

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

The mental representation of syntax: Interfaces with production, comprehension, and learning

Permalink

<https://escholarship.org/uc/item/09d61510>

Author

Morgan, Adam Milton

Publication Date

2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

The mental representation of syntax: Interfaces with production, comprehension, and learning

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Experimental Psychology

by

Adam Milton Morgan

Committee in charge:

Professor Victor S. Ferreira, Chair
Professor David Barner
Professor Tim Brady
Professor Grant Goodall
Professor Eva Wittenberg

2019

Copyright

Adam Milton Morgan, 2019

All rights reserved.

The Dissertation of Adam Milton Morgan is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2019

TABLE OF CONTENTS

Signature Page	iii
Table of Contents	iv
List of Figures	vii
List of Tables	ix
Acknowledgements	xii
Vita	xiii
Abstract of the Dissertation	xiv
Introduction	1
Chapter 1 A weird structure that we wonder why English speakers produce it: Multi- paradigm evidence that resumptive pronouns hinder comprehension	5
1.1 Introduction	5
1.1.1 Resumptive pronouns in English	7
1.1.2 Explaining the acceptability-production paradox	9
1.1.3 The present study	16
1.2 Experiment 1: Sentence-Picture Matching	17
1.2.1 Method	18
1.2.2 Analysis	20
1.2.3 Results	21
1.2.4 Discussion	23
1.3 Experiment 2: Self-Paced Reading	24
1.3.1 Method	24
1.3.2 Analysis	30
1.3.3 Results	30
1.3.4 Discussion	34
1.4 Experiment 3: Visual World Eyetracking	35
1.4.1 Method	36
1.4.2 Analysis	41
1.4.3 Results	42
1.4.4 Discussion	46
1.5 Experiment 4: Ordinary Pronoun Comprehension	48
1.5.1 Method	49
1.5.2 Results	52
1.5.3 Discussion	53
1.6 Experiment 5: Production	55
1.6.1 Method	55

1.6.2	Data Coding and Analysis	58
1.6.3	Results	59
1.6.4	Discussion	59
1.7	General Discussion	60
1.7.1	Chance Performance Supports Production-Based Accounts	62
1.7.2	Other reasons to doubt the Facilitation Hypothesis	63
1.7.3	The relationship between comprehension and production	65
1.7.4	Ways to salvage the Facilitation Hypothesis	67
1.7.5	A new perspective on the competence/performance distinction	69
1.7.6	Resumption: A cautionary tale for language comprehension research ...	70
1.8	Conclusion	70
1.9	Acknowledgment	71
Chapter 2	Learning untrained syntactic structures: Behavioral evidence for abstract representations of abstract representations	72
2.1	Introduction	72
2.1.1	Relative clauses	76
2.1.2	Language learning as a tool to study syntactic representation	77
2.1.3	The present study	79
2.2	Experiment 1a	83
2.2.1	Method	84
2.2.2	Results	91
2.2.3	Discussion	96
2.3	Experiment 1b	97
2.3.1	Method	97
2.3.2	Results	98
2.3.3	Discussion	100
2.4	Experiment 2	102
2.4.1	Method	102
2.4.2	Results	107
2.4.3	Discussion	108
2.5	Experiment 3	109
2.5.1	Method	110
2.5.2	Results	111
2.5.3	Discussion	114
2.6	General Discussion	115
2.6.1	Implications for theories of language acquisition	118
2.6.2	Asymmetries in generalization	120
2.7	Conclusion	121
Chapter 3	Individual differences reveal gradience in syntactic representations	123
3.1	Introduction	123
3.1.1	Individual differences	125
3.1.2	The present study	128

3.2	Experiment 1	132
3.2.1	Method	132
3.2.2	Analysis.....	136
3.2.3	Results.....	138
3.2.4	Discussion.....	141
3.3	Experiment 2	145
3.3.1	Norming task	145
3.3.2	Method	147
3.3.3	Pre-registered analyses	150
3.3.4	Results.....	152
3.3.5	Discussion.....	158
3.4	General Discussion	160
3.4.1	What is syntactic strength?	162
3.5	Conclusion	162
Chapter 4	Conclusion	164
Bibliography	168

LIST OF FIGURES

Figure 1.1.	Two production models.	15
Figure 1.2.	Sample display from Experiment 1.	18
Figure 1.3.	Experiment 1 results: (A) all responses and (B) just <i>target</i> and <i>local</i> responses — i.e., those included in the analysis	22
Figure 1.4.	Experiment 2 results: Accuracy on filler trials, by filler type (see Table 1.5)	31
Figure 1.5.	Experiment 2 results: (A) all responses and (B) just <i>target</i> and <i>local</i> responses — i.e., those included in the analysis	31
Figure 1.6.	Experiment 2 self-paced reading results.	33
Figure 1.7.	Experiment 3 results of the multiple choice (interpretation) question: (A) all responses and (B) just <i>target</i> and <i>local</i> responses — i.e., those included in the analysis	43
Figure 1.8.	Experiment 3 gaze data.	44
Figure 1.9.	Gaze data during the gap or resumptive pronoun for all three island types.	45
Figure 1.10.	Screenshot from Experiment 4.	49
Figure 1.11.	Experiment 4 results: (A) all responses and (B) just <i>distant/target</i> and <i>local</i> responses — i.e., those included in the analysis	53
Figure 1.12.	Screenshot of Experiment 5. The trial shown here is the <i>strong island</i> condition. Participants were instructed to type in the box to complete the new sentence using the information in the context sentence.	56
Figure 1.13.	Experiment 5 results: (A) all responses and (B) just <i>gap</i> and <i>resumptive pronoun</i> responses — i.e., those included in the analysis	60
Figure 2.1.	Experiment 1a results.	92
Figure 2.2.	Experiment 1a: The relationship between learning trained structures (horizontal axis) and generalization to untrained structures (vertical).	94
Figure 2.3.	Proportion well-formed responses of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 1b.	99
Figure 2.4.	Experiment 1b: The relationship between learning trained structures and generalization to untrained structures.	100

Figure 2.5.	Proportion of well-formed productions of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 2.	107
Figure 2.6.	Experiment 2: The relationship between learning trained structures and generalization to untrained structures.	109
Figure 2.7.	Proportion of well-formed productions of the elicited type as a function of GROUP and TRIAL TYPE in Experiment 3.	114
Figure 2.8.	Experiment 3: The relationship between learning trained structures and generalization to untrained structures.	115
Figure 3.1.	Screenshot of a production trial from Experiment 1.	133
Figure 3.2.	Experiment 1 results. Best-fit lines plotted in green; density plots in purple along axes. Point size is proportional to the observation's weight in the model.	139
Figure 3.3.	Experiment 2: Production as a function of preference in data from the pre-test phase. Point size is proportional to the observation's weight in the model. Best-fit lines plotted in green; density plots in purple along axes. . .	152
Figure 3.4.	Experiment 2: Production data for dative (left) and wh-island (right) stimuli.	153
Figure 3.5.	Experiment 2: Acceptability ratings for all four structure types by condition.	155
Figure 3.6.	Experiment 2: The amount of change in production rate of DOs (left) or gaps (right) from pre-test to post-test.	157

LIST OF TABLES

Table 1.1.	Experiment 1 stimuli. Sentences appeared in a 2×3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.	19
Table 1.2.	Experiment 1 results.....	22
Table 1.3.	Experiment 2 stimuli. Sentences appeared in a 2×3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.	25
Table 1.4.	Response options for Experiment 2: Options listed here correspond to the item set given in Table 1.3 and were the same for all six conditions	26
Table 1.5.	Experiment 2 fillers.	27
Table 1.6.	Summary of the possible comprehension heuristics that the fillers were designed to prevent. The specific filler types that were designed to prevent each strategy are listed in the right hand column.	28
Table 1.7.	Experiment 2 results: Multiple choice interpretation responses.	32
Table 1.8.	Experiment 2 results: Reading speed.	34
Table 1.9.	Written versions of the auditory Experiment 3 stimuli.....	38
Table 1.10.	Response options for Experiment 3.	38
Table 1.11.	Animal characters for Experiment 3	39
Table 1.12.	Experiment 3 results: Multiple choice interpretation responses.	43
Table 1.13.	Experiment 3 results: Gaze.	46
Table 1.14.	Stimuli for Experiment 4.	51
Table 1.15.	Response options for Experiment 4. As in previous experiments, the comprehension question was, “Who did what to whom?”	52
Table 1.16.	Experiment 4 results: Multiple choice interpretation responses.	52
Table 1.17.	Experiment 5 stimuli and target responses. Participants were instructed to complete the sentence started by the prompt using the information given by the context sentence.	57

Table 1.18.	Experiment 5 coding rubric with frequency of each response type in each of the three conditions and examples.	59
Table 1.19.	Experiment 5 results.....	59
Table 2.1.	Relative clauses vary along two dimensions: position of the relative clause relative to the head noun (columns) and position of the gap (rows).	77
Table 2.2.	Sample materials from Experiments 1a and 1b.	86
Table 2.3.	Percentage of the most common errors in Experiment 1a, by condition. ...	90
Table 2.4.	Experiment 1a results.....	93
Table 2.5.	Experiment 1b results.	99
Table 2.6.	Experiment 2: sample stimulus items from Training Phase 1.	105
Table 2.7.	Experiment 2: sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.....	105
Table 2.8.	Percentage of the most common errors in Experiment 2, by condition.	106
Table 2.9.	Experiment 2 results.....	108
Table 2.10.	Experiment 3: Sample stimulus items from Training Phase 1.....	111
Table 2.11.	Experiment 3: Sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.....	112
Table 2.12.	Percentage of the most common errors in Experiment 3, by condition.	113
Table 2.13.	Experiment 3 results.....	113
Table 3.1.	Examples of each of the four structural alternates we used in our stimuli. ...	130
Table 3.2.	Experiment 1 production stimuli and target responses. Participants used the information given by the context sentence to complete the sentence begun by the prompt, as in Figure 3.1.	135
Table 3.3.	Experiment 1 results: all four structure types.	140
Table 3.4.	Experiment 1 results: Individual differences in the pre-test data.....	153
Table 3.5.	Experiment 2 results: Priming.	154
Table 3.6.	Experiment 2 results: Satiation.....	156

Table 3.7. Experiment 2 results: Relationship between the size of the priming effect and the size of the satiation effect. 157

ACKNOWLEDGEMENTS

Thanks to Vic Ferreira for his invaluable guidance and encouragement throughout my graduate education. I have regularly marveled at his constant positive attitude and ability to find something interesting in every domain of psychology, even the comprehension of resumptive pronouns.

Thanks to Eva Wittenberg, who has been a constant source of support and insight for the majority of my graduate education. I am proud to consider myself a founding member of her Language Comprehension Lab and very sad to be leaving it.

Thanks to Titus von der Malsburg, who has been a joy to collaborate with, and from whom I have learned many valuable lessons about writing and with his helpful comments.

Thanks to the rest of my committee for their helpful feedback and patience with the many iterations of some of the projects reported here.

Thanks to all of the dedicated, diligent, and generally wonderful RAs who have helped with the research presented here.

Thanks to all of my current and former lab mates, without whom my graduate education would have been far more productive (but much more boring).

Thanks to my previous mentors Matt Wagers and Masha Polinsky, who both instilled in me a fascination for linguistic structure and a love of formal analysis while guiding me toward the field I have finally (sort of) settled in.

Thanks also to my friends and family for their constant love and support, and especially for putting up with me during the preparation of this dissertation.

Chapter 1 is a reprint of the material submitted to *Cognition*. Morgan, Adam M.; von der Malsburg, Titus; Ferreira, Victor S.; and Wittenberg, Eva. The dissertation author was the primary investigator and author of this paper.

VITA

- 2008.5 Bachelor of Arts, Physics, Music, Middlebury College
2011–2013 M.A., Linguistics, UC Santa Cruz
2013–2019 Ph.D., Experimental Psychology, UC San Diego

ABSTRACT OF THE DISSERTATION

The mental representation of syntax: Interfaces with production, comprehension, and learning

by

Adam Milton Morgan

Doctor of Philosophy in Experimental Psychology

University of California San Diego, 2019

Professor Victor S. Ferreira, Chair

In three sets of experiments, this dissertation investigates the mental representation of syntactic structure. Chapter 1 sheds light on an ongoing debate between two models that aim to account for the regular production of an ungrammatical structure, *resumption*. Results support a production account rather than an audience-design account. Chapter 2 uses an artificial language learning paradigm to determine whether various *long-distance dependencies* are represented independently or as the same structure. Results indicate that the representation of these structures is unitary in some sense. Chapter 3 looks at individual differences in the production and comprehension of various structures to determine whether syntax is purely abstract, as generally assumed, or if it exists in a gradient representation space. Results reveal individual differences,

suggesting that syntax does in fact have a gradient component. Together, these studies contribute to a growing body of work indicating that syntactic representations are complex and multifaceted, and require a more nuanced model than is often assumed.

Introduction

This dissertation is comprised of three investigations into the nature of the mental representation of syntax. *Syntax* is, broadly speaking, whatever knowledge speakers have about their language(s) that leads English speakers to utter subjects before verbs, but Hawaiian speakers to utter verbs before subjects. In the present work it is largely used interchangeably with *grammar*.

Throughout the dissertation, the mental representations of particular syntactic structures (e.g., a passive sentence) are referred to as *syntactic representations*. The terms “grammatical” or “syntactically well-formed,” are used to refer to whether speakers have a mental representation of a given structure, not whether a sentence sounds acceptable. *Acceptability* is taken to be a consequence of grammaticality. Therefore, acceptability, along with production, is treated as a behavioral metric which can be used to assess a structure’s grammaticality.

Chapter 1 investigates a particular structure, *resumption*, that is purported to challenge this framework. Typically, if a structure is grammatical, then it is reliably produced and it has high acceptability. However this does not apply to resumption, or the use of *resumptive pronouns*. Take for instance: “a pronoun that we investigate why people say **it**.” Structures like this are regularly produced by native English speakers (Ferreira and Swets, 2005; Morgan and Wagers, 2018; Fadlon et al., 2019), but consistently judged to be unacceptable (Alexopoulou and Keller, 2007; Polinsky et al., 2013; Heestand et al., 2011; Han et al., 2012).

