How Linguistic Exposure Modulates the Acceptability of Long-Distance Dependencies

by

Abigail C. Bertics

B.S. Computer Science and Engineering
Massachusetts Institute of Technology, 2019

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author . . . .
Department of Electrical Engineering and Computer Science
January 21, 2022

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

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

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

How Linguistic Exposure Modulates the Acceptability of Long-Distance Dependencies

by

Abigail C. Bertics

Submitted to the Department of Electrical Engineering and Computer Science
on January 21, 2022, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

A central and still contested question in linguistics is "What makes a sentence good?" This thesis looks into one possible answer: the more you hear it, the better it sounds. More specifically, we are investigating what influences the acceptability of a certain type of long-distance, so-called 'filler-gap' dependency: object-extracted wh-question islands. We take a two-pronged approach. First, we look into how long-term, lifetime exposure to various components of a sentence (estimated from corpora) impacts its acceptability. We find support for the verb-frame frequency hypothesis (VFF; Liu et al. 2021a) and find that the frequency of the matrix-verb frame and the construction type in particular have statistically significant effects on acceptability ratings. It remains an open question how the individual components of a sentence combine to result in a single sentence acceptability judgement. The second prong takes partial inspiration from the mere-exposure effect in psychology (Zajonc, 1968). We investigate how short-term, within-experiment exposure to matrix-verb frame and construction type (the two components that influence overall sentence acceptability) affects the acceptability rating. Although no experimentally robust effect of short-term exposure, or priming, was found, we found a small, but statistically significant effect of order; acceptability ratings increase over the course of the experiment. Amazon Mechanical Turk (mTurk), used judiciously and with proper exclusion protocols, is a robust and powerful tool that is not worse than in-person experiments (Crump et al., 2013; Thomas and Clifford, 2017). Whether this small effect is happenstance or holds more generally has yet to be determined, but it should serve as a cautionary tale about running experiments where the incentive of the participants (e.g., to finish the task as quickly as possible to still get paid) and what the researcher is trying to study (e.g., how the human mind processes language) do not necessarily align.

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

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

Acknowledgments

I have the unique pleasure of thanking (and knowing) Annika Heuser. Thank you, Annika, for your curiosity and for indulging me with discussions on a vast array of topics, not all super relevant to either of our degrees. More concretely, thank you for the structure that you helped to provide when this thesis loomed larger than Mt. Olympus, and for the (much needed) editing that you helped me with. Thank you for being the best rubber duck, and for letting me know gently when I am making no sense. I do not think there exists a finite quantity of ice cream that could express my appreciation for you.

Thank you, Professor Robert C. Berwick, for welcoming me into your lab with open arms, your open-mindedness in allowing me to pursue this thesis topic, and your helpful comments.

Thank you, Professor Edward Gibson, for taking me under your wing and supporting me in this endeavor. Thank you for your guidance in choosing a project and designing and conducting experiments.

Thank you to Professor Francis Mollica for your invaluable guidance in the tumultuous but beautiful world of statistics. Thank you to Dr. Robert Ajemian and Professor Charles Yang for insightful conversation and pushing me to not settle. Thank you to the members of Bob's lab, who have allowed me to participate in academic discourse and entertained my questions.

Thank you to the volleyball team for maintaining my sanity.

Thank you to my family, who probably know a little bit more than they would like about the material in this thesis. Dory and mom, thank you for your unconditional support and love. And, of course, thank you to our three cats, Faina, Utka, and Misha; they emotionally supported me (but not my lungs) through the end of the writing.

Contents

Chapter 1

Introduction

Linguists can probe the human faculty of language by using measures like acceptability judgements, eye-tracking trajectories, or reading times. None of these measures, alone or together, is sufficient to truly get at how humans process and represent language. They are all heavily influenced by memory of prior experience, much like language. Understanding the link between experience and the language faculty is a crucial step towards a model of language acquisition.

Linguists have long focused on what makes certain sentences grammatical, often looking to the syntactic rules of combination and the nature of the lexical items. As such, they use grammaticality as a tool to build theories of the mental linguistic representations that people possess. Traditionally, a speaker's intuition in labeling a sentence as grammatical or ungrammatical has been the primary tool that has allowed linguists to probe at the human language faculty (e.g., Chomsky, 1957). A distinguishing feature of grammaticality is that it has traditionally been seen as binary: a sentence is either grammatical or it isn't. The contrast between these two categories can be used to discover principles about a language or about language as a whole.

Let us take a quick glance at the representations and mechanisms posited by linguists, with the help of adjudicating grammaticality judgments. Human processing of language is a cognitive process, and the brain is a "mental organ" (Chomsky, 1975). Brains "carry out computations," allowing abstract mental representations and the processing of the symbolic system of language. This is often referred to as Computational Theory of Mind and has been a very popular route of scientific inquiry for almost a century (e.g., Fodor, 1975; Gershman et al., 2015; Marr and Ullman, 1981; McCulloch and Pitts, 1943). Language, as processed by a computer-mind, is a static and passive entity. During production and comprehension, the mind builds structured representations from the atomized, building blocks of words. Given what we know about the computational properties of the brain (that it is dynamic, content-addressable, very noisy, and not at all computer-like) a discrete, static, serialized, and context-free approach to language might not be the best one. An alternative considers language processing as taking place in a dynamical system, where the lexicon is structured state space, and grammar can be thought of as the dynamics of the system that constrain movement in the state space. Thoughts, therefore, are not static symbolic entities, but rather paths through state space. This allows for a highly context-sensitive, continuous, and probabilistic conceptualization of language (Elman, 1995). It is clear that there is still a long way to go in determining how the human mind possesses a capacity for language.

While grammaticality is traditionally categorical and binary, acceptability is more nuanced. In this paper, we will investigate the phenomenon of acceptability. To make this tractable, we investigate the modulators of acceptability for two constructions that license long-distance dependencies: object-extracted wh-questions and object-extracted declaratives. We choose those structures because they are in a grey-area. They are neither completely natural, nor completely unnatural, and their acceptability fairly reliably is contingent upon the lexical item that acts as the main verb (Liu et al., 2021a). We also look into priming and its effect on acceptability judgements. Because the average acceptability of the object-extracted wh-questions is not near either boundary in the naturalness scale, we will be able to see changes in acceptability if they occur.

In this introduction, we will first orient ourselves around the difference between grammaticality and acceptability. And then we will discuss why we chose to focus on acceptability.

1.1 Grammaticality vs. Acceptability

First, it is important to put the preceding and following expositions into some greater context. Linguists have not managed to reach a consensus on even the "basics" of language, both in terms of defining it and determining what merits further study. It is not clear that language, at least when defined as the collection of sentences spoken by a community (e.g., Bloomfield, 1926), has any scientific coherence. And even if this abstract collection of strings of words does exist, it is not the same as the "faculty of language." The human capacity for language is innate, and is arguably unique to the human species (Berwick and Chomsky, 2016).

The minimalist tradition, along with proponents of generative grammar, view it as important to split knowledge of language into two categories: innate and learned. The innate component, often referred to as universal grammar, is a simple computation that combines two elements: Merge. This is the combinatorial backbone of language and its hierarchical structure. Initially, universal grammar was posited to contain a lot of linguistic information in the form of various general principles and specific parameters (see Chomsky and Lasnik, 1993), but more recently, the minimalist program has converged on a very impoverished universal grammar, just Merge. All the rest of language's properties that vary from one human language to the next must be acquired from experience. In order to learn these productive rules that seem to abound in external language, humans must leverage the input data they are exposed to and their learning mechanism(s) (e.g., Yang, 2016). For example, case marking, previously argued to be innate and universal, has been argued to be learned and not a part of universal grammar (Legate, 2021).

Grammaticality is a theoretical construct. It is not meant to be studied as it is, rather it is to be used as a tool to evaluate various theories. Acceptability is a phenomenon of external language use, tied up with external behavior and cognitive limitations. A theory of acceptability requires a coherent model of cognitive processing, something not yet realized. An issue here is that grammaticality and acceptability do not exist on the same "level" of analysis, and thus they are not directly comparable.

The claims made in this thesis about language, grammaticality, and acceptability are inherently limited to ensure finite scope and tractability. In any case, this thesis attempts to advance further the considerations of how to assess what makes a sentence good.

1.1.1 The Categorical Notion of Grammaticality

As language wielders, most adult humans have an intuition for whether or not a sentence is good, whether or not it follows the implicit rules of their language.

Linguists have long been interested in this. The first step on their journey was to define language, to characterize which strings of sounds or letters belong to the language and which don't. This computational definition of a formal language, as a set of strings over an alphabet, is still prominent today, used across domains from psycholinguistics to natural language processing.

Another prominent thought touts language to be the infinite use of finite means. If language is generative, then there exists an inverse function capable of discrimination. Perhaps comprehenders/listeners, receiving a string of sounds or letters as input, try to parse it into a structure. If it is indeed possible to generate given the rules / grammar of the language, then it is a full-fledged grammatical sentence. If you can't, then it is ungrammatical, and will sound "off" to a listener.

This distinction can appear unrelated to the meaning of the sentence. In a shocking citation that no one has ever expected, (1) is a grammatical English sentence without a reasonable meaning, while (2) is ungrammatical but with an inferable meaning.

(1) Colorless green ideas sleep furiously. (Chomsky, 1957)

(2) [*] I walk to store and chocolate buy.

1.1.2 Acceptability instead of Grammaticality

Many papers on language from the 1900s through today have relied on this grammatical introspection as a primary source of data. Also, one of the great difficulties of being a student taking a Syntax class is that as you sit on your armchair and ponder the grammaticality of a sentence, the more you repeat it to yourself, the better it sounds (until it ceases to have any meaning to you whatsoever). The lack of reliability and systematicity of introspection as a method has been a serious barrier to principled research in the field of syntax (Schütze, 1996). There has been a push for more quantitative (and therefore replicable) paradigms for linguistic judgements (e.g., Mahowald et al., 2016a).

Linguists have long recognized that grammatical sentences can be unacceptable for a slew of extra-linguistic reasons like memory or contextual limitations. There are grammatical sentences that are not super acceptable (e.g., center embeddings). And reasonably acceptable ones that are not grammatical (e.g., center embeddings missing a verb or certain agreement attraction errors). In most cases, most people agree; in other cases, class-struggle and entrenched power dictate grammaticality. What is becoming less clear is whether there exists grammaticality pure in language, untainted by the extra cognitive factors. What is clear is that there is no way to empirically measure grammaticality (see Schütze, 1996, for a more detailed argument).

Usually, the performance-competence distinction is brought forth to explain why theoretical syntacticians care deeply about grammaticality and eschew investigation into acceptability (Chomsky, 1965). Competence refers to a speaker's knowledge of their language, the rules, the theoretical ability to comprehend and produce. Performance, on the other hand, is the speaker's actual ability to comprehend and produce language. Humans have an ideal grammar in their heads, and comprehension and production are affected by extra-linguistic factors: memory, attention, etc. Chomsky (1986) makes the distinction between internal language and external language. Those in his program take the internal language to be the true object of linguistic study, even though it is, by definition, un-studyable.

Language is in theory a perfect infinite use of finite means that is limited in practice by the worldly constraints of the human mind and brain. If we, as scientists, want to study language, we must deal with the data at hand. There is no way to probe this internal language, in which we have our competence. It seems to me like this pure notion of language is unfalsifiable, as it by definition changes with measurement. If our research is to be grounded empirically, the only information available to us to be measured (i.e., acceptability judgements, elicited production, reading times, etc.) is "tainted" by human performance. Language, after all, exists in the real world, so we might as well study it there.

1.1.3 A Continuous (Rather than Binary) Scale of Acceptability

Although there are some sentences that are clearly natural to a listener, and some sentences that are clearly unnatural, there exist many in between. "The cat that loves mice drink milk." is not great; it is prescriptively ungrammatical. But, it is (at least according to my subjective judgement) much better than "Milk mice the loves that cat drink." There needs to be room for different shades and types of acceptability.

Across domains and levels categorical divisions rarely survive careful scrutiny. When you zoom in on the hard and fast distinction between two categories, you soon realize that that boundary does not exist. Any human-centered theory of language processing must account for gradient acceptability.

If the acceptability of an instance of a given construction is not a binary yes/no decision based on grammatical rules, then what is it? We have four main levels to consider: the constructions (the "grammar"), the lexical components (the "words"), the context, and the interlocutors. It is most probably some combination of the acceptability of all four, but not reducible to simply their individual acceptabilities. Acceptability is subject to all cognitive processes that dare entangle themselves. Acceptability is subjective, and we lose valuable information if we reduce it to ones and zeroes.

1.1.4 Collocations

To drive home the argument that acceptability is an important and different measurement from grammaticality, let us take a gander at a pervasive linguistic phenomenon: collocations. A collocation is a word sequence that appears unusually often in spoken or written language, an intimately connected natural cooccurence of words. It is almost like an idiom, but a collocation's meaning is reducible to the meanings of its parts. For example, "salt and pepper" is a collocation, sounding to most, if not all, native English speakers much better than "pepper and salt." Other examples are as follows: a sense of pride (vs. a feeling of pride), quick brown fox (vs brown quick fox) , blissfully unaware (vs. joyfully unaware), etc. They are cognitively very interesting, because nothing in the collocation's syntax, nothing about its "grammaticality", would indicate why one is better than the other, why one feels awkward. The existence of collocations is an important reason to consider gradient acceptability measurement. Their naturalness does not distill down to the acceptability of their components, nor to the acceptability of the various rules that construct them (McKeown and Radev, 2000).

In addition, collocations come to and maintain their existence in language through repeated use and exposure. They are a key example for how long-term exposure drives acceptability. There is no other reason why "salt and pepper" is more acceptable than "pepper and salt" beyond mere exposure.

