of repairing one's error as well as whether repairs were more likely on target than non-target errors. A logistic

mixed effects regression model was constructed with repairs as the binary dependent variable, and reading condition, word status (target vs. non-target), speech rate, and their interactions, as fixed effects. Participant and paragraph were further included as random effects. However, none of these effects reached significance (*p*s > .05), suggesting that the likelihood of repairing an error is not affected by distracting speech. Further, autocorrections were not more likely to be repaired than non-target errors which indicates that conscious monitoring of autocorrections does not significantly differ than non-target errors. See Table 2 for a full summary of the results.

Table 2.

Summary of Logistic Mixed Effects Analyses.

Effect	*B*	*SE*	*Z*	*p*
Autocorrections				
(Intercept)	-1.53	0.15	-9.97	<.001
Reading Condition (Silent vs. Hebrew)	0.05	0.09	0.59	.558
Reading Condition (Silent vs. English)	0.01	0.09	0.16	.870
Non-Target Function Word Substitutions				
(Intercept)	-5.54	0.13	-41.40	<.001
Reading Condition (Silent vs. Hebrew)	-0.08	0.16	-0.47	.637
Reading Condition (Silent vs. English)	0.39	0.16	2.41	**.016**
Word Position	-0.39	0.21	-1.78	.075
Reading Condition (Silent vs. Hebrew) x Word Position	0.00	0.33	0.01	.991
Reading Condition (Silent vs. English) x Word Position	0.33	0.36	0.94	.348
Non-Target Errors				
(Intercept)	-4.15	0.08	-52.37	<.001
Reading Condition (Silent vs. Hebrew)	0.10	0.06	1.48	.139
Reading Condition (Silent vs. English)	0.11	0.07	1.71	.088
Word Position	-0.33	0.10	-3.36	**.001**
Part of Speech	0.31	0.08	3.74	**<.001**
Reading Condition (Silent vs. Hebrew) x Word Position	0.11	0.13	0.83	.407

Reading Condition (Silent vs. English) x Word Position	0.21	0.13	1.64	.102
Reading Condition (Silent vs. Hebrew) x Part of Speech	0.05	0.13	0.36	.722
Reading Condition (Silent vs. English) x Part of Speech	-0.09	0.13	-0.68	.499
Word Position x Part of Speech	0.16	0.09	1.69	.090
Repairs				
(Intercept)	-0.68	0.09	-8.00	<.001
Reading Condition (Silent vs. Hebrew)	-0.01	0.11	-0.11	.911
Reading Condition (Silent vs. English)	0.15	0.10	1.50	.134
Target Status (Target vs. Non-Target)	-0.09	0.14	-0.65	.514
Reading Condition (Silent vs. Hebrew) x Target Status (Target vs. Non-Target)	-0.05	0.20	-0.26	.797
Reading Condition (Silent vs. English) x Target Status (Target vs. Non-Target)	-0.07	0.21	-0.33	.745

Discussion

The present study investigated the effects of noise on autocorrections and other types of speech errors produced during reading aloud. The analyses revealed distinct patterns of results for errors on targets compared with non-target errors. Whereas there was no effect of distracting speech on the production of autocorrections, analysis of function word substitutions on non-target words revealed that substitutions were more likely in the English than Silent reading condition. When the analysis of non-targets was expanded to include all possible errors on non-target words, results revealed errors occurred more often on words within five positions of a target word and were more likely to occur on content than function words. Finally, there was no effect of reading condition on repairs for any type of error, nor was there a difference in the pattern of repairs for target and non-target words.

The results of this experiment suggest that the production of autocorrections is unaffected by distracting speech, but what makes autocorrections on function words resistant to distraction?

Our original hypothesis was that if the production of function words relies on effortful processing, then we should see a reduction in autocorrections in the distracted reading conditions where speech planning is disrupted. Instead, the similar rates of autocorrections across the three reading conditions indicate that the production of function words is cognitively automatic rather than effortful. One possible explanation for this finding is that because function words are a closed class of words that are higher in frequency, they are more easily selected from the mental lexicon. The ease of processing function words relative to the linguistic content may make it more difficult to produce function words that do not fit the syntactic structure of the sentence. Alternatively, it is possible that autocorrection rates were not affected by reading condition simply because participants failed to notice the autocorrect targets. However, most of the time participants followed instructions and did not correct the autocorrect targets, suggesting that autocorrections are a result of automatic processing instead of a failure of attention.

The finding that autocorrections on function words are unaffected by distracting speech and speech rate contrasts with the pattern of substitution errors on non-target function words. Substitution errors are conceptually similar to autocorrections in that both types of errors involve the production of an entire alternate word in place of what was written. While autocorrections on function words were unaffected by noise, people were more likely to produce substitution errors on non-target function words in the English than Silent reading conditions. The contrast in findings for autocorrections and function word substitutions indicates that autocorrections are invariant to processes that affect spontaneous speech errors. This may be explained by important differences between substitutions and autocorrections. Autocorrections reflect a speaker's tendency to correct written malapropisms to avoid producing syntactically and/or semantically incorrect utterances. Substitutions, however, reflect the production of an alternate word in case

where the written word was appropriate for the context. In this sense, autocorrections and substitutions represent competing forces in language production: autocorrections reflect a tendency to produce syntactically and semantically correct speech, whereas substitutions reflect a failure to produce an intended word. This difference suggests that autocorrections are not influenced by the same linguistic factors such as distracting speech that disrupt typical language production.

Interestingly, as can be seen in Figure 4, the inclusion of speech rate in the analyses revealed that rates of autocorrections remained consistent across varied speech rates (B = -0.11, *SE* = 0.07, *Z* = -1.50, *p* = .133), whereas people who talked faster also produced fewer speech errors on nontarget words (B = -0.21, *SE* = 0.05, *Z* = -4.41, *p* = <.001). The decrease in non-target errors at higher speech rates suggests that speakers who are more efficient are also less likely to make errors. However, the null effect of speech rate on autocorrections indicates that autocorrections are unaffected by individual differences in speech rate. Further, the dissociation between the effect of speech rate on autocorrections and non-target errors strengthens the conclusion that autocorrections are a cognitively automatic process that is invariant to linguistic forces that increase the likelihood of making everyday speech errors.