This pattern of data is confirmed by a number of acceptability judgment studies and a growing number of production studies. It is potentially problematic for work that relies on direct interpretation of metrics like acceptability and production for theory building. However,

two recent proposals seek to explain resumptive pronouns as the result of production processes gone awry. If this is the case, then the fact that resumptive pronouns are produced would not be an indicator of grammaticality, and there would therefore be no disagreement between the production data and the acceptability data. In Chapter 1, we investigate a prediction of this account which would directly contradict a competing proposal, namely, the idea that resumptive pronouns are a tool used by speakers to facilitate comprehension for the listener.

In Chapter 2, an artificial language learning task is employed to better understand the makeup of the mental representation of complex structures. It is widely accepted that language is hierarchically structured — that is, sentences can be decomposed into groups of adjacent units (*phrases*) from the sentence level all the way down to the level of individual words. But according to nearly all models of syntax, hierarchical structure is not the full extent of the complexity of natural language.

Consider a grammatical sentence like “The teacher gave a present to the girl.” Here, “to the girl” forms a complete unit. This readily lends itself to a hierarchical model: “the” and “girl” combine to form a Noun Phrase, and “to” combines with the Noun Phrase to form a Prepositional Phrase. But now consider the sentence “The boy was jealous of the girl who the teacher gave a present to.” Again, the object of “to” is “the girl” — there is no ambiguity about where the present went, even though the boy is an equally plausible recipient. But now, the Prepositional Phrase cannot be formed with adjacent units.

In any given language, there are typically a vast number of structures where, as in the previous example, a phrasal unit is noncontiguous. In principle, language learners could come to model all of these on a case-by-case basis — that is, there could be an independent representation for each such structure, where each representation directly reflects the surface word order. Indeed, a number of psycholinguistic theories tacitly assume this to be the way these structures are represented.

But a number of typological facts — generalizations about what structures do and do not exist among the world’s languages — seem to suggest that there is a more general representation

of these structures, one which would require adding to models of syntactic structure such that they are more complex than just hierarchical structures. Indeed, this is the approach taken by almost all theoretical models of syntax. In Chapter 2, we aim to bring experimental data to bear on the question of whether these structures are represented in this more general, complex fashion.

Chapter 3 calls into question the common assumption that all native speakers of a given language have the same syntactic representations. We contrast recent exemplar-based models of phonemic and lexical representations with the traditional, but still pervasive view of syntactic representations as purely abstract. To determine whether this view still has merit, or whether, as with other linguistic representations, syntactic representations should be thought of as existing, at least partially, in a continuous representation space, we look for signatures of individual differences in syntactic representation.

Previous work has demonstrated that there are cases where speakers' representations differ on a categorical level. For instance, speakers of Polish seem to have different conditions under which the use of a particular genitive suffix should be used (Dabrowska, 2008). But here we ask whether some speakers have *stronger* representations of the same structures, as indexed by production tendencies and acceptability preferences.

Such a finding would not be unexpected on the view that syntax is a system that is constantly changing with experience. This is perhaps the dominant view in the *syntactic priming* literature, which holds that the increased likelihood of (implicitly) choosing to produce a given structure after exposure to that structure reflects learning (Chang et al., 2006; see Pickering and Ferreira, 2008, for a review). Less is known about the nature of *syntactic satiation*, or the increase in acceptability of a given structure after exposure, but it is not hard to imagine that this reflects the same mechanism. Two high-powered studies aim to determine whether individual differences in the strength of syntactic representations might reflect differences in the history of experience.

This dissertation thus investigates syntactic representation using a number of different experimental approaches. Taken together, the findings shed light on the immense complexity of

the system: it varies from person to person; abstract structures that show a family resemblance are represented as the same at some even higher level of abstraction; and a structure's regular production does not necessarily indicate that it is grammatical, even if similar production patterns in other languages does.

Chapter 1

A weird structure that we wonder why English speakers produce it: Multi-paradigm evidence that resumptive pronouns hinder comprehension

1.1 Introduction

A finding from moral cognition is that individuals often hold themselves to different ethical standards than others (Foschi, 2000). For instance, one might quickly condemn a child for a white lie, but then readily commit the same offense when asked what they thought of their boss's new shoes. A parallel situation arises in language, when native English speakers, who do not like the sound of a specific type of pronoun, produce this pronoun anyway. For instance, the word *she* in (1) is reportedly highly unacceptable, but such pronouns are nonetheless readily produced.

- (1) The royal nursemaid, who no one could understand how **she** hadn't noticed that the girl had gone missing, was promptly executed.

As metrics of grammaticality, acceptability and production nearly always align: speakers reliably produce the same kinds of sentences that they judge to be acceptable. So just as understanding why moral judgments change from situation to situation is fundamental to understanding morality, understanding why English speakers produce sentences like (1) may help us understand

fundamental issues about language. For a structure to be unacceptable, but reliably produced, poses problems for the way we typically think of grammar.

In order to model language, one needs to first know which structures are grammatical and which are not. To this end, researchers generally assume that *acceptability*, or how good a structure sounds to native speakers of a language, is a relatively straightforward indicator of grammaticality. Similarly, *production* patterns can also be informative: if a given structure is reliably produced by native speakers, then it is probably grammatical.

Imagine that you read a word, perhaps *resumption*, and you don't know what it means. Consider how you might complete the following sentence: "I read a word that I don't know what..." More often than not, English speakers would produce what is called a *resumptive pronoun* and say: "I read a word that I don't know what **it** means."

Resumptive pronouns, like the "she" in 1, are relatively frequent in English (Cann et al., 2005). Bennett (2009), for instance, found 61 instances of resumptive pronouns in the Switchboard corpus, and examples abound in natural speech:

(2) "We have these things called aircraft carriers, where planes land on them."

-Barack Obama (Davidson Sorkin, 2012)

(3) "...the sale of the uranium that nobody knows what it means."

-Donald Trump (Noble, 2017)

In experimental settings, resumptive pronouns can be reliably elicited, both in speech (Ferreira and Swets, 2005) and in writing (Morgan and Wagers, 2018), and both when speakers are under time pressure to respond and when they are not (Ferreira and Swets, 2005). Taken together, this seems to indicate that resumptive pronouns are grammatical in English.

But in experiments where native English speakers are asked about acceptability, they consistently rate resumptive pronouns as highly unacceptable (Alexopoulou and Keller, 2007; Heestand et al., 2011; Keffala and Goodall, 2011; Han et al., 2012; Polinsky et al., 2013). This is true across a wide variety of sentence types, both with written and auditory stimulus presentation

(Clemens et al., 2012; Heestand et al., 2011). Contrary to what we might conclude on the basis of the production facts, the acceptability data on the surface would seem to indicate that *resumption*, or the use of resumptive pronouns, is ungrammatical. Resumptive pronouns, then, present a case where two generally reliable metrics for grammaticality dissociate.

This apparent paradox is one of the most intriguing puzzles in the field. It indicates that either our assumptions about acceptability and production are flawed, or that resumptive pronouns are an exception to the rule. One prominent account, the *Facilitation Hypothesis*, argues that resumptive pronouns are produced to help listeners keep track of reference. However, in four experiments, we demonstrate that when this hypothesis is tested directly, the data instead indicate that resumptive pronouns *hinder* comprehension, suggesting that the Facilitation Hypothesis should be rejected in favor of production models of resumption.

The remainder of this paper is structured as follows: In the following section, we give a brief theoretical overview of resumptive pronouns and discuss prominent approaches to resolving the acceptability-production paradox, with a particular focus on the Facilitation Hypothesis. We then argue that to adequately test the Facilitation Hypothesis one must investigate how gaps and resumptive pronouns are interpreted, which has yet to be done. We do this in four experiments, which consistently provide evidence against the Facilitation Hypothesis, but are consistent with the predictions of production accounts of English resumption. We conclude by discussing the relationship between comprehension and production and the wider architecture of language.

1.1.1 Resumptive pronouns in English

Resumptive pronouns do not exist in isolation, but are parts of larger structures, such as *relative clauses*. In (4), for example, ‘...*that the fairies kidnapped in the night*’ is a relative clause that modifies ‘*girl*’:

- (4) the girl [that the fairies kidnapped _ in the night]

The modified noun (‘*girl*’) is referred to as the *head noun*. Note that the head noun is not repeated

inside the relative clause, resulting in an empty position referred to as a *gap* (indicated with underscores throughout). Structures like relative clauses, where the meaning of a gap position corresponds to that of the faraway head noun, are known as *wh-dependencies*. Throughout this paper, we will use relative clauses (and more specifically, *clefts*) to create wh-dependencies.

In English, leaving a gap is the only grammatical way to form wh-dependencies. In contrast, other languages employ resumption. Irish (McCloskey, 2002), Hebrew (Shlonsky, 1992), Gbadi (Koopman, 1983), and Cantonese (Lau, 2016), for instance, allow speakers the option of inserting a pronoun, as in (1,2,3) and (5):

(5) the girl [that the fairies kidnapped her in the night]

These resumptive pronouns are different from ordinary pronouns in the way that they are interpreted. Ordinary pronouns, such as ‘*him*’ in (6), may refer to any number of potential referents, including ones in the same sentence (“someone” or “advisor”), or even ones outside the sentence (“king”):

(6) [Context: *The king_i was furious.*]

Someone_j told his advisor_k that the fairies had kidnapped the girl to humiliate him_{i/j/k}.

In contrast, resumptive pronouns have the same referential properties as gaps. By being part of a wh-dependency, their reference is fixed: they must refer to the head noun (e.g., Zaenen et al., 1981). So, in (4) and (5), the object of “kidnap” must be “the girl.”

In a production study, Morgan and Wagers (2018) asked participants to produce relative clauses with different kinds of clausal structures. They found that English speakers produce resumptive pronouns in a class of structures known as *islands*, so called because nouns cannot “escape” from them. In less metaphorical terms, islands are structures which are acceptable under normal circumstances, but unacceptable when a gap appears inside of them. For instance, (4) is a *non-island*, meaning that it is acceptable with a gap in it (and without, as in “*The fairies kidnapped the girl in the night*”). Morgan and Wagers’s participants produced fewer than 5% resumptive pronouns (and more than 95% gaps) in non-islands. Examples (7) and (8) contain

island structures. As demonstrated by (7a) and (8a), they are grammatical and acceptable on their own. However, when gaps appear in islands, as in (7b) and (8b), they become unacceptable (indicated with an asterisk). *Island violations* like these vary in their degree of unacceptability (Ross, 1967). Example (7) is a *weak* island, and is only moderately unacceptable with a gap, while (8) is a *strong* island and is highly unacceptable with a gap.¹

- (7) a. I wonder why the fairies kidnapped the girl. WEAK ISLAND
 b. * the girl [that I wonder why the fairies kidnapped _]
- (8) a. I was scared because the fairies kidnapped the girl. STRONG ISLAND
 b. ** the girl [that I was scared because the fairies kidnapped _]

Morgan and Wagers (2018) showed that English speakers produce resumptive pronouns in a given structure at a rate that correlates strongly with that structure’s degree of unacceptability. In weak islands, English speakers produce close to 50% resumptive pronouns, and over 90% in strong islands.

1.1.2 Explaining the acceptability-production paradox

There are several families of explanation for the acceptability-production hypothesis. Here, we focus on the two dominant approaches: the *Amelioration Hypothesis* and production-based accounts. The Amelioration Hypothesis posits that resumptive pronouns serve to fix some problem – either ungrammaticality (Kroch, 1981; McDaniel and Cowart, 1999), low acceptability (Alexopoulou and Keller, 2007; Heestand et al., 2011; Ackerman et al., 2018; Keffala and Goodall, 2011; Han et al., 2012; Polinsky et al., 2013), or low comprehensibility (Hofmeister and Norcliffe, 2013; Beltrama and Xiang, 2016). Production-based accounts, on the other hand, hold that speakers produce these structures to get themselves out of tricky situations

¹ *Weak* and *strong* islands are actually classes of island structures, each composed of many different structures. In this study, we will use structures called *wh-islands* (as in 7) to operationalize weak islands, and *adjunct islands* (as in 8) as strong islands. For the sake of readability, we use the more intuitive labels *weak island* and *strong island* throughout.

(Asudeh, 2004, 2011; Morgan and Wagers, 2018; Polinsky et al., 2013; Kroch, 1981). We discuss each of these proposals below.

According to the Amelioration Hypothesis, resumptive pronouns might be a way for speakers to make a bad sentence sound a little better. This idea seems to be supported by the finding that English speakers produce resumptive pronouns where gaps would be unacceptable (Morgan and Wagers, 2018). It would resolve the acceptability-production paradox because resumptive pronouns would not have to be acceptable to explain their production; they just have to be *more* acceptable than gaps. Experimental work, however, does not clearly support this hypothesis.

In six two-alternative forced-choice experiments, Ackerman et al. (2018) had participants select the better option between two sentences which were identical but for a gap or resumptive pronoun. Across three types of island structures, participants chose sentences/phrases with resumptive pronouns more often than they chose gaps. Ackerman et al. (2018) argue that their data indicate resumptive pronouns are more acceptable than gaps. However, it is not clear how to interpret data from this unusual paradigm. A participant might consider a sentence's acceptability in making the decision, as Ackerman et al. (2018) argue, but the participant might also choose based on which option they would be more likely to produce, in which case this is closer to a production measure. Indeed, Morgan and Wagers (2018) show that resumptive pronouns are produced at rates similar to those at which Ackerman et al.'s (2018) participants selected resumptive pronoun alternatives. Critically, however, the best predictor of when a resumptive pronoun would be produced in Morgan and Wagers's (2018) data was the acceptability of the sentence with a gap in it, and *not* the acceptability of the sentence with the resumptive pronoun. If this is the case, then Ackerman et al.'s (2018) findings may say more about the acceptability of gaps than of resumptive pronouns. In addition, the fact that Ackerman et al. (2018) found high rates of choosing non-island sentences with resumptive pronouns, which are clearly unacceptable relative to gapped versions, presents a puzzle that renders the overall pattern of results harder to interpret.