1.2 Why Acceptability?

Acceptability of sentences is a well-accepted (pun intended), but not well-understood phenomenon. Understanding how general (corpus) and short term (priming) frequency modulate acceptability judgements will provide insight into human language processing and move us closer to a partial theory of what makes a sentence good. This matters because a comprehensive theory of acceptability would help us better evaluate frequentist, deep learning models like GPT-3 (Brown et al., 2020). Those models learn and operate based solely on exposure to text that they are fed, which accumulates to pretty much the entire internet. If we learn how exposure to sentences affects a human learner's knowledge of language, we can compare, but mostly contrast, that with current state-of-the-art natural language processing models. Also, frequency and priming are potential confounds in many psycholinguistics experiments. Understanding how these affect acceptability can help to mitigate that and allow linguists to better understand the data. This thesis investigates how both long-term and short-term exposure modulates the acceptability of a certain type of ambiguously acceptable structure: the object-extracted wh-question island.

Chapter 2

Related Work

This chapter will introduce the material necessary to understand the three experiments in this thesis. The experiments try to determine how both long- and short-term exposure affects the acceptability of two specific island constructions. First, we introduce the reader to the specific island constructions that compose the experimental items. Then, we discuss various theories that purport to explain their acceptability or lack thereof. Finally, we turn to the effects of short-term exposure and introduce the mere-exposure effect and linguistic priming.

2.1 Long-Distance Dependencies / Islands

Long-distance, or filler-gap, dependencies have long interested linguists because they are hard for humans to process; they complicate language processing. Sentences as easy to read as (3) fascinate the linguist, because of the non-local relationship between the **filler** (i.e., "what") and the **gap** (i.e., the filler's canonical position directly after "eat", written as _, also sometimes referred to as the trace). Filler-gap dependencies can be very long, as in (4); they are theoretically unbounded in size, but as their length increases, the burden and difficulty of successfully processing them does too

(Hawkins, 1999).

(3) What$_i$ did you eat _$_i$ for dessert?

(4) What$_i$ did you say that Kay said that she wanted to eat _$_i$ for dessert?

While the filler-gap dependencies in (3) and (4) are both acceptable, others are less so, as in (5) (Ross, 1967). Those less-acceptable sentences are referred to as "islands." The island traps a constituent on it, because it is unacceptable to extract from the gap's location across certain implicit boundaries (apparently boats don't exist).

(5) a. * What did Bill buy potatoes and _?

b. * What did John fall asleep and Bill wear _?

c. * What did Alex go to the store because he needed to do _?

d. * Who did Oleg ask why Igor was waiting for _?

For the purposes of this thesis, we will focus on a weaker type of island: object-extracted wh-questions. These islands are created when an object is extracted out of sentential clauses that are complements of certain verbs. Certain verbs (often referred to as bridge verbs) license the extraction and create legal sentences, as in (6). Other verbs, as in (7), result in island-like behavior (Erteschik-Shir, 1973).

(6) What did Samia say that Phoebe stole?

(7) *What did Lana shout that Taylor stole?

Now, it would be fallacious to say that (7) is categorically ungrammatical, even though it is referred to in the literature as such; if it was, this thesis would be done here, nothing interesting to read. It is, however, less acceptable than (6). This thesis concerns itself with these non-bridge-verb islands, investigating why some are more acceptable than others and seeing whether repeated exposure to them renders them more acceptable.

There are many proposed explanations to the static (un)acceptability of islands that can be sorted into three main categories: syntactic, pragmatic, and frequency-based. See Liu et al. (2021b) for a thorough review.

2.1.1 Syntactic Explanations

Linguists (i.e., Chomsky et al., 1977; Ross, 1967) have long looked to syntactic rules in order to explain the unacceptability of islands. Chomsky, in particular, thought of islands as ungrammatical, and defined a structural constraint called “Subjacency” to argue for that. Subjacency says that you can not move something across more than one bounding node (Chomsky et al., 1977). This implies that when licit wh-movement occurs, it must be doing so in an iterative manner, only crossing one phase boundary, or one bounding node at a time. Subjacency is a general linguistic principle, and this structural interpretation is construction- and language-agnostic. Movement across two bounding nodes is ungrammatical, regardless of the semantic and lexical properties of the words, and regardless of the type of filler-gap dependency (wh-question, cleft, declarative, etc.). This conveniently fits into Chomsky’s view of language, where syntax and semantics are separated.

There are other syntactic accounts of islands. One more quantitative definition proposed by Sprouse is that island sentences are "super-additively" bad, meaning that the unacceptability of an island sentence is more than the sum of the unacceptability of the components of the sentence (Sprouse, 2007). This super-additivity is interpreted to be syntactic in nature.

However, islands are not categorically ungrammatical or unacceptable. There are many counter-examples of acceptable islands, of multiple flavors, see those in (8) (collected by Liu et al. 2021b but originally from Ross 1967). In addition, the meaning of the verb and the construction type both have an impact on the acceptability of movement, so Subjacency does not adequately account for the island phenomena that we are studying.

(8) a. The funds that I have [hopes the bank will squander _] amount to more than a billion. (Complex NP island)

b. Of which cars were [the hoods _] damaged by the explosion? (Subject island)

c. He told me about a book which I can't figure out [whether to buy _ or not]. (Wh-island)

It doesn't make a lot of sense to pursue a purely syntactic explanation for the badness of islands, because it seems like the meaning of the sentence and the lexical qualities of various components have import, and it is unclear how a structural account would integrate the variability in acceptability.

2.1.2 Pragmatic Explanations

An alternative group of explanations argue that the island effect is not ungrammatical, but rather infelicitous, meaning that it clashes pragmatically, rather than syntactically. Generally speaking, the pragmatic arguments go as follows. (7) is infelicitous because when you use a verb like "shout", the main information delivered by the sentence is the manner of speaking. The verb in-and-of-itself is the focal point; thus, extraction from the complement, which is not dominant, is not licensed. Griceanly speaking, a speaker would only use the word "shout" if they care about conveying the shouting, otherwise, they would violate the maxim of quantity. Trying to introduce two new pieces of information, the shouting and the complement, is too much, so the sentence reads poorly. (6), on the other hand, is felicitous because "say" is a pretty boring linking verb, so the complement of it can be the new information being introduced. You can theoretically nullify the infelicity of (7) by making the manner of speaking old news: Lana was shouting about Taylor stealing something and buying something else. What did Lana shout that Taylor stole?

(9) a. What did Isabella buy a magazine about?

b. *What did Isabella misplace a magazine about

An approach linked to the Gricean view deals with relevance, and states that extracted components must be highly relevant to the discourse (Chaves and King, 2019), see (9). A different approach focuses on focus. Movement out of a clause is an operation that emphasizes that extracted component. This emphasis can only occur from "dominant" clauses, from the element that carries the crux of the information of the sentence (Erteschik-Shir, 1973). Another approach focuses on relevance to discourse. The "Topichood Condition for Extraction" maintains that the only elements that are legal to undergo extraction are those that can plausibly be the topic of the sentence (Kuno, 1987). A later approach focuses instead on "backgroundedness". Extraction is prohibited from backgrounded elements, those that are not the focus or the topic of a sentence (Goldberg, 2013). This backgroundedness is gradient, allowing for gradient acceptability, depending on the context. A similar approach combines previous approaches, hypothesizing that the unacceptability of islands comes from a focused element being part of a backgrounded constituent (Abeillé et al., 2020). Properly defining backgroundness is a crucial step in the development of that hypothesis.

2.1.3 Frequency-based Explanations

A frequency-based explanation puts forth that the acceptability of certain filler-gap constructions is modulated by the frequency of linguistic exposure. As shown in (6) and (7), verbs that infrequently take a sentential component, like "mutter," will result in lower acceptability than verbs that frequently take one, like "say." Liu et al. (2021a) found that there are strong, independent effects of construction type (wh-question, cleft, or declarative sentence) and the verb-frame frequency (the joint probability of the verb and it taking a sentential complement). Unlike the super-additive account (Sprouse, 2007), these effects were found to be independent, or merely additive. Liu et al. proposes the Verb-frame Frequency Hypothesis (VFF), taken directly from the paper and shown in (10). The VFF has to do with correlation; it does not propose any mechanism for how the frequency of the matrix-verb frame and the construction affect the acceptability of the sentence. It does allow for gradient acceptability, depending on the lexical make-up of the sentence.

(10) *The Verb-frame Frequency Hypothesis* (Liu et al., 2021a):
The acceptability of a sentence is best captured by two independent effects: (i) the frequency or the type of the construction (e.g., wh-questions vs. declaratives) and (ii) the frequency of the verb head-structure, P(matrix verb, sentence complement) = P (matrix verb) * P (sentence complement | matrix verb).

A more general way of thinking about the VFF is that exposure affects acceptability. They estimate that long-term, lifetime exposure using the Google books corpus. But, why would some types of exposure be privileged over others, given that a speaker's knowledge of language is dynamic, as they never cease to be a language learner, continually incorporating new information. A natural follow up question is: if long-term exposure increases acceptability, would short-term exposure also increase acceptability?

2.2 Priming

Short-term exposure is also known as priming. Speakers tend to reuse phrases or structures that they recently have encountered (Bock, 1986). There are two main types of priming: syntactic and lexical. Syntactic has to do with the underlying structure or construction, while lexical priming has to do with the actual lexical units, or words. You, yourself, have probably encountered priming in the wild, adopting your friends' turns of phrase unwittingly.

There is no settled-upon model for how these short-term exposures affect the capacities of production or comprehension. However, it makes sense in the context of one known quality of the brain: an associative memory. In addition, syntactic priming provides an interesting window into the mental representations of language, providing evidence of shared structure even when all of the words are ostensibly different (Branigan et al., 1995).

2.2.1 The Effect of Priming on Production

Most experimental research in priming has to do with priming in production. If you are initially exposed to words or syntactic structures, you are more likely to produce that specific structure or word later on, when prompted. To measure the effect of priming, scientists usually calculate the likelihood of producing one structure as opposed to the other, or measure latency to production. For example, some verbs have two ways of use (e.g., "I baked you a cake". vs. "I baked a cake for you.") A participant having heard a sentence with the former structure will be more likely than normal to produce a sentence with that structure than the alternative (Bock, 1989).

The effect of syntactic priming in production, in particular, is robust across different types of structures and across different languages. Lexical priming, when it overlaps with syntactic priming, strongly contributes to the effect (Mahowald et al., 2016b).

2.2.2 The Effect of Priming on Acceptability

It is quite probable that language production has at least some representations or mechanisms in common with comprehension. Therefore, it would make sense for priming to also have an effect on sentence comprehension and acceptability judgements.

Luka and Barsalou conducted a series of five experiments, where participants rated grammatical sentences of English and later rated identical, structurally similar, or novel sentences for grammatical acceptability. They found that participants rate sentences that they had read before as more grammatical. Participants also rate sentences that have shared syntactic structure but no shared content words more highly. The first exposure resulted in the strongest priming effect, but subsequent exposures still increase the effect (Luka and Barsalou, 2005). Zervakis and Mazuka did not find that merely reading sentences leads to a detectable priming effect, but instead that rating sentences for acceptability leads to a detectable priming effect (Zervakis and Mazuka, 2013). Reading times for sentences with previously low acceptability have been shown to decrease with repeated exposure (Hofmeister et al., 2013), showing an analagous effect in the mechanisms of online sentence processing. There is a bit of a replicability crisis with regards to reading times, due to the very small effect size, so that should be taken with a grain of salt.

There is less consensus on the effect of priming in comprehension, and it has been argued that changes in acceptability and reading time can be attributed to an increase in fluency of comprehension due to the mere-exposure effect rather than a true change in acceptability (Zervakis and Mazuka, 2013). That doesn't necessarily mean that priming does not occur, however, because the mere-exposure effect is plausibly a driving force behind the effects of priming (see below in Section 2.2.4).

2.2.3 Syntactic Satiation in Subject Islands

This priming phenomenon in comprehension is sometimes referred to in the literature as "satiation." Syntactic satiation, in particular, refers to the increase in acceptability that occurs due to repeated exposure of "ungrammatical" sentences (Snyder, 2000).

There is a fair amount of controversy about the existence of satiation, especially as it concerns subject islands. Some have found evidence for satiation (e.g., Francom, 2009;

Snyder, 2000). However, some experiments have failed to replicate this result; in particular, Sprouse (2009) argued that the satiation is due to a confound—an unbalanced experimental design. Sprouse (2007) went on to argue that non-satiation should be the result, because islands are ungrammatical, and therefore there should be nothing to satiate; syntactic priming can not occur for ungrammatical constructions. This argument does not hold if one assumes non-syntactic reasons for the relative unacceptability of islands. Crawford (2012) found a satiation effect for some constructions, but not for subject-islands. Chaves and Dery (2014) found that subject-island satiation exists, but does not occur across-the-board; only certain subject islands reliably show priming effects. The failure to satiate occurs when the stimuli are too complex or incoherent to begin with.

Crucial to the detection of priming, or syntactic satiation, is allowing intermediate gradient judgements. Brown et al. (2021) find an initial boost in acceptability that levels off with repeated exposure, ending up in a middle-zone of acceptability, neither completely natural nor completely unnatural. Satiation is an important phenomenon for linguists to understand, because they themselves are very susceptible to it, over the course of an hour, or over the course of their careers. If anything, satiation (which we will refer to as priming from here on out) is another facet to consider when trying to figure out what makes a sentence good.

2.2.4 Mere-Exposure Effect

The mere-exposure effect is a psychological phenomenon in which repeated exposure to something, even without conscious recognition of it, results in people liking it more. This is essentially the same thing as familiarity preference. In his seminal work, Zajonc found a correlation between word frequency and positive connotation. They also manipulated exposure of nonsense words and measured their perceived connotation, concluding that mere repeated exposure to a stimulus increases the positive perception of it (Zajonc, 1968). Later, similar results of a familiarity preference were shown across domains of stimuli, using a variety of methods, and even on different species (Bornstein, 1989).

The proposed explanation behind the mere-exposure effect is as follows: repeated exposure to a stimulus makes it easier to process, or increases *perceptual fluency*. Things that are easier to process, we tend to like more (Bornstein and D'Agostino, 1994).