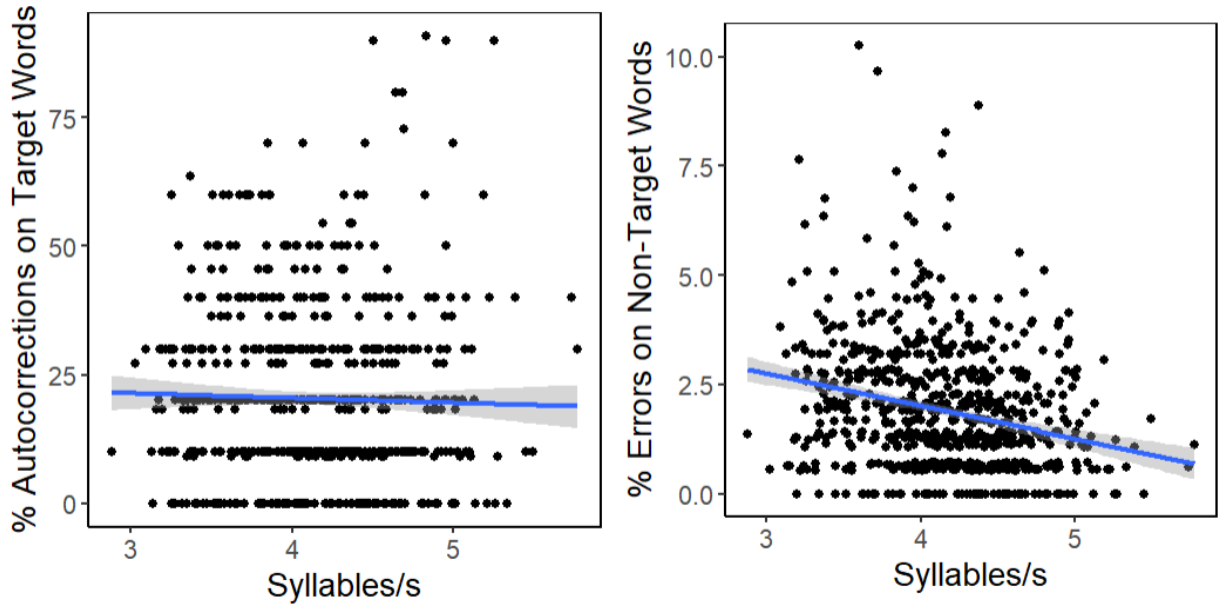

Figure 4. Percentage of Autocorrections (left) and non-Target Errors (right) by Speech Rate.

Although autocorrections were not affected by the reading conditions, autocorrect targets did affect processing of nearby words. Specifically, participants produced more errors on non-target words that were within five positions of a target. This finding indicates that, although the production of autocorrections appears to be cognitively automatic, the presence of autocorrect targets in the text does disrupt linguistic processing. It is possible that producing syntactically incorrect speech taxes the production system, resulting in difficulties in speech planning and monitoring. For example, the autocorrect targets may draw attention away from nearby words, making it more difficult to appropriately monitor for errors on words in close proximity to the targets. Interestingly, the increased number of errors on words within five positions of a target was not restricted to a particular type of error or word class, suggesting that autocorrect targets affect processing of surrounding words more generally, rather than influencing a specific type of error.

A final consideration is the result that non-target errors were more frequent on content than function words. While this is consistent with previous research on spontaneous speech

errors, investigations of autocorrections demonstrate that they occur more frequently on function than content words (Gollan et al., 2020, 2022). We did not include content word autocorrect targets in the present study, but it is possible that distracting speech would affect them differently. Since production of content words may be less automatic, and monitored more extensively than function words, we might expect to see a reduction in autocorrections on content but not function words under distracted conditions. Autocorrect targets represent a syntactic or semantic error within a given sentence, and so disruptions to speech monitoring should make it easier to produce semantically incorrect content words than syntactically incorrect function words. Alternatively, if distracting speech disrupts monitoring processes, individuals may be more likely to produce autocorrections on words that are typically monitored more closely, such as content words. Future research comparing the effects of distracting speech on both content and function word autocorrections is needed to further determine whether the production of function words in particular is cognitively automatic, or if autocorrections more generally are not affected by distracting speech. This would help determine the potential separation of linguistic structure and linguistic content during language processing, particularly by showing how structure and content are differently affected by disruptions to language processing.

Overall, the distinct pattern of errors on autocorrect targets and non-target words suggests that the production of function words is a cognitively automatic rather than effortful process. Further, these results indicate that autocorrection is a unique behavior that produces a different pattern of errors compared to similar substitution errors on contextually appropriate words. This latter point is particularly important given recent findings that function word autocorrections are associated with risk of developing Alzheimer's Disease (Gollan, 2020, 2022). If function word

autocorrections reflect an automatic process, it is possible that reliance on automatic processing increases with aging or disease-related changes in cognition. Thus, further examination of autocorrections can reveal how differences in processing linguistic content and structure result in unique patterns of errors that are unaffected by noisy conditions.

References

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92-111. https://doi.org/10.1016/j.jml.2008.06.003

Bock, K., Levelt, W., & Gernsbacher, M. A. (2002). Language production: Grammatical encoding. *Psycholinguistics: Critical concepts in psychology*, *5*, 405-452.

Chen, X., Branigan, H. P., Wang, S., Huang, J., & Pickering, M. J. (2020). Syntactic representation is independent of semantics in Mandarin: Evidence from syntactic priming. *Language, cognition and Neuroscience*, *35*(2), 211-220. https://doi.org/10.1016/j.jml.2008.06.003

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214-230. https://doi.org/10.1016/S0749-596X(03)00060-3

Cutler, A., & Fay, D. A. (1982). One mental lexicon, phonologically arranged: comments on Hurford's comments. *Linguistic Inquiry*, 107-113. http://www.jstor.org/stable/4178262

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes*, *5*(4), 313-349. https://doi.org/10.1080/01690969008407066

Dell, G. S., & Kim, A. E. (2005). Speech errors and word form encoding. In *Phonological encoding and monitoring in normal and pathological speech* (pp. 29-53). Psychology Press.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic inquiry*, *8*(3), 505-520. http://www.jstor.org/stable/4177997

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of memory and language*, *25*(3), 348-368. http://www.jstor.org/stable/4177997

Garrett, M. F. (1975). The analysis of sentence production. In *Psychology of learning and motivation* (Vol. 9, pp. 133-177). Academic Press.