More standard acceptability rating studies have produced even weaker evidence for the Amelioration Hypothesis. In one representative study, Han et al. (2012) performed a series of acceptability judgment tasks on sentences with various types of non-islands (like 4) and islands (like 7 and 8) with gaps and resumptive pronouns. They showed that across island and non-island conditions, resumptive pronouns were rated similarly low. They noted that there were cases where resumptive pronouns were rated better than gaps (for instance, in strong islands like (8b), but calling this evidence for amelioration overlooks the fact that the resumptive pronouns in these structures were still unacceptable. That is, resumptive pronouns had roughly the same acceptability in these structures as they did in others, where they were rated worse than gaps. Instead, a better characterization is that there are some cases where gaps render a structure so unacceptable that they are even worse than the corresponding resumptive pronouns. Similar findings have been reported in a number of other studies (e.g., Alexopoulou and Keller, 2007; Heestand et al., 2011; Keffala and Goodall, 2011; Han et al., 2012; Polinsky et al., 2013), making it unlikely that that resumptive pronouns serve to improve acceptability (or grammaticality, as argued by Kroch, 1981 and McDaniel and Cowart, 1999; although see Chomsky, 1991; Shlonsky, 1992 for arguments that it may not always be the case that grammaticality implies acceptability, and Miller, 1962; Gibson and Thomas, 1999; Paape et al., 2019 for examples of other structures for which grammaticality and acceptability dissociate).

Another version of the Amelioration Hypothesis that has gained traction is the *Facilitation Hypothesis*, the idea that resumptive pronouns are easier to comprehend than gaps (e.g., Prince, 1990; Dickey, 1996; Polinsky et al., 2013; Hofmeister and Norcliffe, 2013; Beltrama and Xiang, 2016). It is easy to imagine why this might be the case. Resumptive pronouns provide an overt cue for the end of a wh-dependency, along with number, gender, and animacy information which might be helpful for retrieving the correct antecedent (Ferreira and Swets, 2005). Gaps, by comparison, do none of this.

To test the Facilitation Hypothesis, Hofmeister and Norcliffe (2013) conducted a self-paced reading study. They presented subjects with sentences that had gaps or resumptive

pronouns in short (9a) or long (9b) dependencies. Participants read the sentences one word at a time in a moving window paradigm, pressing a button to reveal each word so that researchers could measure how long each word took to read. After each sentence, subjects answered a comprehension question (9c).

- (9) a. The prison officials had acknowledged that there was a prisoner that the guard helped (him) to make a daring escape.
- b. Mary confirmed that there was a prisoner who the prison officials had acknowledged that the guard helped (him) to make a daring escape.
- c. Was a prisoner able to escape?

Hofmeister and Norcliffe’s data revealed that in longer, more difficult dependencies, the words immediately following resumptive pronouns were read faster than those following gaps. They interpreted this result as reflecting “more efficient processing,” a sign that “the resumptive pronoun facilitates processing compared to a gap.”

Faster reading times, however, do not necessarily imply facilitated comprehension. Indeed, a number of possibilities are compatible with this pattern of data. One is that gap and resumptive pronoun dependencies are equally easy or difficult to process, but because resumptive pronouns take longer to read than gaps, they spread the same amount of information across more words. If, as some evidence suggests, the system aims to process a uniform amount of information per unit time (Jaeger and Levy, 2007), then Hofmeister and Norcliffe’s (2013) subjects may have sped up after resumptive pronouns because they had to process less information per word, not because the resumptive pronoun made parsing easier.

Another possibility, which we will return to throughout the paper, is that readers are simply confused by resumptive pronouns. The decrease in reading times doesn’t reflect facilitation, but instead giving up on parsing and clicking through to end the trial. Given that resumptive pronouns are rare in English, and particularly in non-island contexts like those in Hofmeister and Norcliffe’s stimuli, this seems like a more likely interpretation of their results than facilitation.

(See Ferreira et al., 2002; Nicenboim et al., 2016 for related proposals.)

In order to infer that the faster reading times after resumptive pronouns reflect facilitated processing, one would need to minimally establish that participants interpreted resumptive pronoun stimuli at least as correctly as gap stimuli. That is, in order to evaluate the usefulness of resumption in comprehension, one must also measure *interpretation*.

Hofmeister and Norcliffe (2013) did not report the interpretation data collected from comprehension questions. They did, however, remove data from trials that were interpreted incorrectly before performing their analysis on reading times. On the surface, this would seem to ensure that faster reading times were measured only on correctly parsed trials. But, as exemplified by (9), their stimuli were pragmatically rich: Just reading the words *prisoner*, *prison guard*, *help*, and *escape* alone can conjure up a plausible scenario, without needing to establish a syntactic parse (see Mollica et al., 2018 for neural evidence that adjacent words are processed compositionally even when they do not form grammatical strings).

As a result, it may have been possible for participants to correctly answer comprehension questions based on the lexical content alone, and not on the basis of a correctly parsed dependency. If this is the case, then it may be even more likely that faster reading times after resumptive pronouns reflect giving up on parsing and not facilitated parsing. Without more information about exactly how subjects parsed the sentences, Hofmeister and Norcliffe's data cannot definitively answer the question of whether resumptive pronouns facilitate comprehension.

Beltrama and Xiang (2016) also tested the Facilitation Hypothesis by asking subjects to rate sentences for comprehensibility. Their stimuli consisted of context sentences (10a) followed by target sentences, which were manipulated to appear with gaps or resumptive pronouns in non-islands (10b) or islands (10c).

- (10) a. Have you heard? Yesterday there were riots in the streets. Some people were wounded. Look here, they're talking about it in the paper.
- b. This is the boy that the cop who was leading the operation beat (him) up.

c. This is the boy that the cop who beat (him) up was leading the operation.

They found that in non-islands, gaps were rated as more comprehensible than resumptive pronouns.² In islands, on the other hand, resumptive pronouns were rated as more comprehensible than gaps. Beltrama and Xiang took these results to be consistent with a modified version of the Facilitation Hypothesis: that resumptive pronouns serve to facilitate processing, but only in islands.

Like Hofmeister and Norcliffe (2013), Beltrama and Xiang (2016) used stimuli that provided readers with heavy pragmatic cues and did not report how their participants interpreted them. As was true for reading times, comprehensibility ratings alone are not sufficient. It is in principle possible that resumptive pronouns lead comprehenders to interpret sentences less correctly, but to nonetheless feel that they are interpreting them more correctly. Knowing how participants interpret resumptive pronouns is therefore crucial.

What remains to be tested in this literature is the keystone prediction of the Facilitation Hypothesis: that resumptive pronouns result in more accurate interpretation than gaps. For a system whose goal is communication, the worst possible outcome of comprehension processes is incorrect interpretation. One might even consider decreased processing speeds facilitatory if they corresponded to an increase in correct interpretation. But a decrease in interpretation accuracy can never constitute facilitation. If resumptive pronouns make the listener less likely to understand the intended meaning, then it hardly matters whether they do so in less time or with more confidence. Interestingly however, a decrease in interpretation accuracy is exactly the prediction of the second family of explanations of the acceptability-production paradox.

Production-based theories attempt to resolve the paradox by assuming that resumptive pronouns are ungrammatical, straightforwardly accounting for their unacceptability, and explaining their production in terms of difficulties in online production processes. Asudeh (2004)

²Note that Hofmeister and Norcliffe's (2013) stimuli were all non-islands, but they came to the opposite conclusion on the basis of faster reading times. This discrepancy may trace back to the common simplifying assumption that faster reading times mean more efficient processing.

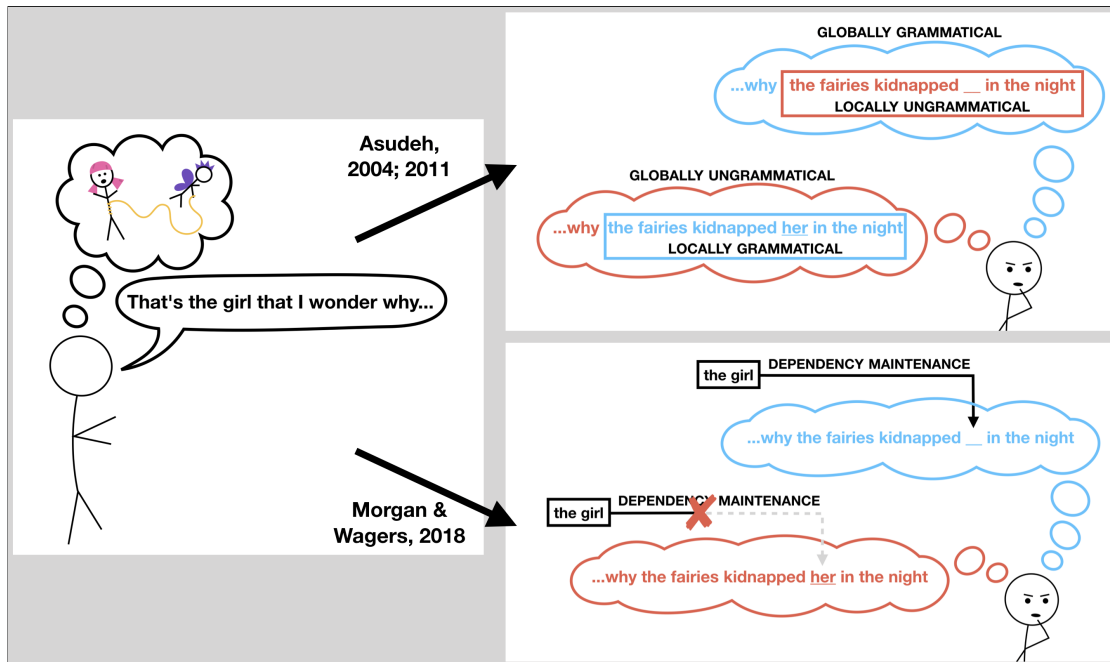


Figure 1.1. Two production models. Asudeh (2004, 2011) (top) argues that when speakers produce resumptive pronouns, they have chosen to create a locally grammatical structure at the expense of global grammaticality. Morgan and Wagers (2018) (bottom) argue that resumptive pronouns occur when the production system fails to maintain a gap dependency.

argues that in the course of producing a filler-gap dependency, people may opt to satisfy local constraints for grammaticality at the expense of global constraints (or vice versa). When they obey local constraints, as in the top panel of Figure 1.1, they produce a locally grammatical phrase containing a pronoun. This strategy results in a globally ungrammatical structure.

In an acceptability rating experiment, Morgan and Wagers (2018) measured the acceptability of gaps and resumptive pronouns in a variety of sentence structures (including non-islands, weak islands, and strong islands). Then, in a production experiment, they elicited the same sentences that participants had rated in the first experiment from a second group of participants. They measured how often gaps and resumptive pronouns were produced in these sentences. The data revealed that the less acceptable a gap-containing structure is, the more likely speakers are to insert a resumptive pronoun when producing sentences with that structure.

This finding led them to suggest that when resumptive pronouns are produced, what we

see is not the result of speakers planning and producing a resumptive pronoun dependency, but the result of speakers giving up on producing a gap dependency midway through the utterance, as roughly schematized in the bottom panel of Figure 1.1. Speakers give up more often in particularly unacceptable domains — presumably where production is more difficult. When speakers do give up on producing the dependency, gaps are no longer licensed, so a pronoun is produced. The result is a structure that looks like those in Irish, Hebrew, and Gbadi, but is in fact just an ordinary pronoun from the perspective of the production system.

For the purposes of this paper, if resumptive pronouns are indeed ungrammatical, then by definition it means that comprehenders cannot parse — or assign a grammatical structure to — resumptive dependencies. But when there is no grammatical structure, comprehension should be impaired. Thus, both Asudeh (2004, 2011) and Morgan and Wagers (2018) predict that resumptive pronouns should lead to *worse* comprehension than gaps. This is the opposite prediction of the Facilitation Hypothesis.

1.1.3 The present study

There is a growing body of data which, on the surface at least, seems to support the Facilitation Hypothesis (Hofmeister and Norcliffe, 2013; Beltrama and Xiang, 2016). Here we present four experiments which directly test the prediction that resumptive pronouns lead to *more*, but never *less* accurate interpretation than gaps. We do so by measuring how participants interpret sentences with gaps or resumptive pronouns in non-islands, weak islands, and strong islands. If speakers do indeed produce resumptive pronouns when they help comprehenders, then in structures where speakers produce them more frequently, resumptive pronouns should facilitate comprehension more. That is, any facilitation effect should be stronger in islands than in non-islands, but also stronger in strong islands than in weak islands.

We have suggested that Hofmeister and Norcliffe’s (2013) and Beltrama and Xiang’s (2016) pragmatically rich stimuli may have made it possible for their participants to rely on non-compositional strategies to interpret sentences with gaps versus resumptive pronouns. In

sentences with gaps, participants may have used both parsing and pragmatic cues to interpret sentences, but in sentences with resumptive pronouns, just relying on pragmatic cues may have sufficed to achieve a high rate of accuracy. To prevent our participants from using pragmatic cues for interpretation, we designed our stimuli using unfamiliar animal characters (e.g., Miss Rabbit, Mr. Froggy). Participants therefore had to rely on bottom-up syntactic processing to interpret gaps and resumptive pronouns.

Experiment 1 is a single-trial sentence-picture matching task where participants were presented with a sentence and four images representing possible interpretations. Experiment 2 is a self-paced reading task, a partial replication of Hofmeister and Norcliffe's (2013) experiment. Experiment 3 is an eyetracking study using a visual world paradigm, which allowed us to assess online sentence interpretation. Experiment 4 is a single-trial sentence comprehension task. In all experiments, the Facilitation Hypothesis makes the same prediction: resumptive pronouns should make the comprehender at least as likely to correctly interpret sentences as gaps. If, on the other hand, resumptive pronouns result in decreased interpretation accuracy, then they cannot be said to facilitate comprehension. This would be inconsistent with the Facilitation Hypothesis, but consistent with production accounts.

1.2 Experiment 1: Sentence-Picture Matching

In Experiment 1, a single-trial sentence-picture matching task, we asked participants to select one of four scenes reflecting possible interpretations of a sentence with a gap or a resumptive pronoun, as in Figure 1.2. The scenes were all equally (im)plausible, such that reasoning over world knowledge would not help participants to identify the correct interpretation.