This evokes a lot of interesting questions. If exposure to a word or construction increases its acceptability, it must be at least partially due to the mere-exposure effect, which is domain-general. How susceptible are linguistic acceptability judgements to extra-linguistic processes? We know for sure of this one domain-general process by which extra exposure yields increased palatability; could this be the foundation of all acceptability judgements? Is the mere-exposure effect a component of linguistic acceptability? To what extent does the mere-exposure effect influence our language acquisition and production?

Chapter 3

Experiment 1 & 1b

This thesis will take you on the journey of the research process. Pitfalls, mistakes, iterations will be discussed. Scientific writing often only references the polished final results, talking about what worked. Arguably more important, and definitely a huge part of growing and learning as a scientist, are the things that didn't go as planned, forcing you to adapt and overcome. If all you care about is publishable results, this thesis is not for you.

3.1 Experiment 1

The purpose of Experiment 1 is two-fold. First, we want to confirm the effects of long-term exposure on acceptability judgements found in Liu et al. (2021a). Their verb-frame frequency hypothesis (VFF) states that the acceptability of filler-gap dependencies is proportional to the frequency at which the main verb takes a sentence complement. A wh-question "island" using a very common S-complement verb, like "said", will be more acceptable than a wh-question "island" using a rare S-complement verb, like "mutter." In addition, the frequency of the construction impacts its acceptability; declaratives are more acceptable than wh-questions. The VFF predicts no interaction between the impact of the verb-frame frequency and the construction type/frequency on the acceptability.

The second and main goal of Experiment 1 is to characterize any short-term exposure effects on those constructions. The frequency-based account described above uses a corpus-based approach to estimate life-time exposure to various verb-frames and constructions. However, given a frequency-based account of acceptability, there is no reason to think that a language wielder's internal frequency estimates and acceptabilities would be static. In fact, we are exposed to new constructions and new lexical items all the time, and we have to continuously integrate that new information into our knowledge of language. This experiment seeks to determine if repeated exposure to particular verb-frames and particular constructions affects their acceptability.

3.1.1 Design & Materials

This design is an extension of Liu et al. (2021a), with a few modifications, the main of which is the restructuring of the experiment into three phases.

Design of the Materials Lists and Experimental Groups:

We ascertain from each subject 180 acceptability judgements, of which 60 are experimental items over 6 matrix verbs (10 judgements per verb). The judgements are ostensibly split into three stages, but the presentation of each item does not change based on the stage. The partitions are:

1. **Initial test.** First, we get baseline acceptability judgements for each matrix verb by soliciting an acceptability judgement for both a declarative sentence and a wh-question (12 sentences total).

2. **Training.** We expose participants to 3 more declarative sentences and 3 more

wh-questions per matrix verb (36 sentences total), getting acceptability judgements each time.

3. **Final test.** To end, we repeat step one, getting final acceptability judgements for each matrix verb in both declarative and wh-question form (12 sentences total).

We use forced-choice binary acceptability judgements, as was done in Liu et al. (2021a). This simplifies data analysis, and previous studies (e.g., Fanselow and Weskott, 2011; Sprouse et al., 2013)) have shown that similar results are given regardless of using Likert or binary scales for acceptability judgements. So, we can simplify analysis without sacrificing experimental quality.

Unlike Liu et al. (2021a), we will use fillers. As this is an experiment conducted via mTurk, strategic effects are a real concern here. We have a 2:1 ratio of distractor to experimental items. Each distractor includes two proper nouns (like the experimental items) and are matched for word length. Half of the distractors are wh-questions and half are declaratives, like the experimental items.

Construction of the experimental sentences:

Each participant will be shown sentences constructed from the following 6 matrix verbs (verb_matrix): hate, realize, say, mumble, whine, suspect. We chose two high-frequency, two medium-frequency, and two low-frequency (in terms of verb-frame) matrix verbs from Liu et al. (2021a) to ensure proper coverage of the possible distribution.

Those matrix verbs are combined with the following 10 embedded verbs (verb_embed), all past-tense forms of 10 of the 25 most frequently used verbs in English: bought, wrote, saw, took, knew, broke, wanted, liked, needed, ate.

The subjects of the matrix and embedded clauses (noun1 and noun2, respectively) will be two names randomly selected from the 50 most common girl and 50 most common boy names from 2020 (see Appendix A).

The declarative and wh-question sentences are constructed as follows:

- declarative:

  noun1 [verb_matrix] that [noun2] [verb_embed] something.

  - ex: Mason noticed that Elena took something.

- wh-question:

  - What did [noun1] [verb_matrix] that [noun2] [verb_embed]?
  - ex: What did Mason notice that Elena took?

We use "something" as the embedded object in the declarative constructions to reduce the possibility of semantic confounds.

Construction of the filler/distractor sentences:

The standard filler sentences are matched to the experimental items in terms of length, number of proper nouns, and percentage of wh-questions. How exactly they are generated and examples of them are shown in the provided code.

Comprehension question design:

In order to make sure that people are minimally reading the sentences and paying attention, we have comprehension questions to answer after each acceptability judgement. These questions are perhaps too easy and of the form: Does the target sentence mention Colton? Half of the comprehension questions have "Yes" as the correct answer and half have "No."

3.1.2 Hypotheses and Predictions

As a quick summary of the experiment, each participant will make acceptability judgements of six matrix verbs in wh-question and declarative constructions at the beginning and at the end of the experiment. They will go through a training stage between the initial and final testing stages, which will expose them to the training stimuli, where learning may or may not occur. This first experiment serves mainly as an exploration, in order to assess the feasibility of using mTurk studies to measure changes in acceptabilities for these constructions.

First, we anticipate that the experimental results will be consistent with the VFF account, in that the acceptability judgements are best explained by two independent, fixed effects: the frequency of the construction type and the frequency of the matrix verb taking a sentential complement.

Separately, we want to see if extra exposures to the verb-frames results in an increased naturalness rating. In other words, we want to see if presentation order is a fixed effect on the acceptability rating (see Section 3.1.4 for a more precise definition of presentation order).

Multiple hypotheses might account for a presentation order effect. Often, in psycholinguistic experiments, presentation order is a possible confound, mitigated by randomizing the order of the materials for each participant; in this case, it is the effect that we are looking for. Given a frequency-based account of acceptability, we might think that the more you are exposed to a given construction, or a given matrix verb, even in the short-term, the more acceptable that construction and matrix verb will become. This notion also follows from the priming literature. When you are exposed to a sentence, you will be more likely to produce words in that sentence or to use constructions in that sentence later. There is also the inclination to rate a sentence higher if you have previously heard it. There is no reason to expect that humans do not learn and adjust their language processing to account for experimental stimuli, just as they do to sentences that they read in The New Yorker. This frequency-based account would suggest that the more you are exposed to a sentence, the better it becomes. In addition, one might even expect that the acceptability of the rarer matrix verbs might change more than the acceptability of the more common matrix verbs. Exposures to frequent items, much like adding one pound to a thousand pounds, would change things much less than exposure to rare items, like adding one pound to ten pounds. To wildly speculate, we might expect a Weber's Law-esque thing to be at play here, where increases in acceptability ratings per exposure are inversely related to the initial verb-frame frequencies. In practice, this might be very difficult to detect through often very messy data. But, this remains an interesting and open question: how does knowledge of language change throughout a life time? How do language users incorporate new information into their language knowledge?

There are other possible hypotheses that could still result in a presentation order effect. Among these include that perhaps participants get nicer as the task goes on, or perhaps they learn something about the distribution of sentences, and they adjust their ratings closer and closer to what they expect that to be.

If there is no presentation order effect, this means that short-term learning does not occur, or at least that priming does not impact acceptability judgements. This could be due to a plethora of reasons, but this is the null hypothesis. If this is the result, perhaps further investigation along these lines is less than prudent.

3.1.3 Participants

We ran this experiment via mTurk with 56 participants. This experiment was exploratory in nature, and we did not have a principled way to estimate the expected effect size, so we chose 56 quite arbitrarily. The experiment was only available to those with U.S. IP addresses. Participants were asked to self-report their native language and country of origin; their responses to those two questions did not impact pay. Each participant was paid 6USD for an estimated maximum of 30 minutes of work.

3.1.4 Results

Our analysis is conducted on the results from 42 participants. We excluded 2 participants who self-reported to be non-native American-English speakers or to not be from the U.S. We also excluded 12 participants who did not answer the comprehension questions with at least an accuracy of 85%. This is an exclusion rate of 25%.

We used mixed-effects logistic regression (the *lme4* package in *R*) to analyze acceptability ratings (Bates et al., 2015). We considered the following variables in our regression:

1. **rating** is the binary decision given by participants, where they chose Unnatural or Natural.
2. **centered_log_freq** is the log-transformed frequency of the matrix-verb frame, centered around zero.
3. **sentence_type** is a binary variable, either declarative or wh-question.
4. **pres_order** is an integer from 1 to 5, the i-th time that particular matrix verb in that sentence type.
5. **WorkerId** is the unique identifier for each participant
6. **item** is a unique identifier for each combination of matrix- and embedded- verbs

The formula was:

$$\textbf{rating} \sim \textbf{centered\_log\_freq}*\textbf{sentence\_type}+\textbf{pres\_order}+ (1|\textbf{WorkerId})+(1|\textbf{item})$$

We first attempted to fit the model using a maximal random effects structure, but, alas, it did not converge, so we minimally reduced the random effects until convergence, which in practice meant leaving only random intercepts and no random slopes (probably due to the paltry size of the data).

The results confirm the VFF. There was a significant main effect of the log-transformed frequency of the matrix-verb frame ($\beta = 0.75$, $z = 9.05$, $p < 0.001$), and there was a significant main effect of sentence type ($\beta = -2.15$, $z = -11.4$, $p < 0.001$). No interaction was found ($p > 0.3$). These beta values are very similar to those found in Liu et al. (2021) ($\beta = .59$ and $\beta = -2.45$, respectively), indicating a successful replication.

The new finding here is that presentation order was a statistically significant predictor of rating ($\beta = 0.23$, $z = 1.6$, $p < 0.001$), meaning that the acceptability of a sentence increases the more times you hear something very similar to it. This is, however, a smaller effect than matrix-verb-frame frequency or sentence type.

3.1.5 Analysis / Discussion

It is important to note (again) that this experiment was intended to be exploratory; its results can intrigue, but must be taken with a large grain of salt.

Experiment 1 shows that verb-frame frequency and sentence type explain acceptability judgments, in an independent, additive manner. But, this experiment also shows that those two predictors are not the only things at play in modulating acceptability.

Figure 3-1 clearly shows a positive, monotonically increasing relationship between presentation order and rating of wh-questions; the more times you see a certain matrix-verb frame in wh-question format, the more natural it sounds. Interestingly, Figure 3-1 does not show a simple linear relationship. The increase in acceptability from exposure 1 to exposure 2 dwarfs the increase in acceptability from exposure 2 to

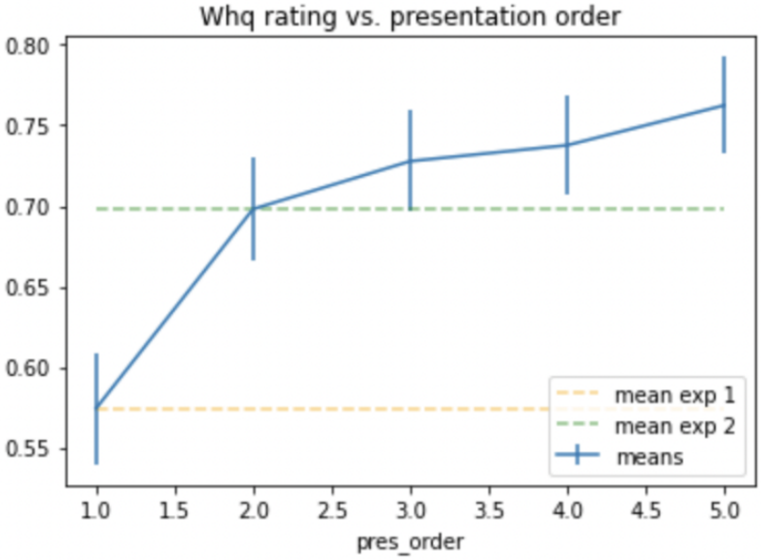


Figure 3-1: Results of Experiment 1: Mean rating of wh-questions plotted against their presentation order. Wh-questions start with a lower acceptability than their declarative counterparts, so any changes would be more qualitatively visible. For easier comparison, the orange dashed line shows the mean acceptability of the participants' first sighting of each matrix verb in wh-question format, and the green dashed line shows the mean acceptability of the second time.

3, 3 to 4, 4 to 5, or even 2 to 5. To the naked eye, absent any hard statistical proof, one might venture to say that it looks logarithmic.

There seems to be something special about the first exposure, priming that specific lexical item (the matrix verb) or bringing that specific syntactic structure (the object-extracted wh-question) to the foreground of the mind. Additional exposures still prime, but to less of an extend than the primary one.

There should be red flags going up, sirens going off right about now, given the wildly speculating nature of our above analysis. There are a few issues with this experiment, the most egregious being that there is no control group. This experiment subjects all participants to "training"; all participants see each matrix verb in wh-question format five times. As it stands, there is no way to know if the increase in acceptability is simply participants getting nicer as the task goes on, or them habituating to the task by better understanding how the mechanics of the survey work, or them catching on to and mimicking the average acceptability of the items (which is above that of wh-questions). We need a control group to see whether the repeated exposures to the stimulus are indeed increasing its acceptability, or whether it is the experimental conditions that are doing that. This brings us nicely to Experiment 1b.

3.2 Experiment 1b

Experiment 1b is essentially the same as Experiment 1 except that we add a necessary control group. Having a control group will hopefully help to tease apart two possibilities, whether the increase in acceptability over the course of the experiment is due to repeated exposures, or whether it is simply due to participants rating things as more natural as they get adjusted to the experimental task.