Gollan, T. H., Li, C., Stasenko, A., & Salmon, D. P. (2020). Intact reversed language-dominance but exaggerated cognate effects in reading aloud of language switches in bilingual Alzheimer's disease. *Neuropsychology*, *34*(1), 88. https://doi.org/10.1037/neu0000592

Gollan, T. H., Schotter, E. R., Gomez, J., Murillo, M., & Rayner, K. (2014). Multiple levels of bilingual language control: Evidence from language intrusions in reading aloud. *Psychological science*, *25*(2), 585-595. https://doi.org/10.1037/neu0000592

Gollan, T. H., Smirnov, D. S., Salmon, D. P., & Galasko, D. (2020). Failure to stop autocorrect errors in reading aloud increases in aging especially with a positive biomarker for Alzheimer's disease. *Psychology and aging*, *35*(7), 1016. https://doi.org/10.1037/neu0000592

Gollan, T. H., Stasenko, A., Li, C., Smirnov, D. S., Galasko, D., & Salmon, D. P. (2022). Autocorrection if→ of function words in reading aloud: A novel marker of Alzheimer's risk. *Neuropsychology*. https://doi.org/10.1037/neu0000592

Haber, R. N., & Schindler, R. M. (1981). Error in proofreading: Evidence of syntactic control of letter processing?. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 573. https://doi.org/10.1037/neu0000592

Malik-Moraleda, S.*, Ayyash, D.*, Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12

language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014-1019. https://doi.org/10.1037/neu0000592

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of memory and language*, *58*(3), 682-700. https://doi.org/10.1016/j.jml.2007.05.002

Martin, R. C., Wogalter, M. S., & Forlano, J. G. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of memory and language*, *27*(4), 382-398. https://doi.org/10.1016/j.jml.2007.05.002

Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1223. https://doi.org/10.1016/j.jml.2007.05.002

Roussel, S., Rohr, A., Raufaste, E., & Nespoulous, J. L. (2016). Eye-movement analysis in reading content words and function words. *Cognition*.

Schotter, E. R., Li, C., & Gollan, T. H. (2019). What reading aloud reveals about speaking: Regressive saccades implicate a failure to monitor, not inattention, in the prevalence of intrusion errors on function words. *Quarterly Journal of Experimental Psychology*, *72*(8), 2032-2045. https://doi.org/10.1177/1747021818819480

Segalowitz, S. J., & Lane, K. C. (2000). Lexical access of function versus content words. *Brain and language*, *75*(3), 376-389. https://doi.org/10.1016/j.jml.2007.05.002

Staub, A., Dodge, S., & Cohen, A. L. (2019). Failure to detect function word repetitions and omissions in reading: Are eye movements to blame?. *Psychonomic bulletin & review*, *26*, 340-346. https://doi.org/10.3758/s13423-018-1492-z

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, *40*(3), 211-228. https://doi.org/10.3758/s13423-018-1492-z

APPENDIX A

	Reading Condition			
	English	Hebrew	Silent	Total
Error Type	Total Errors (%Error)	Total Errors (%Error)	Total Errors (%Error)	Total Errors (%Error)
Autocorrect Targets				
autocorrections	483 (82.28%)	475 (82.32%)	475 (83.48%)	1433 (82.69%)
partial autocorrections	29 (4.94%)	30 (5.20%)	31 (5.45%)	90 (5.19%)
additions	2 (0.34%)	1 (0.17%)	2 (0.35%)	5 (0.29%)
false starts	15 (2.56%)	7 (1.21%)	12 (2.11%)	34 (1.96%)
filled pauses	2 (0.34%)	3 (0.52%)	4 (0.70%)	9 (0.52%)
mispronunciations	6 (1.02%)	8 (1.39%)	3 (0.53%)	17 (0.98%)
morphological errors	3 (0.51%)	1 (0.17%)	1 (0.18%)	5 (0.29%)
omissions	22 (3.75%)	22 (3.81%)	16 (2.81%)	60 (3.46%)
phonological errors	0 (0.00%)	1 (0.17%)	1 (0.18%)	2 (0.12%)
repetitions	24 (4.09%)	23 (3.99%)	24 (4.22%)	71 (4.10%)
word exchanges	1 (0.17%)	6 (1.04%)	0 (0.00%)	7 (0.40%)
Non-Target Function Words				
additions	19 (5.54%)	22 (6.81%)	14 (4.58%)	55 (5.66%)
false starts	11 (3.21%)	16 (4.95%)	18 (5.88%)	45 (4.63%)
filled pauses	8 (2.33%)	5 (1.55%)	8 (2.61%)	21 (2.16%)
mispronunciations	2 (0.58%)	3 (0.93%)	4 (1.31%)	9 (0.93%)
morphological errors	6 (1.75%)	8 (2.48%)	9 (2.94%)	23 (2.37%)
omissions	111 (32.36%)	110 (34.06%)	90 (29.41%)	311 (32.00%)
phonological errors	2 (0.58%)	3 (0.93%)	1 (0.33%)	6 (0.62%)
repetitions	33 (9.62%)	40 (12.38%)	40 (13.07%)	113 (11.63%)
substitutions	126 (36.73%)	98 (30.34%)	96 (31.37%)	320 (32.92%)
partial substitutions	11 (3.21%)	9 (2.79%)	13 (4.25%)	33 (3.40%)
word exchanges	14 (4.08%)	9 (2.79%)	13 (4.25%)	36 (3.70%)
Non-Target Content Words				
additions	29 (7.65%)	26 (6.65%)	23 (6.41%)	78 (6.91%)
false starts	51 (13.46%)	56 (14.32%)	57 (15.88%)	164 (14.53%)
filled pauses	4 (1.06%)	3 (0.77%)	5 (1.39%)	12 (1.06%)
mispronunciations	25 (6.60%)	28 (7.16%)	18 (5.01%)	71 (6.29%)

morphological errors	67 (17.68%)	80 (20.46%)	68 (18.94%)	215 (19.04%)
omissions	39 (10.29%)	42 (10.74%)	30 (8.36%)	111 (9.83%)
phonological errors	32 (8.44%)	28 (7.16%)	35 (9.75%)	95 (8.41%)
repetitions	32 (8.44%)	22 (5.63%)	39 (10.86%)	93 (8.24%)
substitutions	69 (18.21%)	68 (17.39%)	66 (18.38%)	202 (17.89%)
partial substitutions	25 (6.60%)	31 (7.93%)	13 (3.62%)	69 (6.11%)
word exchanges	6 (1.58%)	7 (1.79%)	5 (1.39%)	18 (1.59%)
Total Errors				
Autocorrect Targets	587 (44.84%)	577 (44.69%)	569 (46.11%)	1733 (45.20%)
Non-Target Function Words	343 (26.20%)	323 (25.02%)	306 (24.80%)	972 (25.35%)
Non-Target Content Words	379 (28.95%)	391 (30.29%)	359 (29.09%)	1129 (29.45%)