We used the single-trial method to avoid inadvertently training participants in these unusual structures (Snyder, 2000). Single-trial methods have been previously employed and validated (von der Malsburg et al., 2018). By allowing participants unbounded time to respond and by keeping the sentence on the screen for the duration of the trial, we reduced the possibility

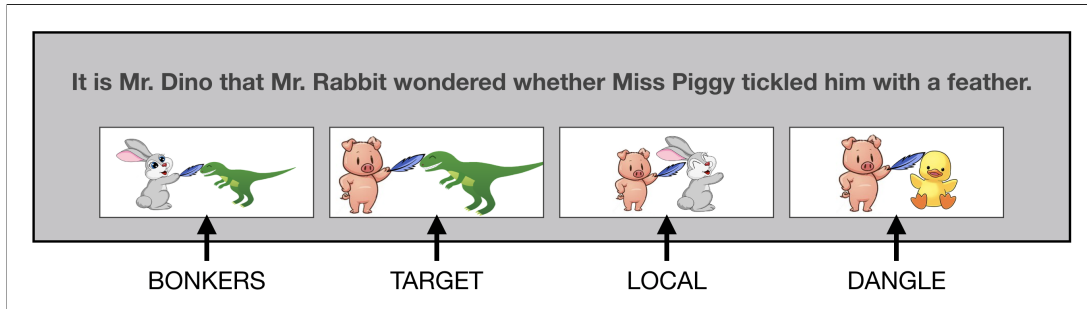


Figure 1.2. Sample display from Experiment 1. The trial shown here is a *resumptive pronoun, weak island* condition. Participants were instructed to read the sentence and click the scene which reflected their interpretation of the sentence. The four response options – *target*, *local*, *dangle*, and *bonkers* (labels not shown to participants) – appeared in random order.

that our results reflect differential forgetting due to varying burdens on working memory between conditions.

1.2.1 Method

Participants.

We paid 300 workers from Amazon’s Mechanical Turk workforce \$0.10 (USD) each for participation. Requirements included that participants learned English before they were 6 years old and that they had not previously participated in the experiment. Subjects were randomly assigned to conditions, such that we collected 50 observations per cell. No participants were excluded.

Factors.

We manipulated two factors resulting in a fully crossed 2×3 design. The first factor, RESUMPTION, had two levels: *gap* or *resumptive pronoun*. The second factor, ISLANDHOOD, had three levels: *non-island*, *weak island*, and *strong island*.

Materials.

The single item set is given in Table 1.1. Each sentence began with an animal character (the head noun; “Mr. Dino”) as the head of a relative clause (or more specifically, a cleft). The

sentence ended with a gap (“tickled _”) or resumptive pronoun (“tickled him”) in direct object position, followed by a prepositional phrase (“with a feather”). The gap/resumptive pronoun appeared in either a non-island, a weak island, or a strong island. Thus, gaps and resumptive pronouns each appeared in environments where participants often hear them and in environments where participants rarely hear them. Each participant read one of the sentences in Table 1.1, and had to match one of the pictures shown in Figure 1.2.

Table 1.1. Experiment 1 stimuli. Sentences appeared in a 2×3 design. The three-level ISLAND-HOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.

Clause Type	Stimulus
Non-island	It is Mr. Dino that Mr. Rabbit said that Miss Piggy tickled (him) with a feather.
Weak Island	It is Mr. Dino that Mr. Rabbit wondered whether Miss Piggy tickled (him) with a feather.
Strong Island	It is Mr. Dino that Mr. Rabbit slept while Miss Piggy tickled (him) with a feather.

Note. Because Experiment 1 was a single-item experiment, Table 1.1 gives all stimuli used in the experiment, not just a representative item set.

The images reflecting different possible interpretations of the sentences were the same for each participant, although the linear positions of the images were randomized. We coded each image according to the type of interpretation it reflected (codes shown in Figure 1.2). The image of the pig tickling the dinosaur is the *target* image, because it reflects an interpretation where the gap or resumptive pronoun refers to its head noun (in this case, the dinosaur). Our best guess about the most likely alternative was that the pronoun would be interpreted as referring to the only other gender- and number-congruent animal in the sentence: the rabbit. We therefore included an image of the pig tickling the rabbit, which we call the *local* interpretation because the rabbit is the closest potential referent. Such an interpretation may reflect a parse favoring local coherence (Tabor et al., 2004). That is, participants may simply disregard the first few words (“It is Mr. Dino that...”) so as to render a clearly grammatical and easy to interpret string (“Mr.

Rabbit said that Miss Piggy tickled him with a feather.”). The final two images were included to ensure that participants were paying attention and not selecting responses at random. The image of the pig tickling the duck reflected a *dangle* interpretation, where the gap or resumptive pronoun refers to a non-sentential referent (the duck was not mentioned in the sentence). Finally, we called the image of the rabbit tickling the dinosaur the *bonkers* interpretation because there should be no ambiguity as to which character was the subject/agent of the verb “tickle.”

Procedure.

Subjects read the instructions, requirements for participation, informed consent, and compensation information on the Mechanical Turk interface. Instructions stated: “You will be presented with a sentence and four pictures. One picture depicts the scene described in the sentence. Your task is to click on the image that matches the sentence. Participation takes about 1 minute.” They then followed a link to the experiment, where they saw one of our six stimulus sentences above four images, as in Figure 1.2, and clicked on the image that corresponded to their interpretation of the sentence. No feedback was given.

1.2.2 Analysis

In this and subsequent analyses, we analyzed responses using Bayesian mixed effects models.³ In contrast to null-hypothesis testing, which dichotomizes all-or-nothing into significant or not, the Bayesian framework enables us to quantify the strength of evidence in a graded fashion. Bayesian inference provides us with estimates of the parameters of interest (e.g., effect of resumption on antecedent choice as a function of sentence type) along with credible intervals indicating the range of plausible values. In addition, the Bayesian approach allows us to directly quantify the probability that the effect of a manipulation is greater or smaller than zero (denoted as $P(\beta > 0)$). If this probability was close to one or zero, we took this as evidence that there was

³All analyses and data can be found on OSF at <<osf.io/9WHN6>>.

a reliable effect.⁴

Bayesian mixed models were fit using the R package *brms* (Bürkner, 2017) which uses the Stan system to obtain posterior distributions (Carpenter et al., 2017). Within Stan, the NUTS sampler (Hoffman and Gelman, 2014) was used to sample from the posterior distributions of the model parameters. We ran four chains each collecting 4000 samples of which the first 1000 were used for warm-up and then discarded. The Gelman-Rubin criterion was used to assess proper mixing of the chains (Gelman and Rubin, 1992).

In all analyses, dependency type was coded using a sum contrast with 0.5 for resumption and -0.5 for the gap condition. As a result, the parameter estimate for dependency indicates the expected increase in the dependent variable when a resumptive pronoun is shown instead of a gap. Clause type was coded using a treatment contrast with non-islands as the base-level and strong islands and weak islands as treatments. The estimates for the main effects therefore indicate the expected effects in the non-island condition and the interactions of clause type with other main effects how the effects differed in strong and weak islands.

In all analyses, unless mentioned otherwise, trials in which *dangle* or *bonkers* responses were selected were excluded. The dependent variable therefore represented whether the response was target (coded as 1) or local (coded as 0); accordingly, a binomial link function was used.

1.2.3 Results

Results are summarized in Table 3.3. In non-island conditions, there was evidence that resumptive pronouns significantly decreased *target* responses and increased *local* responses (see Fig. 1.3). There was some weak evidence that weak islands elicit fewer *target* interpretations than non-islands irrespective of RESUMPTION. Numerically, the effect of resumption was a bit

⁴Note that this quantity should not be interpreted as a Bayesian p-value. It quantifies the probability of the parameter being positive vs. negative, not the probability of the parameter being positive (or negative) vs. the null hypothesis. However, if the probability of a parameter being positive (or negative) is close to one, it is usually safe to assume that the null hypothesis is not the best explanation of the data.

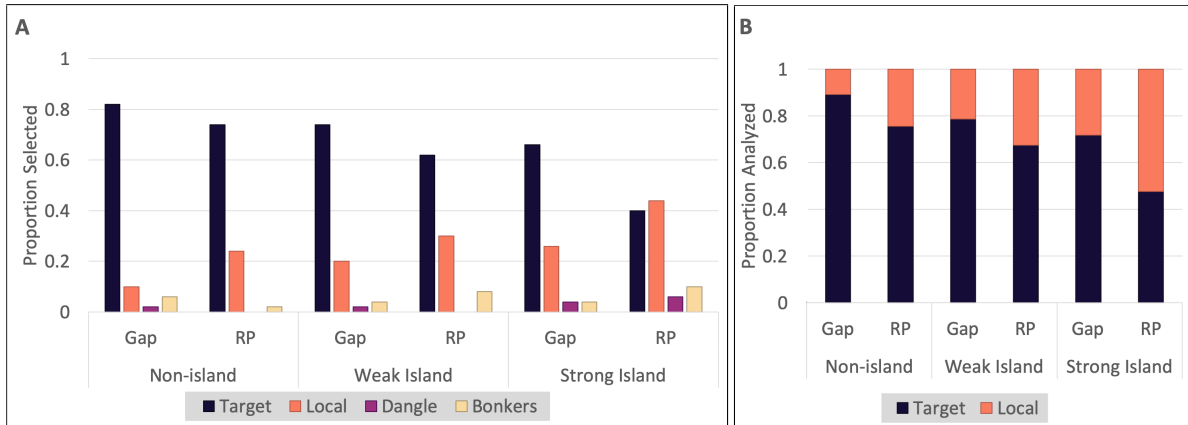


Figure 1.3. Experiment 1 results: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis

smaller for weak islands, but this interaction was also not statistically reliable.⁵ Strong islands elicited fewer *target* interpretations irrespective of RESUMPTION. Numerically, resumptive pronouns reduced *target* interpretations even more for strong islands than for non-islands, but this difference was not statistically reliable.

Table 1.2. Experiment 1 results.

	$\hat{\beta}$	95%-CrI	$P(\beta < 0)$
Intercept (GAP, NONISALND)	1.5	[0.99, 2]	< 0.01
RESUMPTION	-0.75	[-1.6, 0.05]	0.97
ISLANDHOOD:WEAK	-0.44	[-1.1, 0.2]	0.91
ISLANDHOOD:STRONG	-1	[-1.7, -0.39]	> 0.99
RESUMPTION × ISLANDHOOD:WEAK	0.12	[-0.97, 1.2]	0.41
RESUMPTION × ISLANDHOOD:STRONG	-0.25	[-1.3, 0.81]	0.68

Note. Here, $\hat{\beta}$ is the posterior mean and indicates the best estimate of the effect, 95%-CrI indicates the 95% percentile credible interval, i.e., the range of the central 95% of the posterior distribution, and $P(\beta < 0)$ indicates the probability (conditional on the model assumptions) that the true parameter is below zero. $P(\beta < 0) = 0.97$ means that there is a 97% chance that the true parameter is lower than zero, i.e. there is a 97% chance that the effect is negative-going.

⁵The effect of resumption looks larger for weak islands than for non-islands in the plot which shows percentages, but it is smaller in terms of log-odds.

1.2.4 Discussion

Contrary to the prediction of the Facilitation Hypothesis, resumptive pronouns did not lead to more accurate interpretation of sentences than gaps. In fact, they decreased accuracy in interpretation and increased the frequency with which comprehenders selected locally coherent but globally infelicitous interpretations. This was true in non-island conditions, which is perhaps not surprising given that resumptive pronouns are rarely produced in these contexts. But it was also true in island conditions, where resumptive pronouns are produced more often and where both theoretical and experimental work indicated that resumptive pronouns would have a facilitatory effect.

This hindrance effect, which we will refer to as the *resumptive pronoun penalty*, calls into question the interpretation of Hofmeister and Norcliffe's (2013) reading time advantage and Beltrama and Xiang's (2016) subjective comprehensibility rating boost. If resumptive pronouns increase the speed with which comprehenders settle on an interpretation or increase comprehenders' confidence in their interpretation but decrease the likelihood of that interpretation being correct, then they do not in fact facilitate comprehension. Indeed, faster reading times may reflect a number of underlying processes, such as giving up on parsing a confusing string of words.

Experiment 2 addressed a number of follow-up questions. First, we tested whether the resumptive pronoun penalty would extend to data collected in a different paradigm. We also aimed to replicate Hofmeister and Norcliffe's (2013) reading time advantage for resumptive pronouns, and to ask whether this pattern would hold in island conditions as well as the non-island structures they tested. Finally, we tested whether a within-subjects design might stand a better chance of detecting any processing facilitation associated with resumptive pronouns. We addressed these questions by creating 48 item sets manipulated within-subjects and collecting online reading time data as well as offline interpretation data.

1.3 Experiment 2: Self-Paced Reading

Experiment 2 was a self-paced reading task. On each trial, participants pressed a button to read sentences word-by-word and then responded to the multiple choice question, “Who did what to whom?” This experiment was designed to be a partial replication both of Experiment 1 and of Hofmeister and Norcliffe’s (2013) self-paced reading experiment where they found that words after a resumptive pronoun were read faster than words after a gap. Based on Hofmeister and Norcliffe’s (2013) results, we predicted that the words immediately after the resumptive pronoun — that is, the *spillover region* of the resumptive pronoun — will be read faster than that of a gap. If participants are at least as accurate in their interpretations of sentences with resumptive pronouns as with gaps, then faster reading times after resumptive pronouns may be evidence in support of the Facilitation Hypothesis. However, if resumptive pronouns lead to fewer correct interpretations relative to gaps, as they did in Experiment 1, we will take this as evidence against the Facilitation Hypothesis.

1.3.1 Method

Participants.

We paid 96 subjects from Amazon’s Mechanical Turk workforce \$8.00 each for participation. Requirements were that participants learned English and no other language before they were 6 years old and that they had not previously participated in this experiment or Experiment 1. Five participants were excluded: two because their mean accuracy on unambiguous filler trials was below 40%, one for having participated twice (only the data from the second session were excluded), and two for reporting having learned another language before the age of 6. No exclusions were made on the basis of data collected during critical trials.