3.2.1 Design & Materials

The design of Experiment 1b only differs from that of Experiment 1 in the training stage:

2. **Training.** We split the participants into two groups: training and control. The training group is exposed to 3 more declarative sentences and 3 more wh-questions per matrix verb (36 sentences total), getting acceptability judgements each time. For the control group, we expose them to 36 non-island control sentences with approximately the same average naturalness rating as the experimental items.

The filler sentences for the training phase of the control group are slightly different than the normal fillers. It was important to match them to the acceptability of the experimental items to reduce possible confounds. If the acceptability during the "training" phase of the control group is significantly above that of the train group, then their acceptability judgements might increase during the final testing, due to habituation to the task stimulus. In crafting these filler items, we used the average wh-question naturalness rating acquired from Experiment 1 as a target. We combined normal English sentences with a few types of known grammatical but unnatural sentences (e.g., a semantically implausible double-object construction like "Louise and Mia gave the candle the boy.") to reduce the acceptability to something more like that of the wh-questions.

3.2.2 Hypotheses and Predictions

To reiterate Experiment 1b, each participant will make acceptability judgements of six matrix verbs in wh-question and declarative constructions at the beginning and at the end of the experiment. Half of the participants will go through a training stage, where they are exposed to those same matrix verbs three more times in each construction, and half will be the control group, seeing acceptability-matched distractor sentences instead. The purpose of this experiment is to again characterize if and how these additional exposures impact the acceptability judgments.

This time, rather than looking at trends in acceptability over five exposures, we are directly comparing the initial and final acceptability judgements for both the train and control groups. There are multiple possibilities that could be happening with this "learning" paradigm, which distill down to one possible fixed factors (presentation order) and the interaction of that factor and training group. It doesn't make sense for training group to be a main effect, because both groups receive the same materials for the initial testing phase. There is no reason or plausible meaning to one of the groups starting out with a higher acceptability; given enough participants, our random assignment of participants to groups should control for this.

There are $2^2 = 4$ total possible outcomes, described in Table 3.1.

1. Both presentation order and its interaction with training group are significant —Both groups have increased acceptability of the stimuli from initial to final testing, but the training group more so.
2. Presentation order is significant; its interaction with training group is not —Both groups have indistinguishably increased acceptability across the course of the experiment; training does not seem to have an effect.
3. Presentation order is not significant; its interaction with training group is —The control group did not exhibit an increase in acceptability from the first time to the second time they see a matrix-verb-construction combination; the training group does show an increase in acceptability from the first time to the fifth time. This means that presentation order is only significant for the training group.
4. Neither presentation order nor its interaction with training group are significant —There is no difference between groups and there is no improvement with repeated exposure to the stimulus. This is effectively the null hypothesis.

Table 3.1: Possible outcomes of Experiment 1.

Given our results in Experiment 1, and given a frequency-based account of acceptability, we might think that seeing a matrix-verb in a construction five times will yield higher acceptability than only having seen it twice, but both repeated exposures still result in increased acceptability (possible result 1 in Table 3.1).

Another possibility is that both the training and control groups experience the same change in acceptability from the beginning to the end (possible result 2 in Table 3.1). This could be because the increase in acceptability from one exposure to two dwarfs any changes thereafter. This could also be due to habituation to the experiment. Perhaps participants get nicer as the task goes on, or perhaps they learn something about the distribution of sentences, and they adjust their ratings closer and closer to what they expect that to be.

If there is no presentation order effect, but there is an interaction (possible result 3 in Table 3.1), that would most likely be the case if the training group experienced an increase in acceptability, but the control group did not. This could be the case if priming does indeed increase acceptability, but just one exposure is not enough. Or, if priming increases acceptability, but the time between exposures needs to be small (i.e., the 120 sentence training was too long).

If there is no presentation order effect at all, and also no interaction (possible result 4 in Table 3.1); this is the null hypothesis.

3.2.3 Results

We ran this experiment via mTurk with 56 new participants for the control group, with the same meta-conditions as Experiment 1.

Our analysis will be conducted jointly on 85 total participants —42 from Experiment 1, who will be coded to be part of the training group, and 43 from Experiment 1b, who will be coded to be part of the control group. This round, we excluded 2 participants who self-reported to be not from the U.S and 11 participants who did not answer the comprehension questions with at least an accuracy of 85%. This is a 23% exclusion rate, similar to that of Experiment 1

We mentioned the importance of having the control group's distractor sentences matched in terms of naturalness with the experimental items from the training group to reduce possible confounding factors between the groups. It appears that our hack was effective; we used a two-sided t-test to compare the control group's particular fillers against the training group's training items. There was no significant difference in mean acceptability between the control group (mean=0.79; SE=0.024) and the training group (mean=0.81; SE=0.024); t(83)=1.054, p=0.29.

Now onto the experimental data. Similar to what we did in Experiment 1, we used mixed-effects logistic regression to analyze acceptability ratings. We considered the same variables as in Experiment 1, but with one new addition: **with_training** which is a binary variable that is 1 if the participant received training (i.e., extra exposures of the matrix-verb in the wh-question and declarative constructions) and 0 if they are a control.

The formula was similar to last time, but with an added interaction term (in bold):

rating ~
centered_log_freq*sentence_type+pres_order+**pres_order:with_training**
+(1|WorkerId)+(1|item)

We only fit the model on the acceptability judgements from the initial and final testing stages for both the training and the control groups. These two sets of judgments are all that is directly comparable between the groups, and this is to ensure that the model does not overfit to the training group data (because by including the training judgements, there would be more than twice as much of it).

There is a significant main effect of the log-transformed frequency of the matrix-verb frame ($\beta = 0.45$, $z = 6.9$, $p < 0.001$) and of sentence type ($\beta = -1.85$, $z = -11.9$, $p < 0.001$). This time, an interaction between those two was found ($\beta = 0.38$, $z = 3.5$, $p < 0.001$). These main-effect beta values are very similar to those in Experiment 1, which is not unsurprising, given that half of this data is from that experiment. What is surprising is the interaction.

With the addition of a control group, there is a much smaller, yet still (barely) statistically significant effect of presentation order ($\beta = 0.09$, $z = 2.0$, $p < 0.05$), and there is a slightly larger, but still small interaction between presentation order and training group ($\beta = 0.13$, $z = 2.2$, $p < 0.03$).

We also ran a Bayesian Regression (using the *brms* package in *R* (Bürkner, 2018)). The Bayesian approach to hypothesis testing differs quite significantly under the hood from frequentist approaches like logistic regression. Bayesian regression uses sampling to estimate each beta value and outputs a probability distribution over each parameter. Instead, in frequentist statistics, each parameter is assumed to be unknown but fixed (van de Schoot et al., 2014). Bayesian regression also allows the statistician to specify prior distributions for the parameters. For all intents and purposes (at least for this thesis), both types of regression yield similarly-interpretable results. We dabble in Bayesian Regression here so that we can include the maximal random effects structure without worrying about convergence issues. For frequentist regression, you look to the p-value to determine the significance of a parameter; if $p < 0.05$, then the parameter is generally statistically significant. For Bayesian regression, if the 95% confidence interval does not contain zero, the parameter is statistically significant. If it does, it isn't. The significant population-level effects were the same as, with slightly different values than, the results we got from the logistic regression (see Table 3.2).

Parameter	β estimate	l-95% CI	u-95% CI
centered_log_freq	**0.68**	0.45	0.93
sentence_type	**-2.63**	-3.41	-1.93
centered_log_freq:sentence_type	**0.41**	0.05	0.77
pres_order	**0.14**	0.02	0.27
pres_order:with_training	**0.24**	0.05	0.47

Table 3.2: Results of Bayesian Regression for Experiment 1. Statistically significant β estimates are in bold, in this case all of them.

Just for fun, we reran the same analyses above, but with just the wh-question acceptability ratings. We were concerned that the declaratives were already pretty maxed out on the scale (with a starting mean acceptability of around 0.9) and any statistically visible gains in acceptability would be be limited by boundary effects. Running a logistic regression with the predictors of centered_log_freq, pres_order, and pres_order:with_training with random intercepts for item and participant found statistically significant effects of centered_log_freq ($\beta = 0.71$, $z = 8.2$, $p < 0.001$) and pres_order ($\beta = 0.16$, $z = 2.7$, $p < 0.01$). The interaction of training group with presentation order was not found to be super significant ($p = .09$). A Bayesian regression more or less confirms these results.

3.2.4 Analysis / Discussion

The regression analyses above provide tentative support for potential result 1, that both groups have increased acceptability of the stimuli from initial to final testing, but the training group more so. The effect sizes for the main predictor and the interaction are quite small, but given that both of those have positive β values, it shows that

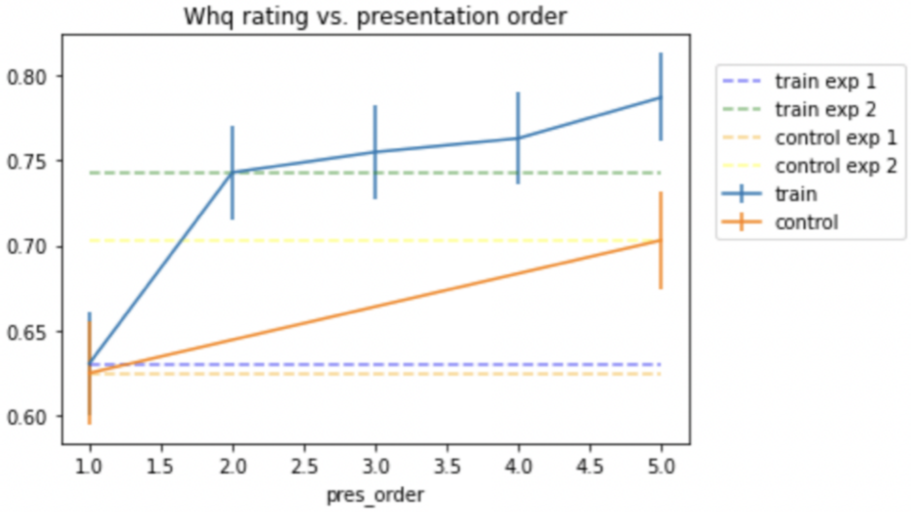


Figure 3-2: Results of Experiment 1b: Mean rating of wh-questions plotted against experimental stage, by group. For easier comparison, the orange and blue dashed lines shows the mean acceptabilities of the participants' first sighting of each matrix verb in wh-question format, and the green and yellow dashed lines show the mean acceptabilities of the second time. Note that the control group's second exposure is shown at pres_order=5.0.

training and exposure both increase acceptability.

Let us now draw our attention to Figure 3-2, which shows wh-question acceptability judgements. The blue (and green) data is from the train group that received five exposures of each matrix-verb-construction combination. The orange (and yellow) data is from the control that only received two exposures of each. We have already talked about the training group data individually in the Experiment 1 analysis in Section 3.1.5. What is fairly interesting here is that the control group also experienced a significant increase in acceptability from the first exposure to the second (even though the second occurred over 100 other acceptability judgements and 100 comprehension questions away. The standard error bars overlap for the second exposure of both the train and control groups, so at least on this sub-30 minute time scale, the increase of acceptability given one exposure is independent of time between that first exposure and the second. Second, you can see that there is a difference between the mean acceptability of the final exposure of the training group and the final exposure of the control group. Both groups saw the same number of sentences between the initial and the final exposures, so it appears at least like the extra three exposures that the train group receive increases the acceptability.

There are multiple weaknesses of this experiment. First, a post-hoc power analysis tells us that we were at 65% power, which is not optimal. Statistical power is the probability of finding a significant effect, given the size of our data and the size of the effect. It is post-hoc because we run the analysis given the effect that we found in the actual data. Word on the street is that you should shoot for power of > 80%. Second, the effect sizes are quite small, barely reaching the threshold of normative statistical significance ($p < 0.05$). In addition, these was no sequestration of hypothesization, data collection, and analysis. This was treated as an iterative, exploratory analysis. There was a bit of double-dipping with the datasets, which is scientifically dubious, so this leads us right into Experiment 2 —a principled replication!

Chapter 4

Experiment 2

Experiment 2 is a replication of the amalgamation of Experiments 1 and 1b using an expanded set of matrix verbs.

4.1 Design & Materials

This time, because we are generating all of the data in advance, and not sequentially deciding on a control group, we will make it a paired experiment, meaning that one train participant and one control participant will see the same exact initial test and final test stimuli. The goal is to measure the effect of those three extra exposures on the change in acceptability.

For this experiment, each participant will see experimental stimuli composed of a subset of 6 of the following 12 verbs:

- claim, learn, notice, deny, hint, guess, shout, maintain, yell, conceal, whine, mutter

4.2 Hypotheses and Predictions

See Experiment 1b Hypotheses and Predictions in Table 3.1 for a more thorough summary, but as a quick reminder, we think there are four possible outcomes, reiterated in Table 4.1. We are comparing initial vs. final testing rounds and the group with training vs. the control group.

1. Both groups increase in acceptability from initial to final, but the training group more so.
2. Both groups indistinguishably increase in acceptability from initial to final.
3. The training group exhibited an increase in acceptability, but the control group did not.
4. There is no difference between groups or between initial and final testing rounds (H_o).

Table 4.1: Possible outcomes of Experiment 2.

Priming (both lexical and structural) could result in outcomes 1, 2, or 3. If exposures additively increase the effect of priming, then that would result in outcome 1. If all that matters in priming is the first exposure and time between exposures does not matter, then that would result in outcome 2. If the exposures must be close together for there to be a priming effect, then that would result in outcome 3. Lack of any priming or short-term learning would result in the null hypothesis, outcome 4.

Experiment 1b resulted (weakly) in outcome 1, so we predict the same outcome here. Given the weak results, however, we might not be super surprised if one or both of the effects fails to appear.

4.3 Participants using Power Analysis of Experiment 1b

Since we now have an expected effect size from previous experiments, we can estimate the number of participants we need to rerun this experiment on. We do this via a power analysis, using the *simr* package in *R* (Green and MacLeod, 2016).