CHAPTER 5: GENERAL DISCUSSION

The goal of the present thesis was to further our understanding of the process of lexical access by examining the linguistic and contextual factors that influence access to distinct categories of words. Across three chapters, we investigated the effects of semantic competition on proper names (Chapter 2), the effect of semantic constraint on object naming (Chapter 3), and the cognitive automaticity of function word production (Chapter 4). Together, these studies used behavioural methods that required participants to produce individual words across different linguistic contexts to highlight important differences in lexical retrieval for different types of words.

Beginning in Chapter 2, we investigated how proper names compete for retrieval during lexical selection. Proper names are especially susceptible to retrieval difficulties, and TOTs in particular, compared with object names. While this is often attributed to their typical combination of 2-3 words, low frequency, and lack of semantic information about their referent (Brédart, 1993; Cohen, 1990; Hanley & Chapman, 2008), it is less clear how semantic relationships between words contribute to difficulties in proper name retrieval. Conflicting evidence has found that related names can either facilitate retrieval (e.g., Oberle & James, 2013; White et al., 2013; Vitkovitch et al, 2006), or induce competition (e.g., Brédart & Dardenne, 2015; Davis & Abrams, 2016; Marful et al., 2014). To address this debate, we conducted two experiments that required participants to either read a famous or non-famous name aloud (Experiment 1), or to classify a name as famous or non-famous (Experiment 2), prior to naming a celebrity picture. Results revealed that participants experienced more TOTs after classifying a famous than non-famous name, but this difference was not present when participants simply read the prime name. However, across both studies, successful retrievals decreased with increasing

trial number. Follow-up analyses revealed that this may be due to proactive interference from previous attempts to retrieve related celebrity names.

The findings reported in Chapter 2 demonstrate that the effects of competition between related proper names are not uniform, but instead depend on both depth of processing in a single trial, and the build-up of proactive interference across trials. When participants were asked to name a celebrity picture directly after processing a prime name, TOTs increased only when the prime names were sufficiently processed. However, the robust effect of accumulating interference across successive attempts to retrieve proper names indicates that interference from earlier names results in fewer successful retrievals. While the effects of proactive interference in memory have been well established (e.g., Anderson & McNeely, 1996; Underwood, 1957; Watkins & Watkins, 1975), these findings are the first of their kind to report proactive interference in proper name retrieval without explicit instructions to remember previous items in the list. Further, the finding that successful retrievals decrease with increasing trial number has not been previously reported in studies investigating proper name or object name retrieval. The difference in results from previous studies investigating object name retrieval is particularly interesting because it seems to suggest that proactive interference may be stronger for proper names than common objects.

In Chapter 3, we expanded our investigation of the semantic influences on word retrieval to investigate how the broader semantic context of sentences facilitates or inhibits retrieval of object names. It is commonly accepted that individuals can use the semantic context of a sentence to make predictions about upcoming words, particularly when presented with a highly constraining context in which there are fewer reasonable alternatives that can fit the sentence (e.g., Altmann & Kamide, 1999; DeLong et al., 2005; Heilbron et al., 2022; Van Berkum et al.,

2005). Accurate predictions are often considered to benefit processing, but it is less clear if these benefits extend to situations where predictions are only almost correct, or completely incorrect. We investigated this question in Chapter 3 by asking participants to either produce an object name (Experiments 1 and 2) or perform a semantic classification task (Experiment 3) for an object presented at the end of a high or low constraint sentence. We were specifically interested in whether producing a word following a highly constraining sentence fragment would facilitate retrieval even in cases where the target word did not match the most expected word, relative to when predictions about upcoming words could not be generated. The results of the first two studies that required participants to produce the word showed that small prediction errors in highly constraining contexts resulted in a smaller processing cost relative to larger prediction errors. However, the same words were still processed fastest in a low constraint context where predictions could not be generated. These results were replicated in a third experiment that required participants to perform a semantic classification of the target word instead of producing the word. Together, these findings suggest that across both language production and comprehension, predicting almost correctly is better than predicting completely incorrectly, but not better than not predicting at all.

The results of Chapter 3 speak to an important debate within psycholinguistics: what is the utility of predicting during language production? Specifically, if predictive processing only benefits lexical access when predictions are exactly accurate – which, according to recent research is a relatively rare occurrence in natural language (Luke & Christianson, 2016) – why predict at all? One possible explanation is that prediction is used in language learning. Dell and Chang (2014) proposed the p-chain framework which argues that prediction is useful for language learning because generating predictions naturally leads to prediction failure, which

results in implicit learning. On this account, the processing cost associated with failed predictions provides an error signal that is used to tune the language system. The results presented in Chapter 3 demonstrate a clear processing cost to incorrect predictions, even in cases where the target was semantically similar to the most expected term. Although these studies weren't designed to assess the utility of prediction in learning, the findings support the idea that there is a cost to prediction failure, which may be used as an error signal leading to implicit learning.

Chapters 2 and 3 together highlight how semantic constraints affect the process of lexical access. However, it is important to note that these studies both investigate lexical access of content words. Content words, which contribute to the meaning of sentences, are distinct from function words, which support the linguistic structure of sentences. Theories of language processing often argue that linguistic content and structure are processed separately (e.g., Chen et al., 2020; Cleland & Pickering, 2003; Ferreira F. & Clifton, 1986; Peterson et al., 2001). This leads to questions regarding whether function words themselves are processed differently than content words. Specifically, is the process of lexical access for function words more cognitively automatic? This question was explored in Chapter 4. Participants were asked to read aloud paragraphs while listening to English speech, Hebrew speech, or without distraction. 10 words in each paragraph were replaced with a grammatically incorrect target, and participants were told to read the paragraphs exactly as they were written. We were particularly interested in whether individuals would produce autocorrections – a phenomenon in which individuals spontaneously correct the grammatically incorrect target words – more frequently in the distracting reading conditions, and whether the pattern of autocorrections would differ from non-target speech errors. The results showed that the distracting speech did not affect the likelihood of producing

an autocorrection on the target words. However, participants were more likely to produce substitutions on non-target function words in the English reading condition, and were more likely to make errors on non-target words that were within five positions of an autocorrect target. Additionally, more errors were produced on non-target content words than non-target function words, confirming previous research that content words are more susceptible to speech errors than function words.