Factors.

We manipulated the same factors as in Experiment 1: RESUMPTION (*gap* or *resumptive pronoun*) and ISLANDHOOD (*non-island*, *weak island*, *strong island*). This resulted in a fully-

crossed 2×3 design.

Materials.

We created 48 item sets, an example of which is given in Table 1.3. Aside from our experimental manipulations, every sentence was structurally identical. Each began with a clefted animal character and ended with a gap or resumptive pronoun in direct object position followed by a prepositional phrase introducing an instrument. Characters in the sentence were pseudo-randomly drawn from a pool of eight animal characters such that each character appeared in each argument position a roughly equal number of times across items and such that the filler (“Miss Piggy” in Table 1.3) and the middle subject (“Miss Cat” in Table 1.3) were always one gender, while the lowest subject (“Mr. Dog” in Table 1.3) was always the other gender. Other elements that varied across critical items included the tense of the root clause (half of the critical items began, “It is. . .,” and the other half, “It was. . .”); the gender of the head noun and resumptive pronoun, where applicable (half were feminine and half masculine); and subordinator (half used “that” and half “who”). All logical possible combinations of these features appeared a roughly equal number of times across items. All clause boundaries contained an overt subordinator (“that/why/while” in Table 1.3).

Table 1.3. Experiment 2 stimuli. Sentences appeared in a 2×3 design. The three-level ISLAND-HOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line.

Clause Type	Sample stimulus
Non-island	It was Miss Piggy that Miss Cat reported that Mr. Dog poked (her) with a pencil.
Weak Island	It was Miss Piggy that Miss Cat understood why Mr. Dog poked (her) with a pencil.
Strong Island	It was Miss Piggy that Miss Cat snacked while Mr. Dog poked (her) with a pencil.

After every sentence was read word-by-word, the question “Who did what to whom?” appeared on the screen with four response options (Table 1.4). Responses were systematically

created in the same manner as in Experiment 1, except in this experiment they were presented as sentences and not images. The four options always included a *target* interpretation (where the gap or resumptive pronoun refers to the head noun), a *local* interpretation (where the gap or resumptive pronoun refers to the most local gender-agreeing noun, i.e., the middle subject), a *dangle* interpretation (where the gap or resumptive pronoun has an extra-sentential referent), and a *bonkers* interpretation (where the gap or resumptive pronoun is correctly interpreted as the head noun but the subject of the low verb is wrong).

Table 1.4. Response options for Experiment 2: Options listed here correspond to the item set given in Table 1.3 and were the same for all six conditions

Label	Sample response options
Target	Mr. Dog poked Miss Piggy with a pencil.
Local	Mr. Dog poked Miss Cat with a pencil.
Dangle	Mr. Dog poked Miss Rabbit with a pencil.
Bonkers	Miss Cat poked Miss Piggy with a pencil.

We also included 60 fillers that were specifically designed to deter subjects from developing heuristic parsing or response strategies (e.g., ‘the first animal in the sentence is always the one that pronoun refers to’). These consisted of five sets of twelve items. Examples of each of these types and the types appear in Table 1.5, and the types of parsing/response strategies they were meant to prevent are summarized in Table 1.6. We designed the fillers to match the critical items in several ways so that the critical items would not stand out to participants. Each filler type included two embedded clauses which were comprised of roughly equal numbers of non-islands, weak islands, and strong islands. All fillers began a *wh*-dependency with a clefted animal character and the dependency ended with a gap in the middle clause. Note that because there were no gaps in the lowest clause (i.e., inside the island), there were no island violations in the fillers.

We controlled for many of the same features of these stimuli as with the critical items. Similarly, the pattern of genders across the characters within a given filler type remained the

Table 1.5. Experiment 2 fillers.

Filler label	Sample stimulus
Type A	<p>It is Mr. Dino who Mr. Bear asked _ if Miss Cat immobilized Miss Rabbit with a rope.</p> <p>Who did what to whom?</p> <ul style="list-style-type: none">a) Miss Cat immobilized Miss Rabbit with a rope. (✓)b) Mr. Bear immobilized Miss Rabbit with a rope.c) Mr. Bear immobilized Miss Duckie with a rope.d) Miss Cat immobilized Miss Duckie with a rope.
Type B	<p>It was Miss Rabbit that Mr. Bear told _ that he jabbed Miss Piggy with a fork.</p> <p>Who did what to whom?</p> <ul style="list-style-type: none">a) Mr. Bear jabbed Miss Piggy with a fork. (✓)b) Mr. Bear jabbed Miss Cat with a fork.c) Miss Duckie jabbed Miss Piggy with a fork.d) Miss Duckie jabbed Miss Cat with a fork.
Type C	<p>It is Miss Rabbit who _ informed Mr. Dino that she whacked Mr. Dog with a bottle.</p> <p>Who did what to whom?</p> <ul style="list-style-type: none">a) Miss Rabbit whacked Mr. Dog with a bottle. (✓)b) Miss Rabbit whacked Mr. Froggy with a bottle.c) Mr. Dino whacked Mr. Dog with a bottle.d) Mr. Dino whacked Mr. Froggy with a bottle.
Type D	<p>It was Mister Bear who _ saw Miss Duckie when he hit Mr. Froggy with a rock.</p> <p>Who did what to whom?</p> <ul style="list-style-type: none">a) Mr. Dog hit Mr. Froggy with a rock. (✓)b) Mr. Dog hit Miss Piggy with a rock.c) Miss Duckie hit Mr. Froggy with a rock.d) Miss Duckie hit Miss Piggy with a rock.
Type E	<p>It was Miss Cat that _ understood why Mr. Dino swore that Mr. Froggy cleaned him with a loofa.</p> <p>Who did what to whom?</p> <ul style="list-style-type: none">a) Mr. Froggy cleaned Mr. Dino with a loofa. (✓)b) Mr. Froggy cleaned Ms. Cat with a loofa.c) Miss Cat cleaned Mr. Dino with a loofa.d) Miss Cat cleaned Mr. Dog with a loofa.

Note. Gaps (underscores) were not shown to participants.

Table 1.6. Summary of the possible comprehension heuristics that the fillers were designed to prevent. The specific filler types that were designed to prevent each strategy are listed in the right hand column.

Concern to be addressed	Solution
<p>All pronouns in the critical stimuli are resumptive pronouns, which means they all refer to the first animal character in the sentence (i.e. the head noun). Participants may develop a strategy of interpreting all pronouns as referring to the first animal, which would allow participants to correctly interpret resumptive pronouns without parsing them.</p>	<p>We designed fillers that contain pronouns with referents in various positions so as to diversify the types of pronouns in the experiment. Some referred to the subject of middle clause (Types B & E), others to the high subject (Type C), and some even to extra-sentential referents (Type D).</p>
<p>Similarly, if all pronouns in accusative case (i.e., <i>her/him</i>, as opposed to <i>she/he</i>) are resumptive pronouns, participants may come to rely on case cues to compensate for any difficulties with parsing resumptive pronouns.</p>	<p>Pronouns in Type E fillers have accusative case and are not resumptive pronouns.</p>
<p>If all pronouns in the study have referents within the sentence, participants may learn to disregard the dangle multiple choice option for critical items.</p>	<p>The answer choices for Type D fillers require participants to settle on an extra-sentential referent for a pronoun.</p>
<p>Participants may develop task-specific parsing strategies if they only ever encounter gaps in one position.</p>	<p>Fillers contain gaps in various positions: direct object of middle clause (Types A & B), subject of middle clause (Types C, D, & E), in addition to gaps in the critical items which are in the lowest object position.</p>
<p>Some fillers should have embedding verbs that take clausal complements so that the critical items do not stand out in this regard. (Filler Types A, B, & C take two complements each: a NP followed by a clause.)</p>	<p>Filler Type E contains embedding verbs which take single clausal complements.</p>

same. For instance, every filler of Type B had three characters: feminine, masculine, feminine (in that order) or masculine, feminine, masculine (in that order). Within each filler type, half of the twelve items began with a feminine character and half with a masculine character. Also as with critical items, within each type of filler root clause tense was half present, half past; subordinators were “that” for half of the stimuli and “who” for the other half; and all clause boundaries were marked with an overt subordinator. All logical possible combinations of these features appeared a roughly equal number of times across each filler type.

Procedure.

Subjects read the requirements for participation, a brief summary of the task, and compensation information on the Mechanical Turk interface. They then followed a link to the experiment, hosted on Ibex Farm (Drummond, 2013), where they completed an informed consent form, language background and demographics questionnaire, and read task instructions: “In this experiment, you will read about 100 sentences. After each sentence, you will answer a comprehension question. Sentences will be presented to you one word at a time. To go on to the next word, press the spacebar. Please read carefully and do your best to select the correct response. Some sentences will be difficult, so don’t worry if you aren’t sure. Go with your best guess.”

After three practice trials, participants saw a progress bar appear and began the actual experiment. Each trial began with a row of dashes and spaces. Participants pressed the spacebar to reveal the first word, and then pressed the spacebar again to replace the first word with dashes and reveal the next word. This continued until every word in the sentence had been revealed once, at which point all words and dashes representing the stimulus sentence disappeared and the multiple choice question and response options appeared. No feedback was given and the next trial began after a 500 ms pause.

1.3.2 Analysis

For Experiments 2 and 3, where the experimental design entailed repeated measurements (subjects and/or items), all population-level parameters were also allowed to vary on the group-level (maximal random-effects structure; see Barr et al., 2013). To avoid overfitting, all parameters received regularizing priors (Matuschek et al., 2017; Nicenboim and Vasishth, 2016). For population-level predictors (i.e. fixed effects), we used normal priors with $\mu = 0$ and $\sigma = 1$. Group-level parameters (i.e. random effects) were modeled in terms of a correlation matrix and a vector of standard deviations. For the standard deviations we used half-normal priors with $\mu = 0$ and $\sigma = 0.5$. For the correlation matrix an LKJ prior with $\eta = 8$ was used such that smaller correlations among group-level parameters were favored over values closer to the extremes (-1 and 1) without preventing the inference of high correlations if there was evidence for that in the data. All other details of the analysis were the same as in Experiment 1.

1.3.3 Results

Filler items all included one unambiguously correct interpretation among the four multiple choice options. Overall, accuracy on these trials was high (Figure 1.4), indicating that participants performed the task as intended. A notable exception is in Type D fillers, which were answered correctly only 32.8% of the time (a conservative estimate of chance for these trials would be 25%). Type D fillers were designed so as to require participants to establish a referent for the pronoun that was not present in the sentence – a *dangle* interpretation. This indicates that participants felt a strong pressure to choose a referent for the pronoun from among the characters mentioned in the sentence, a point which we will return to in the General Discussion.

Multiple choice data from critical trials (Figure 1.5) were analyzed as in Experiment 1, except that we now included crossed random effects for subjects and items. The overall pattern was similar to Experiment 1, thus also serving as a successful replication.

Multiple choice interpretation results are summarized in Table 1.7. In non-islands,

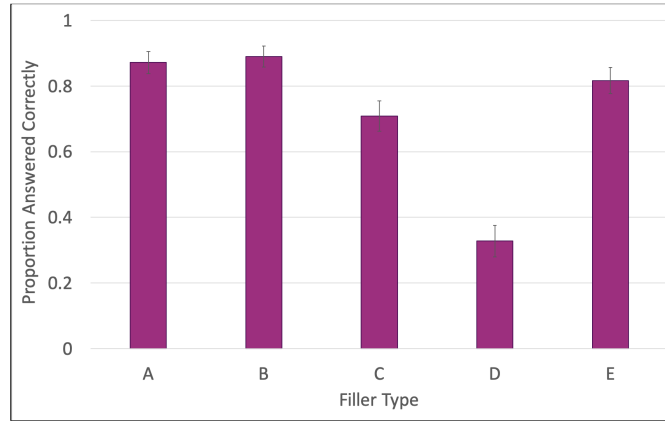


Figure 1.4. Experiment 2 results: Accuracy on filler trials, by filler type (see Table 1.5)

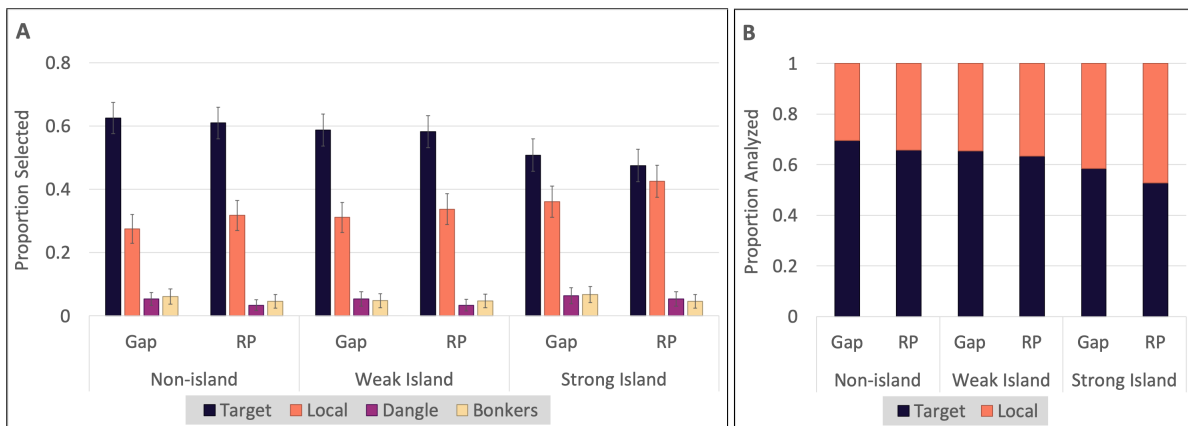


Figure 1.5. Experiment 2 results: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis

resumptive pronouns elicited fewer target responses and more local responses. Weak islands elicited slightly fewer target responses than non-islands and the reduction in target responses due to resumption was numerically smaller than in non-islands. Strong islands elicited fewer target responses than non-islands overall and the reduction in target responses due to resumptive pronouns was numerically bigger than in non-islands.