We want to determine how many data points we need to have sufficient power to detect our interaction of focus. The β values that we estimated in Experiment 1b were 0.09 for pres_order and 0.13 for the interaction of with training group. These are very small effects, and often the empirical estimation of a detected effect is larger than the true value of that effect, so the first step is to underestimate the effect size. To calculate a conservative estimate of sample size, we, in essence, pretend that the effect is smaller than it was to see how many samples we would need just in case the effect is that small. We run a power analysis, assuming the effect size is 0.08 (which is 2/3 of the empirical value for the interaction and just under the empirical value for the main effect of presentation order). The choice of exact number here is somewhat arbitrary; we could have chosen a similarly scaled multiple by which to reduce the empirical effect size.

There are two ways in this experiment to increase the number of samples. We can increase the number of mTurk workers, or we can increase the number of matrix verbs. We generated a power curve along both axes, as shown in Figure 4-1.

Although statistics is deeply rooted in theory and mathematics, in practice, statistics is more of an art than a science. A “good” and agreed-upon power number to shoot for is 80%; there is no mathematical grounding to this, but there is precedence in the literature (besides, in order to get NIH funding, you need to demonstrate that your study has 80

4.4 Results

We ran Experiment 2 on mTurk with 150 participants, as per the computations above, with the same meta-conditions as Experiments 1 and 1b. We had to exclude 2 participants who self-reported to not be native American-English speakers and 19

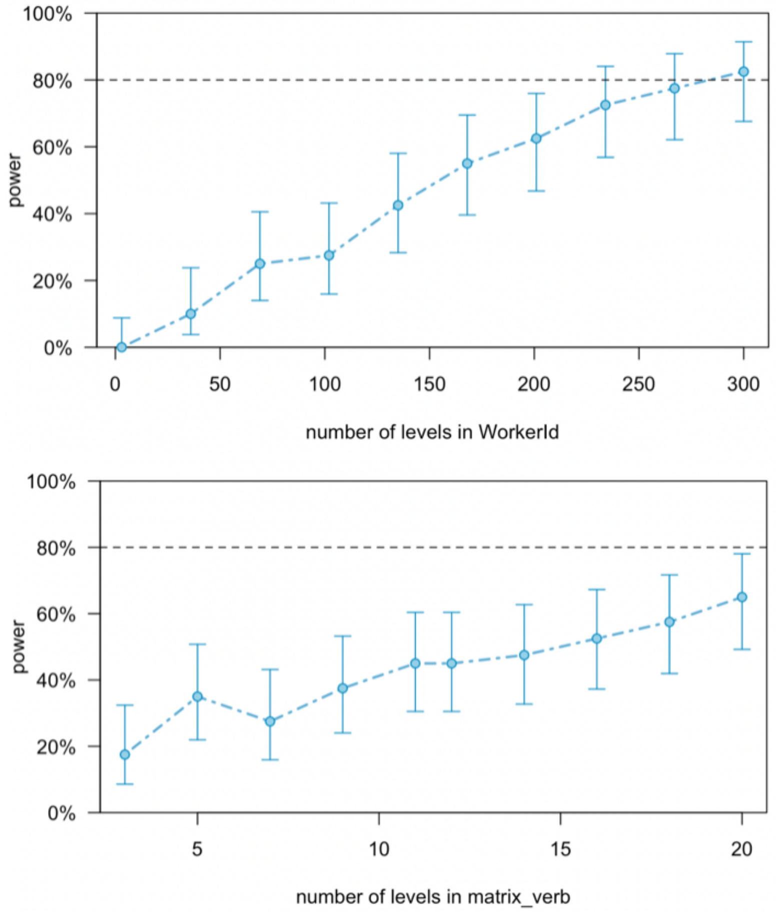


Figure 4-1: Simulated power curves to detect a small effect ($\beta = 0.08$).

participants who answered the comprehension questions incorrectly more than 15% of the time. So, the analysis below will be on the survey results from 129 participants. This is a 14% exclusion rate, which is apparently rather low.

First, we verify that the train and control groups were exposed to items of comparable acceptability during the "training" phase (rounds 2-4). Like with Experiment 1b, a two-sided t-test was used to compare the control group's particular fillers against the training group's training items. This time, there was potentially a significant difference in mean acceptability between the control group (mean=0.80; SE=0.017) and the training group (mean=0.74; SE=0.026); t(127)=1.698, p=0.09. The p-value does not pass the 0.05 threshold, but it is much closer than it was in the previous experiment (0.29).

Now onto regressions. We used mixed-effects logistic regression and Bayesian regression to predict acceptability ratings. We only considered the acceptability judgements from the initial and final testing stages for both the training and the control groups. We used the same formula as described in Experiment 1b (using a maximal Random Effects structure for the Bayesian analysis):

$$\text{rating} \sim \text{centered\_log\_freq*sentence\_type+pres\_order+pres\_order:with\_training} \\ \text{+(1|WorkerId)+(1|item)}$$

We again confirm the VFF, finding a significant main effect of the log-transformed frequency of the matrix-verb frame ($\beta = 0.72$, $z = 8.89$, $p < 0.001$) and of sentence type ($\beta = -2.65$, $z = -24.7$, $p < 0.001$). No interaction between those two was found ($p > 0.5$).

Last time, we found small effects of presentation order and the interaction between presentation order and training group. This time, we tragically don't. There is an almost significant main effect of presentation order ($\beta = 0.07$, $z = 1.8$, $p = 0.07$) and no significant interaction ($p > 0.15$).

Running a Bayesian regression confirms these results, see Table 4.2.

Parameter	β estimate	l-95% CI	u-95% CI
centered_log_freq	**0.81**	0.58	1.05
sentence_type	**-3.13**	-3.61	-2.68
centered_log_freq:sentence_type	0.13	-0.17	0.48
pres_order	0.11	-0.01	0.22
pres_order:with_training	0.09	-0.06	0.25

Table 4.2: Results of Bayesian Regression for Experiment 2. Statistically significant β estimates are in bold, in this case just centered_log_freq and sentence_type.

We had concerns about the integrity of the declarative ratings, given the discrepancy in their starting points (see Figure 4-2), so we also ran regressions on just the wh-question data. In any case, wh-questions is where the bulk of the "learning" would take place, given their lower starting acceptability. The logistic regression estimates a significant main effect of verb-frame log frequency ($\beta = 0.81$, $z = 8.5$, $p < 0.001$) and a significant main effect of presentation order ($\beta = 0.11$, $z = 2.4$, $p = 0.016$), but no significant interaction of presentation order and training group ($p > 0.30$). A Bayesian regression concurs qualitatively in terms of significance.

Another way that we considered analyzing the data that would mitigate the effect of the strange initial declarative rating group difference would be to analyze the two groups separately. We did that. We ran a logistic regressions on both groups' data, with the following predictors: centered_log_freq, sentence_type, their interaction, and pres_order. For both groups, there was a statistically significant effect of pres_order (train: $\beta = 0.11$, $z = 2.9$, $p < 0.005$; control: $\beta = 0.08$, $z = 2.1$, $p < 0.05$). The issue with running this analysis separately is that there is less insight to be gained from comparing the two groups. In addition, the power is lower, because there are fewer sentence types. For instance, the estimate of the effect of presentation order seems larger for the train group than the control group, but there is no reasonable quantitative comparison to be made between the two. What it does allow is to further cement the reasonableness of the presentation order effect as a within-group effect, while allowing us to ignore any possible confounding and strange among-group effects.

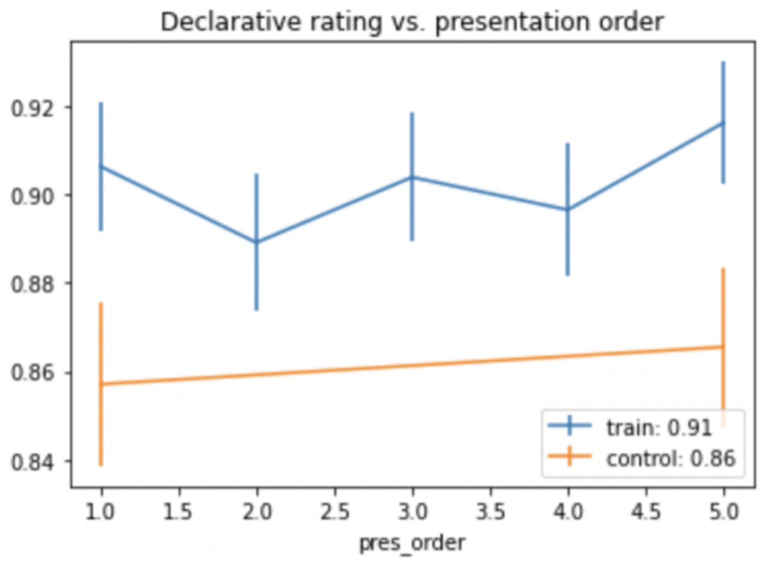


Figure 4-2: Mean rating of declarative sentences through the course of the experiment by group. The bars show standard error. The upper blue curve shows the train group's ratings from the initial testing round, through training, and then from the final testing round. The lower orange curve shows the control group's from the initial and final testing rounds. The mean ratings during the initial round are shown in the legend.

A quick note: doing these types of ad-hoc analyses as a ploy to save ourselves from bad data is, for lack of a better word, sketchy. Another word for it could be p-hacking. We do think, however, that without a proper explanation for the initial group differences in declarative rating, we don't have a better option. The best thing to do is another replication or an extension, which we will get to in Experiment 3.

4.5 Analysis / Discussion

There were a couple suspicious bugs/features of the data collected for this experiment, two of which we would like to address before diving into the meat of the analysis: a low participant rejection rate and a bizarre initial difference in rating between the two groups.

This experiment had a rejection rate of 14%, which is anecdotally quite low. It is not

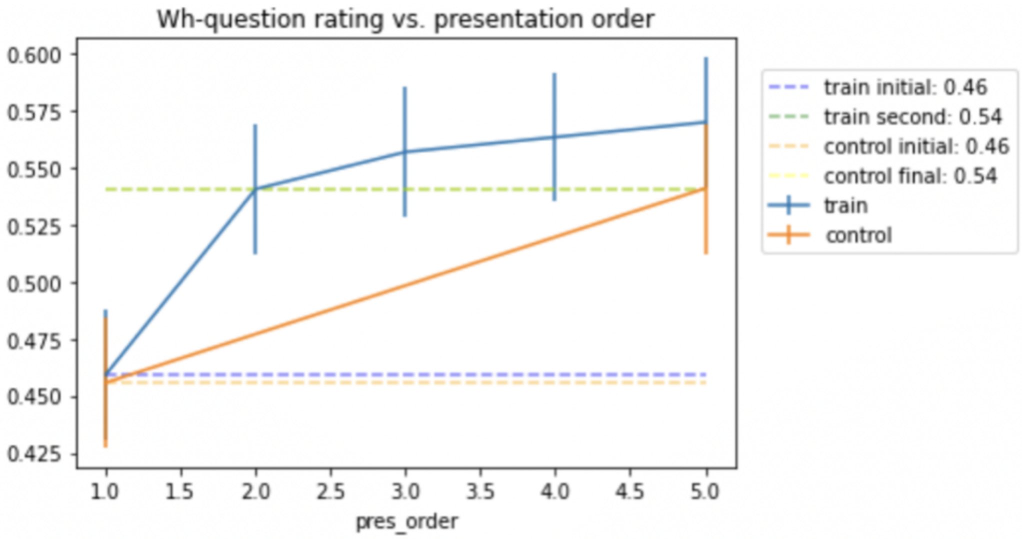


Figure 4-3: Mean rating of wh-questions through the course of the experiment by group. The bars show standard error. The upper blue curve shows the train group's ratings from the initial testing round, through training, and then from the final testing round. The lower orange curve shows the control group's from the initial and final testing rounds.

uncommon for the rejection rate to be 20-30% of participants, due to the nature of the mTurk market. Many users attempt to game the system and/or use bots to fill out surveys; the concrete incentive is to finish the survey to get the money, not to provide the researchers with quality data. Obviously, this does not hold for all participants, but it underlines the importance of having robust methods to determine the quality of a participant's survey answers. We will do two additional things in Experiment 3 —add completion questions and make the comprehension questions tougher —in order to try to improve the quality of our data.