An important question that was addressed in Chapter 4 concerned whether access to function words is cognitively automatic. The finding that autocorrections were unaffected by distraction seems to suggest that function words are indeed cognitively automatic rather than effortful. However, the results also showed that function word substitutions – a class of errors that are conceptually similar to autocorrections – were more frequent in the English than Silent reading conditions. The contrasting pattern of results for autocorrections and non-target function word substitutions seems to indicate that autocorrections are not affected by the same linguistic factors as everyday speech errors. Instead, autocorrections appear to be resistant to distractions that disrupt monitoring during language production. A remaining question concerns whether autocorrections themselves are cognitively automatic, or whether these results are specific to function word autocorrections. Content words may be monitored more extensively than function words, which may affect the extent to which content word autocorrections are affected by noise. Follow-up research will be needed to determine if function word autocorrections are uniquely automatic, or whether these findings can be generalized to content word autocorrections.

Overall, the studies reported in this thesis contribute to important theoretical debates in the psycholinguistic literature – namely, how do linguistic factors such as semantic similarity, semantic context, and distraction differently affect the process of lexical access? First, we

demonstrated that the level of competition between proper names during retrieval varies based on both the depth of processing and the build-up of proactive interference across successive attempts to retrieve multiple loosely related names. Next, we presented evidence of a clear processing cost to failed predictions, even in contexts where individuals make only small prediction errors. Finally, we showed that autocorrections on function words reflect automatic cognitive processing that is resistant to linguistic factors that disrupt everyday speech errors. Overall, these findings are broadly supportive of a framework of lexical access in which both semantic relationships between words and different linguistic representations across words plays an important role in successful lexical selection.

REFERENCES

Abrams, L., & Davis, D. K. (2017). Competitors or teammates: how proper names influence each other. *Current Directions in Psychological Science*, *26*(1), 87-93. http://dx.doi.org/10.1177/0963721416677804

Altmann, G. T., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247-264. https://doi.org/10.1016/S0010-0277(99)00059-1

Anderson, M. C., & Neely, J. H. (1996). Interference and inhibition in memory retrieval. In E. L. Bjork & R. A. Bjork (Eds.), Memory: Handbook of perception and cognition (2nd ed., pp. 237–313). San Diego, CA: Academic Press.

Aristei, S., & Rahman, R. A. (2013). Semantic interference in language production is due to graded similarity, not response relevance. *Acta Psychologica*, *144*(3), 571-582. http://dx.doi.org/10.1016/j.actpsy.2013.09.006

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*

Bates, D., Maechler, B., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1—48. http://dx.doi.org/10.18637/jss.v067.i01

Bell, A., Brenier, J. M., Gregory, M., Girand, C., & Jurafsky, D. (2009). Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language*, *60*(1), 92-111. https://doi.org/10.1016/j.jml.2008.06.003

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior research methods*, *42*(3), 665-670. https://doi.org/10.3758/brm.42.3.665

Bock, K., Levelt, W., & Gernsbacher, M. A. (2002). Language production: Grammatical encoding. *Psycholinguistics: Critical concepts in psychology*, *5*, 405-452.

Boudewyn, M. A., Long, D. L., & Swaab, T. Y. (2015). Graded expectations: Predictive processing and the adjustment of expectations during spoken language comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 607-624. https://doi.org/10.3758/s13415-015-0340-0

Brédart, S. (1993). Retrieval failures in face naming. *Memory*, *1*(4), 351-366. http://dx.doi.org/10.1080/09658219308258243

Brédart, S. (2017). The cognitive psychology and neuroscience of naming people. *Neuroscience & Biobehavioral Reviews*, *83*, 145-154. http://dx.doi.org/10.1016/j.neubiorev.2017.10.008

Brédart, S., & Dardenne, B. (2015). Similarities between the target and the intruder in naturally occurring repeated person naming errors. *Frontiers in psychology*, *6*, 1474. http://dx.doi.org/10.3389/fpsyg.2015.01474

Brédart, S., & Valentine, S. B. T. (1998). Descriptiveness and proper name retrieval. *Memory*, *6*(2), 199-206. https://doi.org/10.1080/741942072

Brothers, T., Dave, S., Hoversten, L. J., Traxler, M. J., & Swaab, T. Y. (2019). Flexible predictions during listening comprehension: Speaker reliability affects anticipatory processes. *Neuropsychologia*, *135*, 107225. https://doi.org/10.1016/j.neuropsychologia.2019.107225

Brothers, T., & Kuperberg, G. R. (2021). Word predictability effects are linear, not logarithmic: Implications for probabilistic models of sentence comprehension. *Journal of memory and language*, *116*, 104174. https://doi.org/10.1016/j.jml.2020.104174

Brothers, T., Swaab, T. Y., & Traxler, M. J. (2017). Goals and strategies influence lexical prediction during sentence comprehension. *Journal of memory and language*, *93*, 203-216. https://doi.org/10.1016/j.jml.2016.10.002

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological bulletin*, *109*(2), 204. http://dx.doi.org/10.1037/0033-2909.109.2.204

Brown, A. S., & Nix, L. A. (1996). Age-related changes in the tip-of-the-tongue experience. *The American journal of psychology*, 79-91. https://doi.org/10.2307/1422928

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior research methods*, *41*(4), 977-990. https://doi.org/10.3758/brm.41.4.977

Burke, D. M., Locantore, J. K., Austin, A. A., & Chae, B. (2004). Cherry pit primes Brad Pitt: Homophone priming effects on young and older adults' production of proper names. *Psychological Science*, *15*(3), 164-170. http://dx.doi.org/10.1111/j.0956-7976.2004.01503004.x

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults?. *Journal of memory and language*, *30*(5), 542-579. http://dx.doi.org/10.1016/0749-596X(91)90026-G