Table 1.7. Experiment 2 results: Multiple choice interpretation responses.

	$\hat{\beta}$	95%-CrI	$P(\beta < 0)$
Intercept	0.82	[0.59, 1]	< 0.01
RESUMPTION	-0.19	[-0.44, 0.053]	0.94
ISLANDHOOD:WEAK	-0.14	[-0.34, 0.06]	0.92
ISLANDHOOD:STRONG	-0.58	[-0.81, -0.36]	> 0.99
RESUMPTION \times ISLANDHOOD:WEAK	0.071	[-0.29, 0.43]	0.35
RESUMPTION \times ISLANDHOOD:STRONG	-0.089	[-0.43, 0.26]	0.7

For the reading time data (Figure 1.6), we followed Hofmeister and Norcliffe (2013) in defining the critical region as the second word after the gap or resumptive pronoun. In our stimuli, this was always the determiner of the instrument (e.g., the “a” in “with a pencil” in Table 1.3). Prior to the reading time analysis, we excluded 9 trials where this word was read in more than 5000 ms and an additional 9 trials in which any word in the sentence was read in less than 100 ms (likely indicating that a participant accidentally skipped a word by holding the space bar down for too long). A total of 4590 trials were included in the analysis.

Reading times were analyzed on the reciprocal scale (Kliegl et al., 2010; Wu et al., 2018; Baayen and Milin, 2010). The dependent variable therefore was *reading speed* (measured in words per second) at the second word after the gap/resumptive pronoun which always was a determiner. The distribution of the residuals was assumed to be Gaussian, which was confirmed using posterior predictive checks.⁶

⁶It is common to analyze the logarithm of self-paced reading times. However, examining the residuals usually suggests that this violates the distributional assumptions of Gaussian linear models, and this can in turn distort the results (Kliegl et al., 2010). We therefore analyzed reciprocal reading times, which also has the benefit of providing parameter estimates that are transparently interpretable in terms of reading speed.

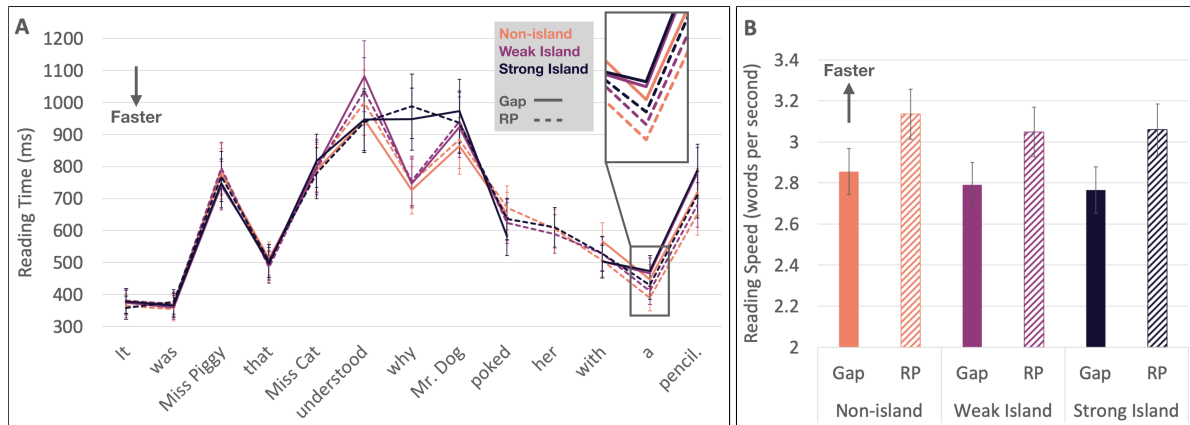


Figure 1.6. Experiment 2 self-paced reading results. (A) All reading times (zoom box on the critical region). The word regions are shown with the sample weak island stimulus sentence from Table 1.3. The corresponding non-island sentence would have “reported” and “that” in place of “understood” and “why.” The corresponding strong island sentence would have “snacked” and “while.” (B) Reading speeds (i.e., the DV in our statistical model) at the critical region.

The prior for the intercept was a normal distribution with $\mu = 3$ and $\sigma = 0.5$. This captures the idea that the average reading speed is likely between 2 and 4 words per second. Priors for the other fixed effects normals with $\mu = 0$ and $\sigma = 0.5$. The prior on the residuals was a normal with $\mu = 0$ and $\sigma = 1$. The other priors were the same as before. This means that if the average reading speed was 3 words per second, most of the data would be between 1 and 5 words per second (i.e. the reading times would be between 200ms and 1000ms). Note that these priors can and will be overruled by the data if necessary.

Results of the reading speed analysis appear in Table 1.8. Reading speed in the non-island condition (intercept of the model) was estimated to be 3 words per second. There was strong evidence suggesting that resumptive pronouns increased reading speed in non-islands. Reading speeds in weak islands were slower than in non-islands, but there was no evidence suggesting that the effect of resumption was different. Similarly, reading speed was overall slower in strong islands than in non-islands, and again there was no evidence suggesting that the effect of resumption was different.

Table 1.8. Experiment 2 results: Reading speed.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept	3	[2.8, 3.1]	> 0.99
RESUMPTION	0.28	[0.17, 0.38]	> 0.99
ISLANDHOOD:WEAK	-0.08	[-0.14, -0.01]	0.01
ISLANDHOOD:STRONG	-0.08	[-0.15, -0.02]	< 0.01
RESUMPTION \times ISLANDHOOD:WEAK	-0.02	[-0.15, -0.02]	0.4
RESUMPTION \times ISLANDHOOD:STRONG	0.02	[-0.12, 0.15]	0.59

Note. A positive coefficient means more button presses per second, i.e. faster reading. For instance a coefficient of 1 would mean participants read one additional word per second. Accordingly, the final column of probabilities is now the probability that β is *greater* than 0, not *less* than as in Tables 3.3 and 1.7.

1.3.4 Discussion

The interpretation data from Experiment 2 again indicate that resumptive pronouns lead to less accurate offline interpretation than gaps. Specifically, resumptive pronouns led readers to select fewer target responses and more locally coherent (but incorrect) responses than gaps in both non-islands and islands. These data thus replicate the finding of Experiment 1, mitigating concerns related to the single-item, between-subjects nature of that study.

The reading time data in Experiment 2 replicated Hofmeister and Norcliffe’s (2013) finding that in non-islands, readers read the words after a resumptive pronoun faster than they read the words after a gap. There was no evidence that this effect was different in islands. Taking less time to perform the same process undoubtedly constitutes more efficient processing. However, participants in this study did not perform the same processes when they read a gap as when they read a resumptive pronoun. If they had, resumptive pronouns would have shown the same pattern of interpretation as gaps, as evidenced by the fact that resumptive pronouns led to more incorrect interpretations of the wh-dependency than gaps. It is therefore impossible to conclude from these data that resumptive pronouns constitute an improvement relative to gaps from the point of view of the comprehender.

As mentioned in the Introduction, reading times alone do not offer insight into the

underlying mechanisms involved in sentence comprehension. It is possible for different processes to result in the same pattern of reading times. Indeed, this may be trivially true for any dependent measure, indicating a need for more multi-paradigm studies like the one we present here. Even when a difference is detected (as in the faster reading times following resumptive pronouns relative to gaps), it is not usually possible to attribute it to any specific process in the absence of other data: Faster reading times could not only indicate ease of comprehension, as is most often assumed, but also that readers abandon a parse (e.g., Nicenboim et al., 2016).

We therefore ran Experiment 3, a visual world experiment where we tracked comprehenders' eyes while they listened to auditory stimuli. We measured which animal characters they looked at and when they looked at them while they comprehended sentences with gaps and resumptive pronouns in order to better understand the online processing data.

1.4 Experiment 3: Visual World Eyetracking

Experiment 3 was a visual world paradigm in which we investigated how gaps and resumptive pronouns are processed online. We used the same stimulus sentences from Experiment 2 (with minor modifications to accommodate the paradigm; see below), but presented sentences auditorily to subjects through headphones while they looked at four animal characters in the corners of a monitor. As in Experiments 1 and 2, we asked participants how they interpreted the sentence at the end of each trial. Response options were identical to those in Experiment 2. If the resumptive pronoun penalty we saw in the previous two experiments is independent of modality (i.e., whether participants read or heard the sentences), then the multiple choice data should again show fewer target interpretations in resumptive pronoun conditions than in gap conditions.

In visual world comprehension studies, the comprehender's gaze indicates the focus of attention, which in turn is mechanistically driven by comprehension processes.⁷ Thus, if

⁷Gaze can therefore serve as a reliable indicator of whether and when a given referent is being processed (Altmann and Kamide, 2004; Altmann, 2004; Huettig et al., 2011; Huettig and Altmann, 2005; Altmann and

we want to know how processing differs when parsing a gap dependency versus a resumptive pronoun dependency, we can compare looks to potential referents while subjects listen to gaps and resumptive pronouns. (Altmann and Kamide, 2004; Altmann, 2004; Huettig et al., 2011; Huettig and Altmann, 2005; Altmann and Kamide, 2009; Altmann and Mirković, 2009)

To determine whether resumptive pronouns facilitate online comprehension relative to gaps, we compare how accurate referent identification is when processing gaps and resumptive pronouns. If comprehenders' gazes more accurately pick out the target interpretations after resumptive pronouns than after gaps, then this would constitute evidence in support of the Facilitation Hypothesis in online processing.

Given the findings of Experiments 1 and 2, however, we predict that the more likely outcome will be the opposite. We have suggested that comprehenders are simply confused by resumptive pronouns. In self-paced reading data, this would be reflected in decreased reading times as readers try to end the trial more quickly. In visual world data, the confusion account predicts that resumptive pronouns will lead to less accurate referent identification than gaps in the gaze data. Specifically, we predict that comprehenders' looks after hearing a resumptive pronouns will approach chance between the two plausible referents of the pronoun (both the *target* and the *local* referents).

1.4.1 Method

Participants.

We ran subjects from the UC San Diego undergraduate population until we reached our target of 96 participants who met a priori criteria for being included in analyses.

These criteria included: (1) across all trials, participants looked more to the first noun

Kamide, 2009). When listening to a sentence that begins, "It was Miss Piggy who Miss Cat," attention will first be drawn to *Miss Piggy* while semantic information is accessed and bound to syntactic information (e.g., direct object; head noun of a cleft construction). During this process, the eyes will be drawn to the pig. Soon thereafter, processing of *Miss Cat* will begin, and the eyes will be drawn away from the pig to the cat. How exactly looking at a referent supports processing is still a matter of some debate (Huettig and Altmann, 2005; Huettig et al., 2011; Altmann, 2004; Altmann and Mirković, 2009).

while hearing the first noun than they looked to any other character (1 exclusion); (2) that participants stayed awake for the duration of the experiment (3 exclusions); (3) that participants' responses to multiple choice interpretation questions on unambiguous fillers exceed 80% accuracy (6 exclusions); (4) that each participant provided at least one trial's worth of eye-tracking data during the region of interest (gap/resumptive pronoun and spillover) in each of the 6 cells of the experiment (i.e. the eye-tracker detected their eye and they were not looking at the fixation cross or away from the screen during this portion of every critical item; 2 exclusions); and (5) that the experimental computer and the eye-tracker ran smoothly and without the need for frequent recalibration (as reported by experimenters; 41 exclusions). No exclusions were made on the basis of behaviors contingent on the experimental manipulations.

Subjects received course credit for participation. Pre-screen requirements were that participants were over 18 years old, that they learned English and no other language before they were 7 years old, and that they had normal or corrected-to-normal vision.

Factors.

We manipulated the same factors as in Experiments 1 and 2: RESUMPTION (*gap* or *resumptive pronoun*) and ISLANDHOOD (*non-island*, *weak island*, *strong island*). This resulted in a fully-crossed 2×3 design.

Materials.

Critical stimuli for Experiment 3 were the same as those in Experiment 2 except for two modifications. First, instead of using "that" as the first subordinator for half of the stimuli, we exclusively used "who" in Experiment 3 (to facilitate the stimulus recording). Second, so as to control for duration of the auditory stimuli across islandhood conditions, we changed several embedding verbs so that within an item set all embedding verbs had the same number of syllables. Thus, where the sample item for Experiment 2 (Table 3) had "that" immediately after the head noun and contained the embedding verbs, "reported, understood, snacked," the same item in

Experiment 3 (Table 1.9), had “who” after the head noun and used embedding verbs “reported, understood, exercised,” each of which has three syllables. Response options (Table 1.10) were identical to those in Experiment 2.

Table 1.9. Written versions of the auditory Experiment 3 stimuli. Sentences appeared in a 2×3 design. The three-level ISLANDHOOD manipulation is shown across rows, while the two-level RESUMPTION manipulation is shown in-line. Stimuli were almost identical to those in Experiment 2, save for minor changes related to creating controlled auditory recordings.

Clause Type	Sample stimulus
Non-island	It was Miss Piggy who Miss Cat reported that Mr. Dog poked (her) with a pencil.
Weak island	It was Miss Piggy who Miss Cat understood why Mr. Dog poked (her) with a pencil.
Strong Island	It was Miss Piggy who Miss Cat exercised while Mr. Dog poked (her) with a pencil.

Table 1.10. Response options for Experiment 3. For every trial, after the audio recording of the stimulus sentence was played, the comprehension question “Who did what to whom?” was displayed on the screen, along with these response options. The response options correspond to the item set given in Table 1.9 and were the same for all six conditions. They were created in the same formulaic way as in Experiments 1 and 2; so ‘target’ refers to the interpretation where the resumptive pronoun (correctly) refers to the head noun, etc.

Label	Sample response options
Target	Mr. Dog poked Miss Piggy with a pencil.
Local	Mr. Dog poked Miss Cat with a pencil.
Dangle	Mr. Dog poked Miss Rabbit with a pencil.
Bonkers	Miss Cat poked Miss Piggy with a pencil.