An additional component of these data's funkiness was the consistent differences in the ratings by the train and control groups. In the early stages of analysis, we noticed that the train group's mean ratings were significantly (0.05) and consistently above the control group's. It might make sense for the train group to end up with a higher rating than the control group (that was one of the hypotheses, after all), but the data show that the train group even started with higher ratings. This makes no sense. There should be no significant differences in the groups' ratings during the

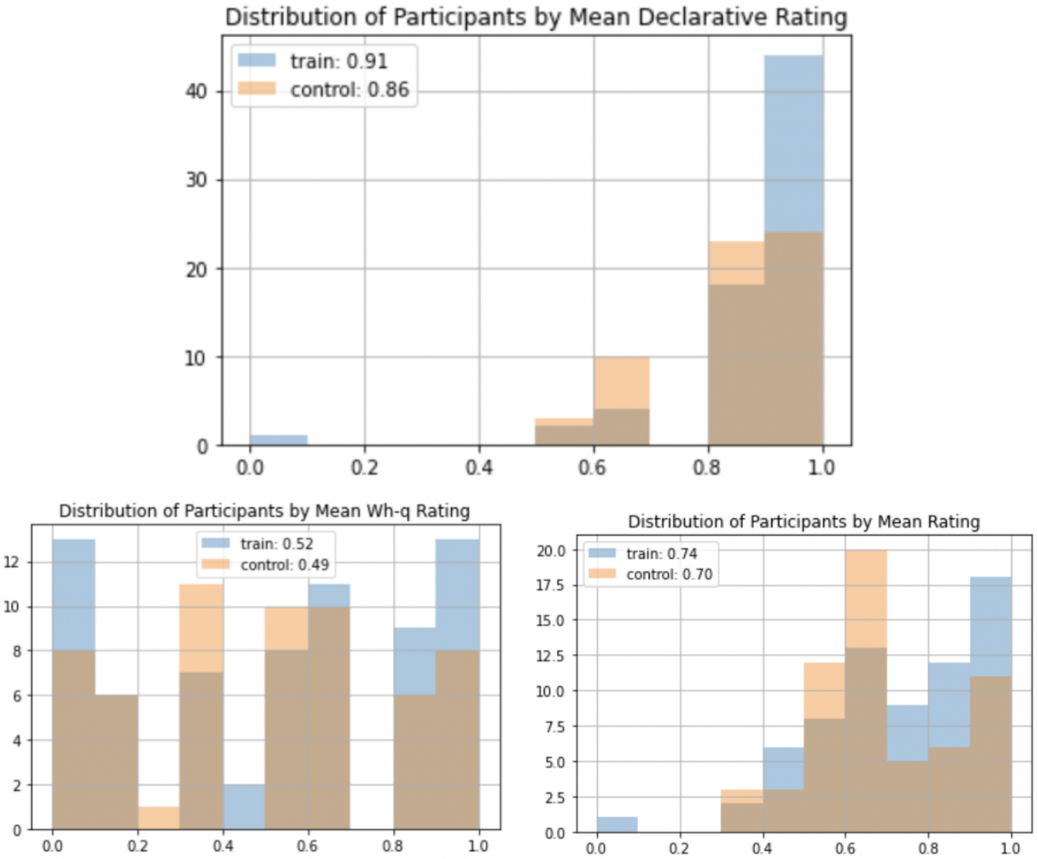


Figure 4-4: Participant distributions by mean rating of all declarative sentences, wh-questions, and all sentences. Population means are shown in the legends.

initial testing stage. They see exactly the same material and no group has undergone training yet. The only difference is the randomly assigned group. Though not for lack of trying, we have not found a way to attribute this to anything beyond quite an inopportune feat of randomness.

First, we assumed that there was a bug somewhere in our analysis pipeline. We spent more than a few hours trying to figure out what the problem was, and we are pretty confident it isn't the code, even though that would be the most likely cause for the discrepancy. We tried multiple ways of skinning the cat, but it always resulted in the same difference. Through this careful consideration, however, we were able to locate the source of the discrepancy: the declarative sentences. There is no real difference in the mean ratings of the filler items (train=0.947; control=0.941), there is no real difference in the initial mean ratings of the wh-questions (train=0.459; control=0.456), but there is a significant difference in the initial mean ratings of the declarative sentences (train=0.906; control=0.849).

Then, we tried to visualize the distributions of the participants' mean ratings for the various sentence types, thinking that there could be a few outliers that were randomly placed all on one group. This visualization is shown in Figure 4-4. Looking at just the declaratives, there are about twice as many train-group members that have a mean rating of >0.95. The number of participants in the train group that gave all natural ratings (1.0 average) for declarative items was 24, while in the train group it was only 14. There is no reasonable explanation for this, because both groups were shown the exact same materials in the initial phase. The wh-question distributions seem qualitatively more similar to each other (and so are the population means) than the declarative ones, which means that at least we have managed to localize the discrepancy visually.

To try to "fix" this, we tried some additional data scrubbing. We removed participants who rated all sentences as natural (1.0), those who rated all fillers and declaratives as natural and all wh-questions as unnatural (0.0), and those who rated all declaratives as natural, all wh-questions as unnatural,, and fillers at chance (0.5; fillers were half questions, so their strategy might have been to mark all question marks as 0 and all periods as 1). This removed an additional 26 participants, which is kind of a lot. It makes us wonder how many of the participants are making good-faith linguistic acceptability judgements and how many are just completing a task, which is what they are incentivized to do. In any case, mTurk is both a blessing and a curse. Unfortunately this extra participant removal did not result in the initial declarative rating means converging. They were still the same distance apart. This means that these "outliers" who might be gaming the system were not the cause of the strangeness. This makes sense, because we expect peopole with these strategies to be evenly distributed in the groups. We leave finding the actual reason as an exercise for the reader.

Figure 4-2 shows the average ratings of declarative sentences across the course of the experiment for both the train and control groups. Yes, it is rather suspicious that the control group starts and ends below the train group by a consistent amount (∼0.05); we will talk about this discrepancy in detail in the subsequent analysis. Declarative ratings, in any case, are less interesting due to their proximity to the upper boundary of the rating system. We are interested in changes in acceptability, and any changes are likely to be very subtle and not super significant due to the already very acceptable nature of these declarative constructions.

Wh-questions, on the other hand, are the object of interest, given their feeble but not atrocious starting acceptability. If there is any priming to be found, it would be found on these blurrily acceptable constructions. We have three main conclusions to draw from Figure 4-3: priming occurs, priming is strongest with the first exposure, and priming is relatively time-independent.

First, we can tell that priming occurs by the monotonically increasing curves for both training groups. Let us provide a caveat that we can not yet causally attribute the increase in acceptability to priming, because there is still the plausible non-linguistic explanation that participants get better at the task the more sentences they rate, or they simply get nicer the longer they spend on the task. We will address this possibility with Experiment 3.

Second, to compare the priming effect of each exposure, we have to look at the train group's curve in blue, as they are the only ones in which the effect of multiple exposures can be seen. Although the acceptability increases with each exposure, the magnitude of the increase decreases significantly after the first jump. The increase from the first to the second was 0.081, while the increase from the second to the fifth was much less, 0.029. In fact, we used a two-sided t-test to compare the ratings during the second and fifth exposures, and there was no significant difference in the train group's mean acceptable rating during their second exposure (mean=0.54; SE=0.028) versus during their fifth (mean=0.57; SE=0.028); t(612)=0.73, p=0.47.

Third, in addition to the first exposure being the most powerful priming force, there is no time-dependent difference in the priming effect of that first exposure. We also saw this in Experiment 1. The delta acceptability from the first to the second exposure for the train group was 0.081 and for the control group it was 0.085. Note that the control group's final acceptability rating is also their second. The change in acceptability is virtually the same, even though the control group had upwards of 100 sentence rating and comprehension tasks in between ratings! On a short-term time scale, at least, the effect of the first exposure does not diminish with elapsed time.

Although Figure 4-3 provides fruitful ground for tantalizing analysis, the parameter estimations from the regressions are much less convincing. We found a very small effect of presentation order and no significant impact of training group on the effect of presentation order. This is not what we found in Experiment 1 & 1b. It is weakly in line with possible option number 2: both groups indistinguishably increase in acceptability from initial to final. It is also weakly in line with the fourth option, a.k.a. the null hypothesis, that there is no short-term learning occurring. In spite of the strange declarative data, the clear visual relationship of exposure and wh-question acceptability and the extra regressions that we ran provide evidence that acceptability is increasing with exposure. However, as the experiment stands, there is no way to disentangle exposures with the natural time course of the experiment. Hence we march on to Experiment 3.

Chapter 5

Experiment 3

The purpose of Experiment 3 is to more concretely determine whether priming affects acceptability, or whether participants rate more generously as the experiment goes on. We do this by simplifying the experiment and by adding a different type of control group.

5.1 Design & Materials

The experimental sentences are identical to those in Experiment 2. The changes regard the structure of the Experiment, the design of the groups, and some meta-experiment changes to address concerns with mTurk.

Experiment 3 is a simplified, shorter version of Experiment 2. From Experiment 2, we know that the biggest change in acceptability occurred between the first and second exposures and that time between those two exposures didn't have much of an impact, if at all. We will just do the initial and final testing rounds, forgoing the "training" phase.

In addition, we will have two control groups, one who does not receive experimental items at all in the initial round so that we can determine whether acceptability increases regardless of exposure. The second control group will see declarative sentences with the matrix verbs in question, but not the island, wh-question constructions. This will allow us to see if it is exposure to the matrix-verb in the specific construction that contributes to priming, or if it is just familiarity with the lexical items.

To reiterate more concretely, each participant will now provide acceptability judgements and comprehension question answers for 72 sentences. There are three groups: train, lexical control, and control. All participants will see a subset of 6 out of the 17 possible matrix verbs (the set union of those in Experiments 1 and 2).

- Initial testing round:
  - **Train** sees 6 wh-questions, 6 declaratives, and 24 fillers
  - **Lexical control** sees 12 declaratives and 24 fillers
  - **Control** sees 36 fillers
- Final testing round:
  - **All groups** see the same thing (6 wh-questions, 6 declaratives, and 24 fillers)

In addition to the experiment redesign, and as discussed previously, we needed to find a way to reduce the number of participants that were task-hacking rather than providing good-faith linguistic judgments. We did this in a couple of ways. First, we made the comprehension questions more challenging so that you could no longer simply pattern match to answer them. For example, if the experimental item was "Josiah maintained that Olivia knew something." Before, the comprehension question would ask: Did the target sentence mention Olivia? (A: Yes). Now, we ask: "Did the target sentence mention Josiah knowing anything?" (A: No). These questions are still obvious, but now they require more careful reading of the sentence and cannot be answered quickly by string-matching. This will enable us to better rule out participants gaming the system. Second, we added five simple completion tasks to the beginning of the survey. These are of the form: "It is raining so…" or "I hope…" and the task is to complete the sentence. These are a good litmus test for fluency, because they are fairly straightforward for any native English speaker to do, but are fairly difficult (and it is pretty evident) if you don't know the language. This eliminates people who might otherwise be able to pass the other requirements, but who are not from our target population of native American-English speakers.

5.2 Hypotheses and Predictions

This experiment's purpose is to really distill down Experiment 2 and determine what we should attribute the increase in acceptability to priming or just a general experimental increase. We will consider the ratings in the final testing round for our analysis. There are three main possibilities:

1. There is no difference in final wh-question acceptability ratings among all three groups (H_o).

2. The train group's ratings are higher than the other two groups.

3. The train and lexical control groups' ratings are higher than the pure control group.

If possibility 1 is true, this does not provide any evidence for priming affecting acceptability. Under possibility 2, increases in acceptability can be explained by structural priming, showing that exposure to the verb in the construction increases acceptability on a short-term time scale. Under possibility 3, increases in acceptability can be explained by lexical priming, meaning that exposure to that specific verb increases acceptability.

This experiment attempts to answer the following question: does priming (lexical and/or structural) modulate linguistic acceptability?

5.3 Participants using Power Analysis of Experiment 2

We conducted a power analysis using the empirical results of Experiment 2 to determine how many samples we need to have enough statistical power to detect our desired presentation-order effect. This time, though, given the different participant-grouping structure, this computation might lean more art than science. We start by using the computed presentation order effect size from Experiment 2 ($\beta = 0.065$), and we divide by 1.5 to ensure a conservative estimate. We then compute the power curve (Figure 5-1). The tricky part about this is that in Experiment 2, the presentation order effect affected all groups; it was a population-level effect. In Experiment 3, only one of the three groups is under the same experimental learning conditions as the participants in Experiment 2; so, for all we know, it might be the case that only one of the three groups has a presentation order effect. This means we need to have enough samples in that *one group alone* to detect the effect. Looking at the curve, a good number might be a little less than 150, and we settled on 120 participants in each group, so 360 participants total.

Because this priming effect should be matrix-verb-independent, we will use a pool of 17 matrix verbs (the set union of the verbs used in the first two experiments). This number did not significantly affect the power calculations.

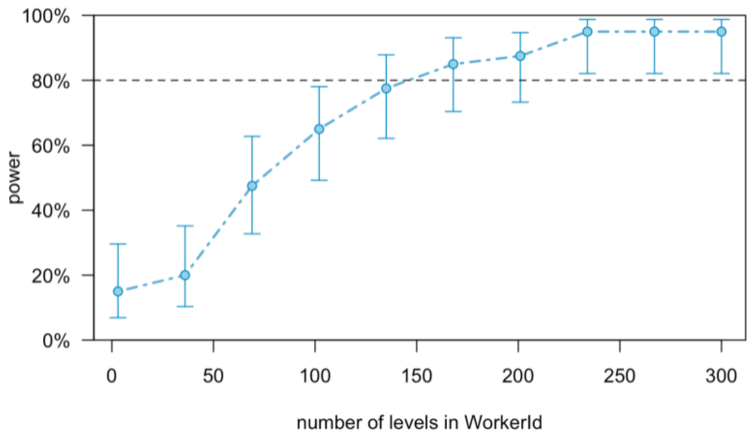


Figure 5-1: Simulated power curves to detect a small effect ($\beta = 0.065$).

5.4 Results

We ran Experiment 3 on mTurk with 360 participants with the same meta-conditions as Experiments 1 and 2, but with 3USD compensation due to the shorter duration of the experiment (less than half of the judgements, but with more challenging completion questions and an additional completion task).

We had to exclude 89 total trials form analysis —2 non-native speakers, 2 duplicate trials, and 85 trials had a comprehension question accuracy below our threshold of 85% (64 trials were answered at chance). In addition, we excluded an additional 9 trials which had particularly sketchy completion-task responses (e.g., InMySchool… "I am the topper of the student." or IHope… "god very well."). We only excluded those that we considered to be particularly egregious. So, the analysis below will be on the survey results from 262 participants. This is about a 27% exclusion rate, which is more in line with what usually happens.

One important thing that we are controlling for in this experiment is the mean acceptability in the first half of the experiment. We do not know how participants' judgements are affected by the distribution of acceptabilities. Although it is interesting and bears further study, for this experiment, we want reduce the possibility of that being a confounding factor. Unlike with Experiments 1 and 2, there are now three groups to compare. We use three pair-wise two-sample t-tests to compare each combination of groups. There was no significant difference in the mean initial round acceptability between the train (mean=0.89; SE=0.010) and the control (mean=0.91; SE=0.010) groups; t(167)=-1.40, p=0.16, nor between the train and the lexical control (mean=0.90; SE=0.0094) groups; t(170)=-0.41, p>0.50, nor between the control and the lexical control groups; t(181)=1.05, p=0.29.