Chen, X., Branigan, H. P., Wang, S., Huang, J., & Pickering, M. J. (2020). Syntactic representation is independent of semantics in Mandarin: Evidence from syntactic

priming. *Language, cognition and Neuroscience*, *35*(2), 211-220. https://doi.org/10.1080/23273798.2019.1644355

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, *39*. https://doi.org/10.1017/s0140525x1500031x

Cleland, A. A., & Pickering, M. J. (2003). The use of lexical and syntactic information in language production: Evidence from the priming of noun-phrase structure. *Journal of Memory and Language*, *49*(2), 214-230. https://doi.org/10.1016/S0749-596X(03)00060-3

Cohen, G. (1990). Why is it difficult to put names to faces?. *British Journal of Psychology*, *81*(3), 287-297. http://dx.doi.org/10.1111/j.2044-8295.1990.tb02362.x

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407. https://doi.org/10.1037/0033-295X.82.6.407

Conley, P., Burgess, C., & Hage, D. (1999). Large-scale databases of proper names. *Behavior Research Methods, Instruments, & Computers*, *31*(2), 215-219. http://dx.doi.org/10.3758/BF03207713

Cross, E. S., & Burke, D. M. (2004). Do alternative names block young and older adults' retrieval of proper names?. *Brain and language*, *89*(1), 174-181. http://dx.doi.org/10.1016/S0093-934X(03)00363-8

Cutler, A., & Fay, D. A. (1982). One mental lexicon, phonologically arranged: comments on Hurford's comments. *Linguistic Inquiry*, 107-113. http://www.jstor.org/stable/4178262

Damian, M. F., & Als, L. C. (2005). Long-lasting semantic context effects in the spoken production of object names. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1372. https://doi.org/10.1037/0278-7393.31.6.1372

Dahlgren, D. J. (1998). Impact of knowledge and age on tip-of-the-tongue rates. *Experimental Aging Research*, *24*(2), 139-153. http://dx.doi.org/10.1080/036107398244283

Davis, D. K., & Abrams, L. (2016). Here's looking at you: Visual similarity exacerbates the Moses illusion for semantically similar celebrities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(1), 75. http://dx.doi.org/10.1037/xlm0000144

Deffler, S. A., Fox, C., Ogle, C. M., & Rubin, D. C. (2016). All my children: The roles of semantic category and phonetic similarity in the misnaming of familiar individuals. *Memory & cognition*, *44*(7), 989-999. http://dx.doi.org/10.3758/s13421-016-0613-z

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and cognitive processes*, *5*(4), 313-349. https://doi.org/10.1080/01690969008407066

Dell, G. S., & Chang, F. (2014). The P-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

Dell, G. S., & Kim, A. E. (2005). Speech errors and word form encoding. In *Phonological encoding and monitoring in normal and pathological speech* (pp. 29-53). Psychology Press.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117-1121. https://doi.org/10.1038/nn1504

Dupont, M. (2018). The similarities between the target and the intruder in naturally occurring person naming errors: A comparison between repeated and single naming confusions. *Journal of psycholinguistic research*, *48*(1), 33-42. http://dx.doi.org/10.1007/s10936-018-9586-3

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551. https://doi.org/10.1016/S0022-5371(81)90165-1

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic inquiry*, *8*(3), 505-520. http://www.jstor.org/stable/4177997

Fazekas, J., Jessop, A., Pine, J., & Rowland, C. (2020). Do children learn from their prediction mistakes? A registered report evaluating error-based theories of language acquisition. *Royal Society Open Science*, *7*(11), 180877. https://doi.org/10.31234/osf.io/3phxu

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of memory and Language*, *41*(4), 469-495. https://doi.org/10.1006/jmla.1999.2660

Federmeier, K. D., Wlotko, E. W., De Ochoa-Dewald, E., & Kutas, M. (2007). Multiple effects of sentential constraint on word processing. *Brain research*, *1146*, 75-84. https://doi.org/10.1016/j.brainres.2006.06.101

Ferreira, F., & Chantavarin, S. (2018). Integration and prediction in language processing: A synthesis of old and new. *Current directions in psychological science*, *27*(6), 443-448. https://doi.org/10.1177/0963721418794491

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of memory and language*, *25*(3), 348-368. https://doi.org/10.1016/0749-596X(86)90006-9

Fogler, K. A., & James, L. E. (2007). Charlie Brown versus Snow White: The effects of descriptiveness on young and older adults' retrieval of proper names. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *62*(4), 201-207. http://dx.doi.org/10.1093/geronb/62.4.P201

Frisson, S., Harvey, D. R., & Staub, A. (2017). No prediction error cost in reading: Evidence from eye movements. *Journal of Memory and Language*, *95*, 200-214. https://doi.org/10.1016/j.jml.2017.04.007

Garrett, M. F. (1975). The analysis of sentence production. In *Psychology of learning and motivation* (Vol. 9, pp. 133-177). Academic Press.

Gauvin, H. S., Jonen, M. K., Choi, J., McMahon, K., & de Zubicaray, G. I. (2018). No lexical competition without priming: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology*, *71*(12), 2562-2570. https://doi.org/10.1177/1747021817747266

Gollan, T. H., & Brown, A. S. (2006). From tip-of-the-tongue (TOT) data to theoretical implications in two steps: when more TOTs means better retrieval. *Journal of Experimental Psychology: General*, *135*(3), 462. http://dx.doi.org/10.1037/0096-3445.135.3.462

Gollan, T. H., Ferreira, V. S., Cera, C., & Flett, S. (2014). Translation-priming effects on tip-of-the-tongue states. *Language, Cognition and Neuroscience*, *29*(3), 274-288. http://doi.org/10.1080/01690965.2012.762457

Gollan, T. H., Li, C., Stasenko, A., & Salmon, D. P. (2020). Intact reversed language-dominance but exaggerated cognate effects in reading aloud of language switches in bilingual Alzheimer's disease. *Neuropsychology*, *34*(1), 88. https://doi.org/10.1037/neu0000592

Gollan, T. H., Schotter, E. R., Gomez, J., Murillo, M., & Rayner, K. (2014). Multiple levels of bilingual language control: Evidence from language intrusions in reading aloud. *Psychological science*, *25*(2), 585-595. https://doi.org/10.1177/0956797613512661