Stimuli for Experiment 3 were recorded by a native speaker of American English. Critical stimuli were spliced such that within a given item, all of the lexical content that remained constant across conditions was acoustically identical (i.e., “It was Miss Piggy who Miss Cat,” “Mr. Dog poked,” and “with a pencil” for the item in Table 1.9). Content that varied across conditions was manipulated using Audacity and Praat so as to have the same durations. For embedding verbs and subordinators (i.e., “reported that,” “understood why,” and “exercised while”), this was achieved by using the *Lengthen* function in Praat to stretch/compress each clip to the mean duration of

the original three clips for that item. In cases where this resulted in one or more recordings sounding clearly artificially manipulated, the two-word clips were re-recorded and the process was repeated until recordings were judged by undergraduate RAs to sound like unaltered speech. For gap conditions, silence was spliced in where a resumptive pronoun otherwise appeared such that the duration between the offset of the lowest verb and the lowest preposition was identical in all conditions. To mitigate the oddness of silence in this position, as well as to eliminate coarticulation effects that might make spliced in material sound unnatural, the speaker who recorded the materials produced pauses between words throughout the recording while attempting to approximate normal prosody. The resulting sentences were spoken slowly, with single-syllable words like “who” and “him” averaging 414 ms in duration (including the brief periods of silence mentioned above).⁸

Fillers were identical to those in Experiment 2. In order not to render fillers more or less unnatural sounding than critical items, the recordings were created in a similar way. The speaker produced pauses throughout, and undergraduate RAs used Praat to swap strings with identical lexical content from recordings of other filler (e.g., “It was Miss Cat who”).

Eight animal characters with distinguishing colors, features, and gender-typical clothing, colors, and accessories were digitally drawn (Table 1.11).

Table 1.11. Animal characters for Experiment 3

Mr. Bear	Miss Cat	Mr. Dino	Mr. Dog	Miss Duckie	Mr. Froggy	Miss Piggy	Miss Rabbit
							

⁸Auditory recordings are available on OSF at <<osf.io/9WHN6>>.

Procedure.

We began each experimental session by familiarizing subjects with the animal characters. The experimenter started by showing the participant each animal character printed out on laminated cards one at a time, saying the character's name and instructing the subject to try to remember what the character looked like because they would be asked questions about its physical appearance. Once all eight characters had been introduced, they were placed face down on a table in front of the subject. On the back of each card was a label specifying what animal appeared on the reverse, for instance, "Duck" or "Dinosaur." In a first round of questions, the experimenter asked 16 questions which required the subject to recall the name of each character twice, as in "Who was wearing a yellow bowtie?" If the subject responded incorrectly, or did not correctly produce the name of the character (e.g., "frog" or "Mr. Frog" instead of "Mr. Froggy"), the subject was corrected and shown the image of the intended character. The experimenter then shuffled a different deck of 16 cards with harder questions. These questions required subjects to recall features of each character's physical appearance, as in "What color was Miss Duckie's umbrella?" Subjects were shown the correct character's picture when they responded incorrectly. Incorrectly answered question cards were set aside and repeated after the experimenter had exhausted the deck. This process was repeated until all questions had been answered correctly once.

The subject was then seated at a tower-mounted EyeLink 1000 eyetracker and asked to put on headphones. After calibration, the subject read instructions on the screen: "In this experiment you will listen to about 100 sentences through headphones while we track your eye movements. We will show you pictures of some of the characters in the sentence. After each sentence, you will answer a comprehension question. To answer comprehension questions, press the number associated with your response. Some sentences will be difficult, so don't worry if you aren't sure. Go with your best guess." They then proceeded to three practice trials, after which they were instructed to let the experimenter know if they had questions. If not, they went

on to 108 experimental trials (48 critical, 60 filler), each separated by a brief fixation check. Experimenters monitored calibration for accuracy throughout the experiment and paused to recalibrate as necessary.

On each trial, four images of characters appeared in the corners of the screen 500 ms before the onset of the audio recording. On critical trials, the images depicted the three characters in the sentence and the extra-sentential referent of the dangle multiple choice response option (i.e., Miss Rabbit for the item in Table 1.9). For filler trials, all characters mentioned in the sentence (up to four) were present on the screen; for stimuli with only three characters, a character from a multiple choice response option for that item was selected to appear on the screen. The position of characters on the screen was pseudo-randomized such that each interpretation option was equally represented in each corner of the screen across every experimental session (e.g., each participant saw the head noun in each corner an equal number of times).

A 200 ms pause occurred at the end of the audio recording after which the four animal characters disappeared from the screen and the question, “Who did what to whom?” appeared in the center of the screen followed by the four response options. Participants pushed the number corresponding to their selection on the keyboard, and then the trial concluded.

1.4.2 Analysis

The analysis of eye-movements in visual world experiments is an open problem and various approaches have been used in the literature, from growth-curve analyses, to permutation tests, and logistic regression with polynomial predictors. Our analysis loosely follows the approach described by Barr (2008). Barr’s approach is to use multilevel logistic regression to model the proportion of looks to some target (versus other locations on the screen) as a function of time (starting from the onset of the critical word) and the manipulated factors. In a preprocessing step, Barr bins eye-tracking data temporally in order to reduce redundancy and computational complexity such that the dependent variable is the proportion of time spent looking at the target in a given bin. To account for repeated measuring, Barr includes random

effects for subjects and items in the model.

This approach has two shortcomings which led us to slightly deviate: First, the dependent variable is a proportion that is 0 or 1 (gaze on target or not) most of the time but not always. This property requires the use of empirical logit approximations which have no solid grounding in statistical theory and which are known to introduce biases (Gart et al., 1985; Donnelly and Verkuilen, 2017). To address this, we analyzed not the proportion of time on target in a time bin but the position of the gaze (on target or not) in 100 ms intervals (non-averaged). Second, the approach by Barr does not adequately capture within-trial auto-correlation. It assumes that the bins within a trial are statistically independent, which due to the relative slowness of eye movements is clearly not the case. To capture this auto-correlation, we include by-trial random effects (Barr et al., 2013), which in the Bayesian framework are easy to fit.

The model had the predictors: dependency, clause type, time of the measurement (centered on zero) and all interactions of these factors. The time window of the analysis began 200 ms after the offset of the word preceding the gap/resumptive pronoun and ended 1000 ms later (i.e. 1200 ms after gap/resumptive pronoun onset). In the interest of parsimony, and in keeping with Barr (2008), the effect of time was assumed to be linear, which is probably a good approximation given the small time window. In this analysis, we only included trials where subjects looked at either the target or the image corresponding to the local noun.

1.4.3 Results

Multiple choice responses were analyzed as in Experiment 2. Data are shown in Figure 1.7. Results, summarized in Table 1.12, were largely the same as in Experiments 1 and 2. There was strong evidence that resumptive pronouns reduced target responses in non-islands, compared to gaps (the *resumptive pronoun penalty*). Weak islands elicited about as many target responses as non-islands. Numerically, the resumptive pronoun penalty was reduced (i.e., resumption imposed less of a penalty) in weak islands compared to in non-islands, but there was not enough evidence to conclude that this was a statistically reliable effect. In strong islands, target responses

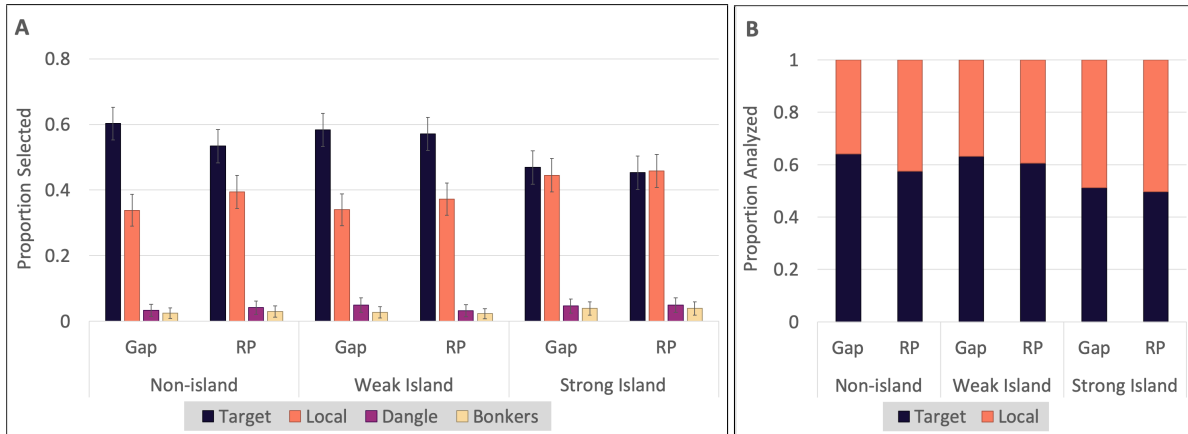


Figure 1.7. Experiment 3 results of the multiple choice (interpretation) question: (A) all responses and (B) just *target* and *local* responses — i.e., those included in the analysis

were also reduced across the board. Interestingly, there was some evidence that resumption did not reduce target responses as much in strong islands as in non-islands.

Table 1.12. Experiment 3 results: Multiple choice interpretation responses.

	$\hat{\beta}$	95%-CrI	$P(\beta < 0)$
Intercept	0.49	[0.29, 0.69]	< 0.01
RESUMPTION	-0.31	[-0.54, -0.07]	> 0.99
ISLANDHOOD:WEAK	0.07	[-0.11, 0.24]	0.24
ISLANDHOOD:STRONG	-0.48	[-0.66, -0.3]	> 0.99
RESUMPTION \times ISLANDHOOD:WEAK	0.17	[-0.15, 0.49]	0.15
RESUMPTION \times ISLANDHOOD:STRONG	0.24	[-0.08, 0.57]	0.07

The gaze data are shown in Figure 1.8 (collapsed across ISLANDHOOD) and Figure 1.9 (the gap/resumptive pronoun region broken down in all six conditions). In general, these data were particularly clean. For instance, the left panel of Figure 1.8 shows that around 500 ms post-onset of the head noun (“Miss Piggy” in the example sentence), participants’ eyes were drawn to the picture of the head noun. The close correspondence between gap and resumptive pronoun conditions indicates that 96 participants was enough to ensure high signal-to-noise ratio. Indeed, one can also see nuanced effects, such as the early drop-off in looks to the gender-incongruent low subject (“Mr. Dog”) after the onset of either of the first two animal characters.

Descriptively, the gaze data from the gap/resumptive pronoun region can be characterized

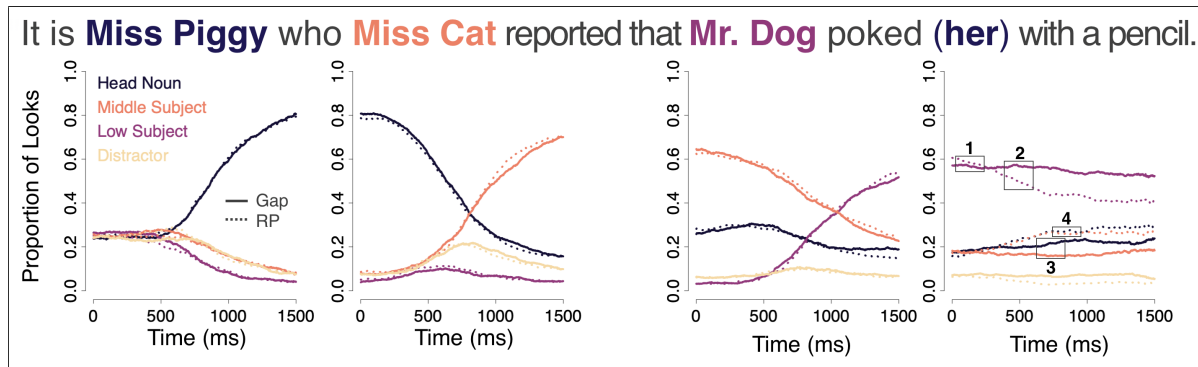


Figure 1.8. Experiment 3 gaze data, collapsed across ISLANDHOOD, from the onset of the head noun (“Miss Piggy” in the example sentence), the middle subject (“Miss Cat”), the low subject (“Mr. Dog”), and the gap/resumptive pronoun. The four descriptive points we outline in the text are numbered in the right-most plot: (1) participants’ gazes remained on Mr. Dog (i.e. the most recently named character) at the onset of the gap/resumptive pronoun; (2) resumptive pronouns resulted in more looks away from Mr. Dog than gaps; (3) gaps led to more looks to the target than the local interpretation; (4) resumptive pronouns led to roughly equal numbers of looks to the target and local interpretations. The latter two points taken together mean that, although resumptive pronouns led to numerically more target looks than gaps, these looks were less accurate. That is, when participants did look away from Mr. Dog, they were more likely to look at Miss Piggy than Miss Cat when hearing a gap than when hearing a resumptive pronoun.

by four observations, labeled in Figure 1.8. First, at the onset of the gap/resumptive pronoun, participants’ eyes remained on the most recently named character in the sentence (i.e. the low subject, “Mr. Dog” in the example). Second, relative to gaps, resumptive pronouns appear to have resulted in more looks away from the low subject (Mr. Dog). Third, after the onset of a gap, the proportion of looks to the head noun (Miss Piggy/the *target* interpretation) increased, but the proportion of looks to the middle subject (Miss Cat/the *local* interpretation) did not appear to change. Fourth, similar to gaps, after the onset of a resumptive pronoun, the proportion of looks to the head noun (Miss Piggy/*target*) increased, but in contrast to gaps, the proportion of looks to the middle subject (Miss Cat/*local*) also increased. Comprehenders who looked away from the low subject (Mr. Dog) while hearing a resumptive pronoun were more likely to look to the target interpretation than when hearing a gap, but they were also more likely to look to the local interpretation than when hearing the gap. In fact, they appeared to look to the target referent and the local referent with roughly the same frequency when hearing a