Next, as with the previous two experiments, we used both mixed-effects logistic regression and Bayesian regression to estimate the effect of training group on the final acceptability ratings. This time, the statistical analysis is simpler, as there is only one parameter of interest to estimate, and the parameter is probably statistically equivalent to some sort of t-test. The variable training_group is a factor with three levels, one for each group. We use dummy coding for the training_group variable, because it compares each level to a reference level (in this case the control group). We used the following formula for the logistic regression (and the same formula, but with a maximal Random Effects structure for the Bayesian analysis).

$$\text{rating} \sim \text{centered\_log\_freq*sentence\_type+training\_group}$$
$$+(1|\text{WorkerId})+(1|\text{item})$$

We, in a truly shocking turn of events, once again confirm the VFF, finding a significant main effect of the log-transformed frequency of the matrix-verb frame ($\beta = 0.58$, $z = 4.80$, $p < 0.001$) and of sentence type ($\beta = -2.81$, $z = -10.4$, $p < 0.001$). No interaction between those two was found ($p > 0.15$).

We find no effect of training group ($p > 0.50$), as would be expected given the results of the t-tests. Last time, we found small effects of presentation order and the interaction between presentation order and training group.

A Bayesian regression with a maximal random effects structure also fails to find a significant difference in acceptability judgments between the lexical control and the control group (95% CI=[-0.55, 1.14]) or between the train and the control group (95% CI=[-0.78, 0.91]). Remember that significance is when the 95% confidence interval

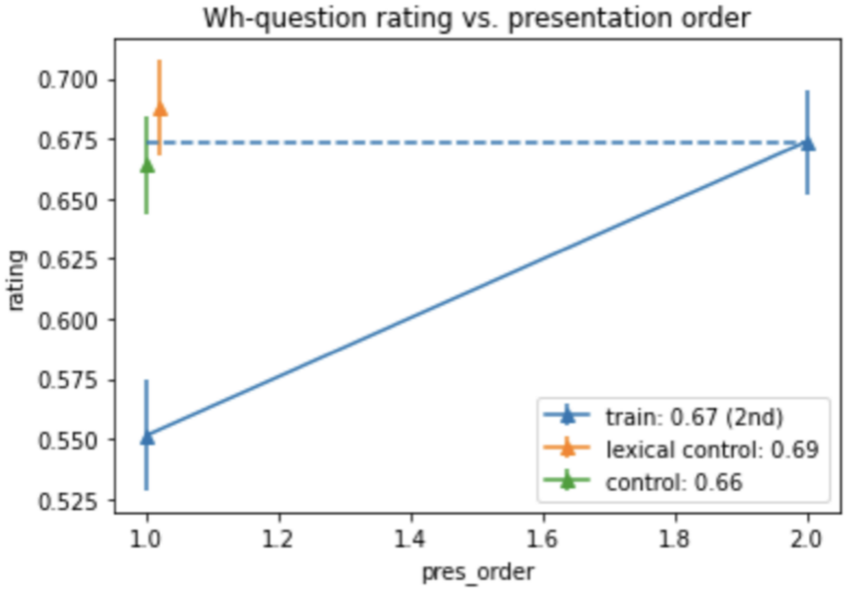


Figure 5-2: Plot showing rating of wh-questions by presentation order and training group. The legend shows the means during the final round of testing. The error bars show standard error. For both of the control groups, the final round is the only round, so it is also the first.

does not contain zero. The Bayesian estimate for the log frequency of the matrix-verb frame (0.65; 95% CI=[0.35, 0.98]), the sentence type (-3.29; 95% CI=[-4.05, -2.60]), and their interaction (95% CI=[-0.17, 0.90]) confirm the logistic regression. One might ask why bother to do both, given that they essentially do the same thing. Logistic regression is how things are usually done, but Bayesian regression is more powerful, allowing for a maximal random effects structure and for specification of priors. Bayesian regressions are not more common, we suspect, due to Newton's first law.

Seeing that all groups had statistically indecipherable average ratings during the second and final stage, we tried a logistic regression using a different predictor than presentation order, or the number of exposures. The "training" is ostensibly useless, and it appears that there is a general increasing trend by the raw order of presentation. We introduce a new predictor variable, raw_order, which is an integer from 36 to 72, coded as the i-th sentence that a participant rates (that we then scaled and centered around 0). The *R* formula that we used for the regressions is as follows:

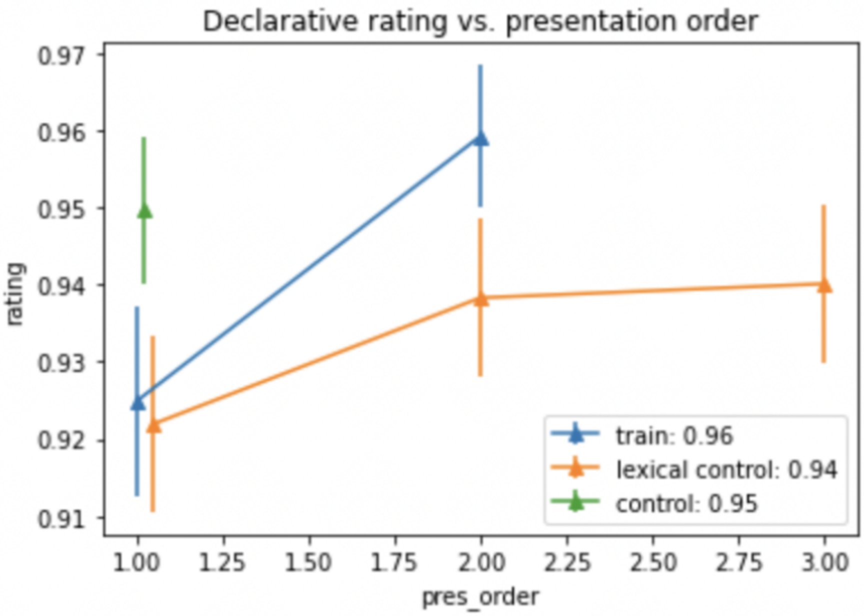


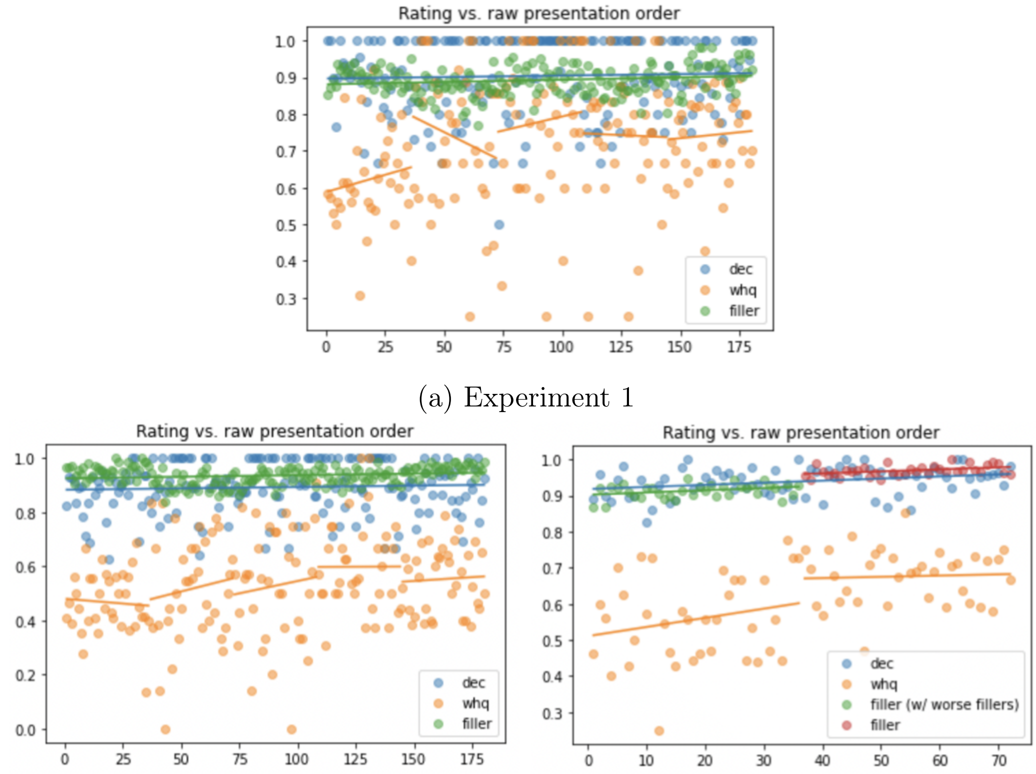
Figure 5-3: Plot showing rating of declaratives by presentation order and training group. The legend shows the means during the final round of testing (so control's first exposure, train's second exposure, and lexical control's third). The error bars show standard error.

$$\text{rating} \sim \text{centered\_log\_freq*sentence\_type+training\_group+}\textbf{raw\_order}$$
$$+(1|\text{WorkerId})+(1|\text{item})$$

Fitting this regression to the final declarative and wh-question ratings did not result in any significant effect of raw_order ($p > 0.50$).

We also tried fitting a slightly more complicated formula on all of the data that will account for the differences in acceptably increase for the different types of sentences. Because this raw-order effect should affect all sentences rated, not just the experimental items, the variable sentence_type is now a three-level factor: declaratives, wh-questions, and fillers. As mentioned before, the declaratives and fillers start off closer to the upper limit of acceptability judgements, so an increase in acceptably may not be as visible:

$$\text{rating} \sim$$
$$\text{centered\_log\_freq*sentence\_type+training\_group+}\textbf{sentence\_type:raw\_order}$$
$$+(1|\text{WorkerId})+(1|\text{item})$$



(a) Experiment 1

(b) Experiment 2

(c) Experiment 3

Figure 5-4: These three plots show average rating based on position in the survey (i.e., raw presentation order). The wh-question ratings are bucketed by the round (pres_order) in which they were seen. Each participant saw 180 total sentences in Experiments 1 and 2, so there are 180 possible positions. For Experiment 2, there are 72. Remember that there are much less participants in the first graph, so there could be noise.

This regression found a significant raw_order effect ($\beta = 0.22$, $z = 2.8$, $p < 0.005$), and significant interactions between filler sentences and raw_order ($\beta = 0.31$, $z = 3.4$, $p < 0.001$) and wh-questions and raw_order ($\beta = 0.24$, $z = 2.3$, $p < 0.05$). Due to the dummy coding, the main effect is the signal for the interaction between declaratives and raw_order.

5.5 Analysis / Discussion

The evidence does not allow us to reject the null hypothesis that priming does not affect acceptability. The regressions fit to the final round of acceptability judgement data do not show a significant effect of training group; also, three pair-wise two-sided t-tests confirm that there is no statistically significant difference among the groups' wh-question ratings at the end of the experiment, even though the final testing round is the control group's first exposure to that verb and verb-construction combination, while the lexical control group has seen that verb before, and the train group has seen that specific verb-construction combination (and the verb) before.

In addition to trusting the math and statistics, looking at Figure 5-2 lends further evidence against priming, lexical or structural, causally affecting acceptability. The mean ratings for each group's first exposure to the wh-question island are not the same. Both control groups have a mean rating much higher than the train group's. The control groups' mean rating, however, is statistically identical to the train group's mean rating during the second/final exposure. The control groups' first rating takes place during the final testing round. In other words, the presentation order or number exposures of each specific matrix-verb-wh-question construction are not predictive of acceptability rating. What is predictive is the experimental round, initial or final. All groups, regardless of their training regimen in the initial round of 36 sentences, rated the object-extracted wh-questions similarly in the final round of 36 sentences. It is clear, also, from the trend of the train group (and from Experiments 1 and 2), that there is an increase in acceptability from the first to the second exposure of the matrix-verb-construction combination.

There is a comparable trend with the declarative sentence acceptability ratings in Figure 5-3. The means do not align based purely on the number of exposures already seen of that matrix-verb-construction combination. Instead, the means are better explained by the current phase of the experiment. The right-most green, orange, and blue marks are from the final testing stage, and they are about the same in rating, even though they are the first, third, and second exposures, respectively.

We tried a regression on the Experiment 3 data, including raw order as a predictor, and it was a significant main effect, and it had significant interactions with training group. It is not clear to us that this is a clean statistical test for the power of that effect, because the experiment was not designed to investigate it.

Looking at Figure 5-4, it does seem like there is a general positive trend, but not within-bucket trend, for each construction over the course of the experiments. Meaning, there is a general increase over the course of the experiment, but not pres_order=i for each i. This general trend, of course, is heavily confounded by the design of the experiments. For the items that we were manipulating exposures of, wh-questions and declaratives, presentation order is essentially the same thing as this raw order, just bucketed. The two factors carry redundant information, because when presentation order is 1, the raw order is going to be 1-36, when it is 2, the raw order is 37-72, and so on. If there is truly a raw order effect, you would expect to see a positive slope within each bucket. So, among all constructions that are the i-th exposure, there should be a trend that those shown at the very beginning bucket are less acceptable than those shown at the end of the bucket. The graphs are less than conclusive towards evaluating this. There does appear to be a general trend of positive slope, but the effect is not super strong. Fillers were not part of the original experimental design, so their trend line is the only one untainted by potential priming confound (or at least controlled for that via randomization). For Experiment 3, however, the fillers for the control groups in the first half are of a different distribution than those of the second half, as per the experimental design. The issue with just looking at fillers is the fact that their acceptability oscillates near 1.0, susceptible to the effects of our chosen scale's ominous boundary. In other words, if somethings starts out very acceptable, any improvement is going to be nearly statistically invisible, but that doesn't mean that it's not there. What is clear is that this potential raw-order effect merits further study; more on this in the discussion on Future Work.