Gollan, T. H., Slattery, T. J., Goldenberg, D., Van Assche, E., Duyck, W., & Rayner, K. (2011). Frequency drives lexical access in reading but not in speaking: the frequency-lag hypothesis. *Journal of Experimental Psychology: General*, *140*(2), 186. https://doi.org/10.1037/e520602012-200

Gollan, T. H., Smirnov, D. S., Salmon, D. P., & Galasko, D. (2020). Failure to stop autocorrect errors in reading aloud increases in aging especially with a positive biomarker for Alzheimer's disease. *Psychology and aging*, *35*(7), 1016. https://doi.org/10.1037/pag0000550

Gollan, T. H., Stasenko, A., Li, C., Smirnov, D. S., Galasko, D., & Salmon, D. P. (2022). Autocorrection if→ of function words in reading aloud: A novel marker of Alzheimer's risk. *Neuropsychology*. https://doi.org/10.1037/neu0000829

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language*, *38*(3), 313-338. https://doi.org/10.1006/jmla.1997.2547

Griffin, Z. M., & Wangerman, T. (2013). Parents accidentally substitute similar sounding sibling names more often than dissimilar names. *PLoS One*, *8*(12), e84444. http://dx.doi.org/10.1371/journal.pone.0084444

Haber, R. N., & Schindler, R. M. (1981). Error in proofreading: Evidence of syntactic control of letter processing?. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(3), 573. https://doi.org/10.1037/0096-1523.7.3.573

Haeuser, K. I., & Kray, J. (2021). Effects of prediction error on episodic memory retrieval: evidence from sentence reading and word recognition. *Language, Cognition and Neuroscience*, 1-17. https://doi.org/10.1080/23273798.2021.1924387

Hanley, J. R., & Chapman, E. (2008). Partial knowledge in a tip-of-the-tongue state about two- and three-word proper names. *Psychonomic Bulletin & Review*, *15*(1), 156-160. http://dx.doi.org/10.3758/PBR.15.1.156

Hantsch, A., Jescheniak, J. D., & Schriefers, H. (2005). Semantic competition between hierarchically related words during speech planning. *Memory & Cognition*, *33*(6), 984-1000. https://doi.org/10.3758/BF03193207

Harley, T. A., & Bown, H. E. (1998). What causes a tip-of-the-tongue state? Evidence for lexical neighbourhood effects in speech production. *British Journal of Psychology*, *89*(1), 151-174. http://dx.doi.org/10.1111/j.2044-8295.1998.tb02677.x

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the*

National Academy of Sciences, *119*(32), e2201968119. https://doi.org/10.1073/pnas.2201968119

Hintz, F., Meyer, A. S., & Huettig, F. (2016). Encouraging prediction during production facilitates subsequent comprehension: Evidence from interleaved object naming in sentence context and sentence reading. *The Quaterly Journal of Experimental Psychology*, *69*(6), 1056-1063. http://dx.doi.org/10.1080/17470218.2015.1131309

Huebert, A. M., McNeely-White, K. L., & Cleary, A. M. (2023). On the relationship between tip-of-the-tongue states and partial recollective experience: Illusory partial recollective access during tip-of-the-tongue states. *Journal of Experimental Psychology: General*, *152*(2), 542. https://doi.org/10.1037/xge0001292

Huijbers, W., Papp, K. V., LaPoint, M., Wigman, S. E., Dagley, A., Hedden, T., ... & Sperling, R. A. (2017). Age-related increases in tip-of-the-tongue are distinct from decreases in remembering names: a functional MRI study. *Cerebral Cortex*, *27*(9), 4339-4349. https://doi.org/10.1093/cercor/bhw234

Kliegl, O., Pastötter, B., & Bäuml, K. H. T. (2015). The contribution of encoding and retrieval processes to proactive interference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *41*(6), 1778. http://dx.doi.org/10.1037/xlm0000096

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1—26. http://dx.doi.org/10.18637/jss.v082.i13

Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT press. http://dx.doi.org/10.7551/mitpress/6393.001.0001

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and brain sciences*, *22*(1), 1-38. https://doi.org/10.1017/S0140525X99001776

Luke, S. G., & Christianson, K. (2016). Limits on lexical prediction during reading. *Cognitive Psychology*, *88*, 22-60. https://doi.org/10.1016/j.cogpsych.2016.06.002

Mahon, B. Z., Costa, A., Peterson, R., Vargas, K. A., & Caramazza, A. (2007). Lexical selection is not by competition: a reinterpretation of semantic interference and facilitation effects in the picture-word interference paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 503. https://doi.org/10.1037/0278-7393.33.3.503

Malik-Moraleda, S.*, Ayyash, D.*, Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience.* https://doi.org/10.1038/s41593-022-01114-5

Marful, A., Paolieri, D., & Bajo, M. T. (2014). Is naming faces different from naming objects? Semantic interference in a face-and object-naming task. *Memory & cognition*, *42*(3), 525-537. http://dx.doi.org/10.3758/s13421-013-0376-8

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of memory and language*, *58*(3), 682-700. https://doi.org/10.1016/j.jml.2007.05.002

Martin, R. C., Wogalter, M. S., & Forlano, J. G. (1988). Reading comprehension in the presence of unattended speech and music. *Journal of memory and language*, *27*(4), 382-398. https://doi.org/10.1016/0749-596X(88)90063-0

Martin, C. D., Branzi, F. M., & Bar, M. (2018). Prediction is Production: The missing link between language production and comprehension. *Scientific reports*, *8*(1), 1-9. https://doi.org/10.1038/s41598-018-19499-4

Meyer, A. S., & Bock, K. (1992). The tip-of-the-tongue phenomenon: Blocking or partial activation?. *Memory & cognition*, *20*(6), 715-726. http://dx.doi.org/10.3758/BF03202721

Morgan, A. M., von der Malsburg, T., Ferreira, V. S., & Wittenberg, E. (2020). Shared syntax between comprehension and production: Multi-paradigm evidence that resumptive pronouns hinder comprehension. *Cognition*, *205*, 104417. https://doi.org/10.1016/j.cognition.2020.104417

Mortensen, L., Meyer, A. S., & Humphreys, G. W. (2006). Age-related effects on speech production: A review. *Language and Cognitive Processes*, *21*(1-3), 238-290. http://dx.doi.org/10.1080/01690960444000278