Chapter 6

Meta-Discussion / Future Work

These three experiments replicate the VFF a few times over for this specific set of matrix verbs and construction types. This hypothesis is not super strong, only detailing a correlation. As such, there is a lot of room for future work here. In addition to testing this effect on other types of sentences, there are other questions left to be answered. One avenue for future study concerns questions that might lead to a partial theory of what makes a sentence "good": Do other sentential components, like the subject or object or embedded verbs, also impact acceptability judgements? (A preliminary and rather cursory statistical analysis of the data collected over the course of the three experiments suggests that the impact, if it exists, is not very strong.) Is the effect of the individual sentential components on the acceptability of the whole additive, or is the composition a more complex process? Why does the matrix verb seem to have a privileged impact on overall sentence naturalness? Is acceptability a purely linguistic phenomenon, or does it involve extra-linguistic processes? Another line of study, more along the lines of language acquisition, might address how knowledge of language is built from and modulated by linguistic exposure. Language considered outside of the human context becomes devoid of meaning, lexical goop floating in the void.

Our hypothesis that would take this frequency-based account a few steps further in the direction of short-term exposure does not appear to be empirically grounded. Although it certainly appears in Experiments 1 and 2 that the rating of a sentence increases significantly the second time a participant sees a specific matrix verb in a specific construction, Experiment 3 finds that the increase occurs just along the course of the experiment; priming doesn't affect it.

There does, however, seem to be a weak generally increasing trend in acceptability throughout the course of all of the experiments. Experiment 3 did some grunt work of trying to disentangle the raw presentation order with the possible effects of structural and lexical priming. Ruling out priming for the time being, there are multiple possible explanations, that could explain a general increase in acceptability over the course of the experiment. First, perhaps participants just get nicer as the experiment goes on. Regardless of the experimental items, they rate them more positively the more sentences they have rated already. Another possibility is that the participants adapt to the stimuli. In completing the task, the participants adjust their rating to fit the general distribution of the acceptability of the sentences. If the stimuli are very natural (like the fillers and declaratives), then this will look like a general increase in acceptability across the course of the experiment. If the stimuli are unnatural, then you might expect the average rating to decrease. This would be an extra-linguistic explanation, consistent with a Bayesian learning framework, where participants start the experiment with a prior belief about how natural the sentences will be. As the experiment goes on, they modify their expectation to suit the actual distribution of the data. This might be more likely to occur if the participants view this as a task to complete quickly for pay, rather than sincere linguistic acceptability judgements. When using mTurk, the following are necessary but not sufficient to ensure the integrity of the results: a rigorous exclusion protocol that consists of non-trivial comprehension questions and a completion task at the beginning, fair pay for workers, interactive surveys, and clear instructions (Crump et al., 2013; Thomas and Clifford, 2017).

In order to disentangle these two possibilities, we would suggest an experiment comparable to Experiment 3, but manipulating the acceptabilities more egregiously. You could present lists varying at intervals between 0% and 100% natural to analyze this raw presentation order trend. If a very unnatural list causes participants to have lower ratings at the end, this would signal that the participants just adjust their expectations to the task. If certain distributions of acceptability cause an increasing trend, and some cause a decreasing trend, you can hone in on what the average prior probability is for mTurk workers taking linguistic acceptability judgement surveys.

Experiment 3 is very similar to the priming experiments in Luka and Barsalou (2005). They found that reading sentences with shared syntactic structure, but not shared content words, earlier in an experiment lead to increased grammaticality ratings (i.e., in Experiment 3, the train group's final ratings would be higher than both of the control groups'). Our data does not support this conclusion. Looking more closely at the differences between our Experiment 3 and their experiment, there seem to be three main differences. First, they use a variety of constructions, while we just look at object-extracted wh-questions and declaratives. Second, their experiments tested around 50 students who provided judgements on sentences printed on separate pieces of paper in person, while our experiments tested over 150 mTurk workers on a survey over the internet. Third, they solicited grammaticality judgements on a scale from 1-7, while we solicited binary "naturalness" judgements. Especially when we are dealing with small, population-level changes in acceptability, perhaps a more fine-grain scale might help to detect them. Those methodological differences seem like they shouldn't affect the empirical results, but they seem to. Further investigation into various methods and their replicability would be very nice.

More generally, further investigation is warranted into the various quantitative behavioral measurements of language and the production and comprehension of it. What information do reading times, eye tracking, acceptability judgements, grammaticality judgments, elicited productions, fMRI analyses give us about language processing? Crucially, what is the difference between the platonic scientific ideal and the empirical reality of these measurements? Do we need to divorce the notion of acceptability and acceptability judgments in an mTurk task? We can't, however, get lost in the sauce. Language is qualia; language describes qualia; language communicates qualia. Although necessary for principled scientific study, quantified language is one level abstracted away from the true phenomenon of language. Mind the gap.

Chapter 7

Conclusion

Grammaticality, as defined by the generative tradition, evades empirical scrutiny. Acceptability, on the other hand, is a phenomenon of performance and lends itself to scientific study. Acceptability is modulated by a language user's linguistic experience. A human's knowledge of language is necessarily shaped by their linguistic input, the words and sentences that they have been exposed to over their lifetime. This knowledge of language is not static, but rather continuously adapting. Language learning is a life-long process that doesn't suddenly halt when a learner reaches some arbitrary measure of fluency. A language learner is continuously integrating new input and new experience into their prior conception of language. A human's ability to make acceptability judgements, to introspect on their knowledge of language, must therefore be affected by their prior exposure to the language, on both a long-term and short-term timescale.

First, concerning the effects of long-term exposure on acceptability, the VFF does seem to be an experimentally robust phenomenon. Acceptability increases with the google-books-corpus frequency of the matrix-verb frame, and acceptability independently increases with the frequency of the sentence type. This hypothesis does not purport to explain why, or posit any causal mechanisms.

Second, the effects of short-term exposure on acceptability judgements are less clear. Experiments 1 and 2 showed an increase in acceptability the second time a particular matrix verb in a particular construction was shown. Experiment 3, however, does not reject the possibility that there is just a general task-based improvement in acceptability. This might skew a tad more to the psychology side of psycholinguistics, but it nonetheless cannot be ignored.

The failure of Experiment 3 to replicate the effects of exposure on acceptability found in Luka and Barsalou (2005) either lends credence to the replicability crisis in psycholinguistics or/and suggests that there are limitations to the method of mTurk studies. Looking into how people judge acceptability and what affects their judgements beyond just linguistic content merits further explanation. Although mTurk is convenient, consideration of the incentives of the participants and robustness in detection of the participants who are taking the surveys with different goals in mind than the researchers is exceedingly important to the scientific integrity of the experiments.

For state-of-the-art natural language processing (NLP) models, exposure ipso facto affects their output, whether it be in the form or "judgements" or "elicitations"; the only input the models receive is text, lots of text. Better understanding how exposure modulates acceptability in humans will allow us to better understand the judgments of NLP models whose essence is predicated on exposure and who have inspired a cottage industry of psycholinguistic experiments, but remain poorly understood. In all likelihood, however, exposure is not the only thing that impacts human acceptability judgements. In fact, my experiments provide concrete evidence for exposure being correlated with acceptability. The underlying mechanisms that could result in that correlation have yet to be solidified. Exposure is but one piece in the puzzle of language and cognition.

Appendix A

Extra Materials Information

Names used in experiments are the 50 most popular girl and boy names from 2020 as claimed by a potentially reputable website[a]:

- Sophia, Olivia, Riley, Emma, Ava, Isabella, Aria, Aaliyah, Amelia, Mia, Layla, Zoe, Camilla, Charlotte, Eliana, Mila, Everly, Luna, Avery, Evelyn, Harper, Lily, Ella, Gianna, Chloe, Adalyn, Charlie, Isla, Ellie, Leah, Nora, Scarlett, Maya, Abigail, Madison, Aubrey, Emily, Kinsley, Elena, Paisley, Madelyn, Aurora, Peyton, Nova, Emilia, Hannah, Sarah, Ariana, Penelope, Lila

- Liam, Noah, Jackson, Aiden, Elijah, Grayson, Lucas, Oliver, Caden, Mateo, Muhammad, Mason, Carter, Jayden, Ethan, Sebastian, James, Michael, Benjamin, Logan, Leo, Luca, Alexander, Levi, Daniel, Josiah, Henry, Jayce, Julian, Jack, Ryan, Jacob, Asher, Wyatt, William, Owen, Gabriel, Miles, Lincoln, Ezra, Isaiah, Luke, Cameron, Caleb, Isaac, Carson, Samuel, Colton, Maverick, Matthew

[a] `https://www.babycenter.com/baby-names/most-popular/top-baby-names-2020`

References

Abeillé, A., Hemforth, B., Winckel, E., and Gibson, E. (2020). Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition*, 204:104293.

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

Berwick, R. C. and Chomsky, N. (2016). *Why only us: Language and evolution*. MIT press.

Bloomfield, L. (1926). A set of postulates for the science of language. *Language*, 2(3):153–164.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Bock, K. (1989). Closed-class immanence in sentence production. *Cognition*, 31(2):163–186.

Bornstein, R. F. (1989). Exposure and affect: overview and meta-analysis of research, 1968–1987. *Psychological bulletin*, 106(2):265.

Bornstein, R. F. and D'Agostino, P. R. (1994). The attribution and discounting of perceptual fluency: Preliminary tests of a perceptual fluency/attributional model of the mere exposure effect. *Social Cognition*, 12(2):103–128.

Branigan, H. P., Pickering, M. J., Liversedge, S. P., Stewart, A. J., and Urbach, T. P. (1995). Syntactic priming: Investigating the mental representation of language. *Journal of Psycholinguistic Research*, 24(6):489–506.

Brown, J., Fanselow, G., Hall, R., and Kliegl, R. (2021). Middle ratings rise regardless of grammatical construction: Testing syntactic variability in a repeated exposure paradigm. *Plos one*, 16(5):e0251280.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1):395–411.

Chaves, R. P. and Dery, J. E. (2014). Which subject islands will the acceptability of improve with repeated exposure. In *Proceedings of the 31st West Coast Conference on Formal Linguistics*, pages 96–106. Citeseer.

Chaves, R. P. and King, A. (2019). A usage-based account of subextraction effects. *Cognitive Linguistics*, 30(4):719–750.

Chomsky, N. (1957). *Syntactic structures.* De Gruyter Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax.* MIT press.

Chomsky, N. (1975). Reflections on language.

Chomsky, N. (1986). *Knowledge of language: Its nature, origin, and use.* Greenwood Publishing Group.

Chomsky, N., Culicover, P. W., Wasow, T., Akmajian, A., et al. (1977). On wh-movement. *1977*, 65.

Chomsky, N. and Lasnik, H. (1993). The theory of principles and parameters. In *Syntax*, pages 506–569. De Gruyter Mouton.

Crawford, J. (2012). Using syntactic satiation to investigate subject islands. In *Proceedings of the 29th West Coast Conference on Formal Linguistics*, pages 38–45. Cascadilla Proceedings Project Somerville, MA.

Crump, M. J., McDonnell, J. V., and Gureckis, T. M. (2013). Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410.

Elman, J. L. (1995). Language as a dynamical system. *Mind as motion: Explorations in the dynamics of cognition*, pages 195–223.

Erteschik-Shir, N. (1973). *On the nature of island constraints.* PhD thesis, Massachusetts Institute of Technology.

Fanselow, G. and Weskott, T. (2011). On the informativity of different measures of linguistic acceptability. *Language*, 87:249–273.

Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard university press.

Francom, J. C. (2009). *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure—evidence from rating and reading tasks.* The University of Arizona.

Gershman, S. J., Horvitz, E. J., and Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349(6245):273–278.

Goldberg, A. E. (2013). Backgrounded constituents cannot be "extracted". *Experimental syntax and island effects*, 221.

Green, P. and MacLeod, C. J. (2016). simr: an r package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4):493–498.

Hawkins, J. A. (1999). Processing complexity and filler-gap dependencies across grammars. *Language*, pages 244–285.

Hofmeister, P., Casasanto, L. S., Sag, I. A., Sprouse, J., and Hornstein, N. (2013). Islands in the grammar? standards of evidence. *Experimental syntax and island effects*, 42.

Kuno, S. (1987). *Functional syntax: Anaphora, discourse and empathy.* University of Chicago Press.

Legate, J. A. (2021). Noncanonical passives: A typology of voices in an impoverished universal grammar. *Annual Review of Linguistics*, 7:157–176.

Liu, Y., Ryskin, R., Futrell, R., and Gibson, E. (2021a). A verb-frame frequency account of constraints on long-distance dependencies in English. *Cognition*, page 104902.

Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., and Gibson, E. (2021b). Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics*, 8.

Luka, B. J. and Barsalou, L. W. (2005). Structural facilitation: Mere exposure effects for grammatical acceptability as evidence for syntactic priming in comprehension. *Journal of Memory and Language*, 52(3):436–459.

Mahowald, K., Hartman, J., Graff, P., and Gibson, E. (2016a). Snap judgments: A small n acceptability paradigm (snap) for linguistic acceptability judgments. *Language*, pages 619–635.

Mahowald, K., James, A., Futrell, R., and Gibson, E. (2016b). A meta-analysis of syntactic priming in language production. *Journal of Memory and Language*, 91:5–27.

Marr, D. and Ullman, S. (1981). Directional selectivity and its use in early visual processing. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 211(1183):151–180.

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

McKeown, K. R. and Radev, D. R. (2000). Collocations. *Handbook of Natural Language Processing. Marcel Dekker*, pages 1–23.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ross, J. R. (1967). Constraints on variables in syntax.

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. Language Science Press.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic inquiry*, 31(3):575–582.

Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1:123–134.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, 40(2):329–341.

Sprouse, J., Schütze, C., and Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua*, 134:219–248.

Thomas, K. A. and Clifford, S. (2017). Validity and mechanical turk: An assessment of exclusion methods and interactive experiments. *Computers in Human Behavior*, 77:184–197.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., and van Aken, M. A. (2014). A gentle introduction to bayesian analysis: Applications to developmental research. *Child Development*, 85(3):842–860.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break the rules of language*. MIT press.

Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1.

Zervakis, J. and Mazuka, R. (2013). Effect of repeated evaluation and repeated exposure on acceptability ratings of sentences. *Journal of Psycholinguistic Research*, 42(6):505–525.