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402-407. https://doi.org/10.3758/bf03195588

Ness, T., & Meltzer-Asscher, A. (2020). Love thy neighbor: Facilitation and inhibition in the competition between parallel predictions. *Cognition*, 104509. https://doi.org/10.1016/j.cognition.2020.104509

Oberle, S., & James, L. E. (2013). Semantically-and phonologically-related primes improve name retrieval in young and older adults. *Language and cognitive processes*, *28*(9), 1378-1393. http://dx.doi.org/10.1080/01690965.2012.685481

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227-252. https://doi.org/10.1016/j.cognition.2009.09.007

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior research methods*, *51*(1), 195-203. http://dx.doi.org/10.3758/s13428-018-01193-y

Peterson, R. R., Burgess, C., Dell, G. S., & Eberhard, K. M. (2001). Dissociation between syntactic and semantic processing during idiom comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(5), 1223. https://doi.org/10.1037/0278-7393.27.5.1223

Pickering, M. J., & Garrod, S. (2013). An integrated theory of language production and comprehension. *Behavioral and brain sciences*, *36*(4), 329-347. https://doi.org/10.1017/s0140525x12001495

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension?. *Trends in cognitive sciences*, *11*(3), 105-110. https://doi.org/10.1016/j.tics.2006.12.002

Ramscar, M., Hendrix, P., Shaoul, C., Milin, P., & Baayen, H. (2014). The myth of cognitive decline: Non-linear dynamics of lifelong learning. *Topics in cognitive science*, *6*(1), 5-42. https://doi.org/10.1111/tops.12078

Resnik, K., Bradbury, D., Barnes, G. R., & Leff, A. P. (2014). Between thought and expression, a magnetoencephalography study of the "tip-of-the-tongue" phenomenon. *Journal of Cognitive Neuroscience*, *26*(10), 2210-2223. https://doi.org/10.1162/jocn_a_00611

Rommers, J., Dell, G. S., & Benjamin, A. S. (2020). Word predictability blurs the lines between production and comprehension: Evidence from the production effect in memory. *Cognition*, *198*, 104206. https://doi.org/10.1016/j.cognition.2020.104206

Rose, S. B., Aristei, S., Melinger, A., & Abdel Rahman, R. (2019). The closer they are, the more they interfere: Semantic similarity of word distractors increases competition in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(4), 753. https://doi.org/10.1037/xlm0000592

Roussel, S., Rohr, A., Raufaste, E., & Nespoulous, J. L. (2016). Eye-movement analysis in reading content words and function words. *Cognition*.

Salthouse, T. A., & Mandell, A. R. (2013). Do age-related increases in tip-of-the-tongue experiences signify episodic memory impairments?. *Psychological science*, *24*(12), 2489-2497. https://doi.org/10.1177/0956797613495881

Schnur, T. T. (2014). The persistence of cumulative semantic interference during naming. *Journal of Memory and Language*, *75*, 27-44. https://doi.org/10.1016/j.jml.2014.04.006

Schotter, E. R., Li, C., & Gollan, T. H. (2019). What reading aloud reveals about speaking: Regressive saccades implicate a failure to monitor, not inattention, in the prevalence of intrusion errors on function words. *Quarterly Journal of Experimental Psychology*, *72*(8), 2032-2045. https://doi.org/10.1177/1747021818819480

Segalowitz, S. J., & Lane, K. C. (2000). Lexical access of function versus content words. *Brain and language*, *75*(3), 376-389. https://doi.org/10.1006/brln.2000.2361

Shafto, M. A., Burke, D. M., Stamatakis, E. A., Tam, P. P., & Tyler, L. K. (2007). On the tip-of-the-tongue: neural correlates of increased word-finding failures in normal aging. *Journal*

of cognitive neuroscience, *19*(12), 2060-2070. https://doi.org/10.1162/jocn.2007.19.12.2060

Shao, Z., & Rommers, J. (2020). How a question context aids word production: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology*, *73*(2), 165-173. https://doi.org/10.1177/1747021819882911

Shipley, W. C. (1940). A self-administering scale for measuring intellectual impairment and deterioration. *The Journal of Psychology*, *9*(2), 371-377. http://dx.doi.org/10.1080/00223980.1940.9917704

Stasenko, A., & Gollan, T. H. (2019). Tip of the tongue after any language: Reintroducing the notion of blocked retrieval. *Cognition*, *193*, 104027. http://dx.doi.org/10.1016/j.cognition.2019.104027

Staub, A., Dodge, S., & Cohen, A. L. (2019). Failure to detect function word repetitions and omissions in reading: Are eye movements to blame?. *Psychonomic bulletin & review*, *26*, 340-346. https://doi.org/10.3758/s13423-018-1492-z

Staub, A., Grant, M., Astheimer, L., & Cohen, A. (2015). The influence of cloze probability and item constraint on cloze task response time. *Journal of Memory and Language*, *82*, 1-17. https://doi.org/10.1016/j.jml.2015.02.004

Szewczyk, J. M., & Federmeier, K. D. (2022). Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of memory and language*, *123*, 104311. https://doi.org/10.1016/j.jml.2021.104311

Underwood, B. J. (1957). Interference and forgetting. *Psychological review*, *64*(1), 49. https://doi.org/10.1037/h0044616

Vainio, S., Hyönä, J., & Pajunen, A. (2009). Lexical predictability exerts robust effects on fixation duration, but not on initial landing position during reading. *Experimental psychology*, *56*(1), 66. https://doi.org/10.1027/1618-3169.56.1.66

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443. https://doi.org/10.1037/0278-7393.31.3.443

Van Petten, C., & Luka, B. J. (2012). Prediction during language comprehension: Benefits, costs, and ERP components. *International Journal of Psychophysiology*, *83*(2), 176-190. https://doi.org/10.1016/j.ijpsycho.2011.09.015

Vitevitch, M. S. (1997). The neighborhood characteristics of malapropisms. *Language and Speech*, *40*(3), 211-228. https://doi.org/10.1177/002383099704000301

Vitkovitch, M., Potton, A., Bakogianni, C., & Kinch, L. (2006). Will Julia Roberts harm Nicole Kidman? Semantic priming effects during face naming. *Quarterly Journal of Experimental Psychology*, *59*(6), 1134-1152. http://dx.doi.org/10.1080/02724980543000178