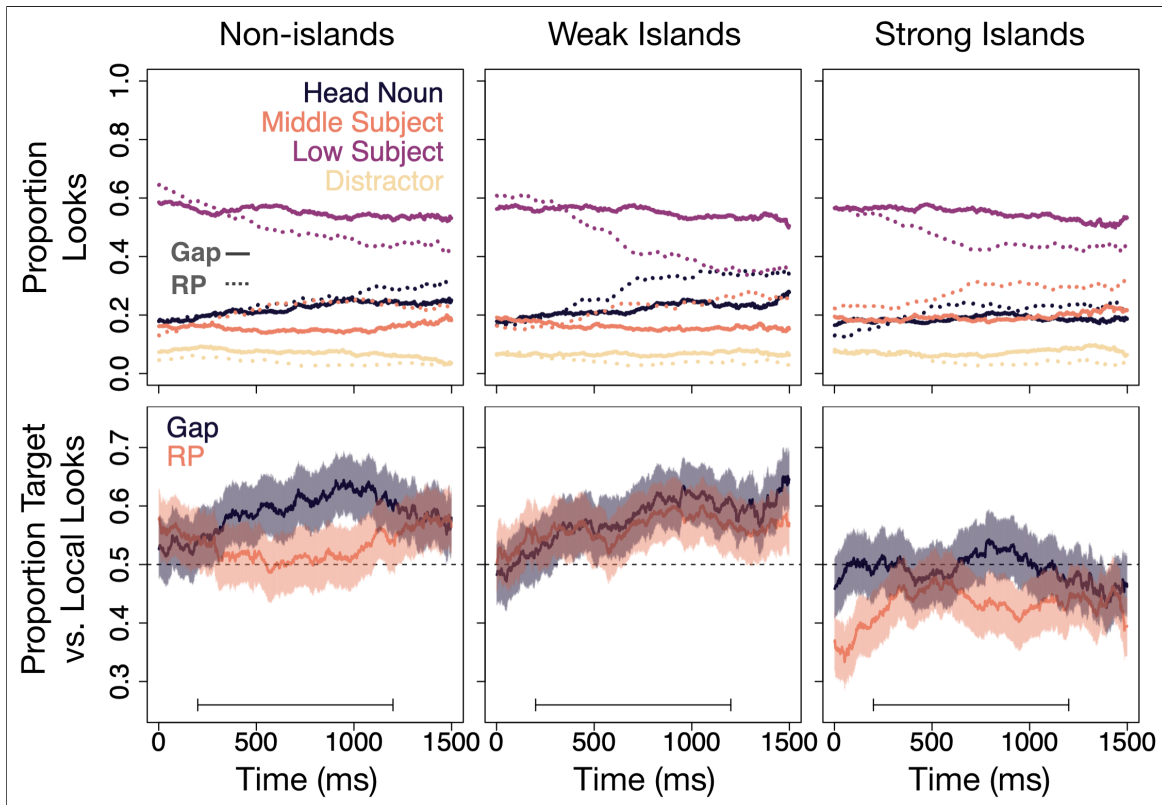


Figure 1.9. Gaze data during the gap or resumptive pronoun for all three island types (columns) from Experiment 3. *Top row:* looks to all four characters on the screen. *Bottom row:* looks to the target interpretation, excluding data where participants were not looking at the target or local interpretation (i.e., the dependent variable in our analysis). Shaded area shows standard error. Chance looking between target and local interpretations is 50% (dashed line). Analysis window (200 to 1200 ms) is indicated with horizontal bar in the bottom of each plot.

resumptive pronoun, suggesting that resumptive pronouns may be fully ambiguous between these two potential referents in online processing.

Results of the eyetracking analysis are presented in Table 1.13. There was some weak evidence suggesting that looks to the target increased over time in the non-island conditions when collapsing across gap and resumptive pronoun conditions. Similar to the pattern we have observed in multiple choice interpretation data, there was strong evidence that resumptive pronouns reduced looks to the target in non-islands. Compared to non-islands, weak islands elicited more looks to the target overall, but there was no evidence suggesting that the resumptive pronoun penalty was different in weak islands than in non-islands. Also similar to what we have

observed in multiple choice interpretation data, compared to non-islands, there were fewer looks to the target in strong islands, and again no evidence that the resumptive pronoun penalty was any different for strong islands than for non-islands. There was also no evidence for any of the other two-way interactions nor the three-way interaction.

Table 1.13. Experiment 3 results: Gaze.

	$\hat{\beta}$	95%-CrI	$P(\beta < 0)$
Intercept	1.1	[0.27, 1.9]	< 0.01
TIME	0.28	[-0.12, 0.69]	0.08
RESUMPTION	-1.1	[-2.1, -0.18]	> 0.99
ISLANDHOOD:WEAK	0.83	[-0.041, 1.7]	0.03
ISLANDHOOD:STRONG	-1.9	[-2.8, -1.1]	> 0.99
TIME \times RESUMPTION	-0.24	[-0.72, 0.65]	0.53
TIME \times ISLANDHOOD:WEAK	0.16	[-0.42, 0.74]	0.71
TIME \times ISLANDHOOD:STRONG	-0.12	[-0.67, 0.44]	0.67
RESUMPTION \times ISLANDHOOD:WEAK	0.22	[-1, 1.5]	0.37
RESUMPTION \times ISLANDHOOD:STRONG	-0.27	[-1.6, 1]	0.66
TIME \times ISLANDHOOD:WEAK	-0.59	[-1.5, 0.38]	0.88
TIME \times ISLANDHOOD:STRONG	0.13	[-0.82, 1.1]	0.39

1.4.4 Discussion

In the multiple choice interpretation task in Experiment 3, resumptive pronouns reduced the number of target interpretations compared to gaps across the board (i.e., in all three levels of ISLANDHOOD). There was some evidence that this effect was attenuated in strong islands. However, even if this attenuation is real, it was not big enough to counteract the resumptive pronoun penalty. Resumptive pronouns still led to numerically fewer target responses than gaps in strong islands, meaning that the attenuation does not constitute evidence for facilitation.

This attenuation is another indicator that the effect of resumption is one of increasing confusion such that performance approaches chance. If the increase in local responses for resumptive pronouns reflected a locality preference, then we might expect to see cases in which resumptive pronouns result in more local responses than target responses. Instead, across

experiments, the highest rates of local responses for resumptive pronouns (i.e., strong island conditions) are cases where participants select target and local responses at roughly the same rate. This suggests that in these cases, participants are selecting from among the a priori plausible responses (*target* and *local*) at chance.

Overall, the multiple choice data are consistent with Experiments 1 and 2. As the stimuli in this experiment were presented auditorily, we can further conclude that the resumptive pronoun penalty is independent of the modality of stimulus presentation (similar to Clemens et al.'s 2012 finding that auditory presentation does not change their acceptability).

The gaze data largely replicated the multiple choice interpretation data from this experiment, as well as Experiments 1 and 2. Critically, gaps resulted in more accurate looking behavior than resumptive pronouns. Even though the overall number of looks away from the low subject was less for gaps than resumptive pronouns (see point (2) in Figure 1.8), when participants did look away, they looked more to the target referent than the local one when they heard a gap than they did when they heard a resumptive pronoun (points (3) and (4) in Figure 1.8). Online as well as offline, then, resumptive pronouns hinder comprehension relative to gaps.

We were particularly surprised by our second descriptive observation (point (2) in Figure 1.8) – that resumptive pronouns induced the comprehender to look away from the low subject more than gaps.⁹ We speculated that the difference may be attributable to the different ways that comprehenders identify the referents of gaps and pronouns. Pronouns trigger a search for a referent (e.g., Hobbs, 1978; Kaiser et al., 2009), reflected in looks away from the low subject, which cannot be the referent because it has the wrong gender and is not reflexive (*Principle B* of Government and Binding Theory; Chomsky, 1993). Gaps, on the other hand, can be anticipated because, once a head noun is encountered, the parser can infer that it is in an open wh-dependency. Indeed, Frazier (1987) showed that gaps are actively predicted during sentence comprehension.

⁹A supplemental analysis at 700 ms post-onset showed that resumptive pronouns reliably reduced looks to the lower subject in the non-island condition ($\hat{\beta} = -0.29$, 95%-CrI: [-0.52, -0.053], $P(\beta < 0) > 0.99$) and that this effect was even larger in weak islands ($\hat{\beta} = -0.39$, 95%-CrI: [-0.73, -0.046], $P(\beta < 0) = 0.99$) and strong islands ($\hat{\beta} = -0.34$, 95%-CrI: [-0.67, 0.0069], $P(\beta < 0) = 0.97$).

When the parser finally comes across the gap, the referent of the gap is already known because gaps are syntactically bound to the head noun. No search is needed.

This account straightforwardly predicts our third observation, that gaps lead to more looks to the head noun (*target*) than the middle subject (*local*), because the referent of the gap is a priori known to be the head noun. Across conditions, when a participant looks away from the low subject, their eyes will move to what they believe to be a likely antecedent. In gap conditions, this is unambiguously the head noun, and indeed looks to the head noun are more common than looks to the local referent when hearing a gap.

In resumptive pronoun conditions, on the other hand, participants' eye movements are at chance between landing on plausible referents of a pronoun: the target and local interpretations. This seems to indicate that comprehenders consider both the target and the local interpretations as plausible antecedents, as they would be for an ordinary pronoun. Perhaps this pattern of data arises because resumptive pronouns simply *are* ordinary pronouns from the perspective of the comprehender.

If resumptive pronouns are in fact ordinary pronouns, then the interpretation of ordinary and resumptive pronouns should pattern together, to the exclusion of gaps. Specifically, ordinary pronouns should display the same preference for local resolution that we have seen for resumptive pronouns. Experiment 4 was designed to test this prediction.

1.5 Experiment 4: Ordinary Pronoun Comprehension

Experiment 4 used a single-item sentence comprehension task to test the hypothesis that resumptive pronouns are in fact ordinary pronouns from the perspective of the parser. If so, this would be an indicator that there is no grammatical representation of a filler-resumptive pronoun dependency. Resumptive pronouns would be ordinary pronouns from the perspective of the comprehension system.

This would be consistent with production models of English resumption, according to

It was Mister Bear that asked Mister Dog why Miss Duckie reported him to the boss.

1. Miss Duckie reported Mister Frog.
2. Mister Bear reported Miss Duckie.
3. Miss Duckie reported Mister Dog.
4. Miss Duckie reported Mister Bear.

Figure 1.10. Screenshot from Experiment 4. The trial shown here is from the *ordinary pronoun* condition. Participants were instructed to read the sentence at the top and select the interpretation “that is most likely to be true based on the sentence.”

which resumptive pronouns are simply ordinary pronouns from the perspective of the production system (Asudeh, 2004, 2011; Morgan and Wagers, 2018). They are the result of the producer giving up on completing a filler-gap dependency and producing a pronoun where a gap would have otherwise appeared had things not gone awry.

If resumptive pronouns are ordinary pronouns from the perspective of the comprehension system, it predicts that ordinary pronouns will show a similar pattern of interpretation to resumptive pronouns. Specifically, ordinary pronouns should show a local preference for resolution as compared to gaps, as we have observed for resumptive pronouns in the previous three experiments. Participants read a sentence with either a gap, an ordinary pronoun, or a resumptive pronoun and then selected a multiple choice option reflecting their interpretation. Both the sentence and the multiple choice options remained on the screen for the entire trial (see Figure 1.10), as in Experiment 1. Participants were allowed as much time as they needed to respond.

1.5.1 Method

Participants.

We continuously ran workers from Amazon’s Mechanical Turk workforce until we reached our target of 150 participants who met a priori criteria for being included in the analysis.

A total of 174 participants were run. We excluded 12 for incorrectly answering the comprehension question in the filler trial preceding the critical trial; 10 for reporting that they learned another language before the age of 7; and 2 for responding to either the filler trial or the critical trial in less than 5 seconds (both of these participants responded in less than 2 seconds and gave incorrect answers on the filler trial). Participants were paid \$0.35 each for participation. Pre-screen requirements included that participants learned English before they were 7 years old and that they had not previously participated in the experiment. Each subject was assigned a different condition from the previous subject, such that we collected 50 observations per cell.

Factors.

We included one factor, REFERRING ELEMENT, which had three levels: *gap*, *resumptive pronoun*, and *ordinary pronoun*.

Materials.

We created one item set, given in Table 1.14, along with four multiple choice interpretation options, Table 1.15, which were the same for each of the three stimulus sentences. In order to compare a gap or resumptive pronoun to an ordinary pronoun, we had to make a significant structural change to the sentence. In the ordinary pronoun condition, when the reader reaches the pronoun, the wh-dependency is already resolved; the pronoun must therefore be interpreted as an ordinary pronoun.

In the gap and resumptive pronoun conditions, on the other hand, there must be an unresolved wh-dependency when the reader reaches the gap/pronoun so that the pronoun is interpreted as resumptive. To do this, we introduced a new argument position by using the verb *ask* to embed the lowest clause inside a weak island. Unlike the previous verbs that we used in weak island stimuli (e.g., *wonder whether*, *understand why*, *consider whether*) *ask* allows an optional direct object before its clausal complement, as in “ask (someone) whether...” In the ordinary pronoun condition, the dependency terminates in this position with a gap, so the

subsequent pronoun is unambiguously an ordinary pronoun.

In the gap and resumptive pronoun conditions, we filled this direct object position of *ask* with the first person pronoun *I*, thereby keeping the dependency open. The choice of *I* was deemed optimal because it cannot be a referent for the resumptive pronoun as the two have different person features (*him* cannot be used to refer to *I*). It also added less of a working memory burden relative to the ordinary pronoun condition than names or full noun phrases would have (Lewis, 1996). Thus, across conditions, the referring element (i.e. the gap, resumptive pronoun, or ordinary pronoun) had the same syntactic role (direct object), thematic role (patient), and semantic role (the character who was reported to the boss).

Table 1.14. Stimuli for Experiment 4.

Clause Type	Stimulus
Gap	It was Mister Bear that I asked Mister Dog why Miss Duckie reported _ to the boss.
Resumptive pronoun	It was Mister Bear that I asked Mister Dog why Miss Duckie reported him to the boss.
Ordinary pronoun	It was Mister Bear that _ asked Mister Dog why Miss Duckie reported him to the boss.

Note. Because Experiment 4 was a single-item experiment, Table 1.14 gives all stimuli used in the experiment, not just a representative item set.

The multiple choice response options reflecting different possible interpretations of the sentences are given in Table 1.15. They were the same across conditions, although their order was randomized for each participant. Because the referent of the ordinary pronoun is ambiguous by design, there is no “target” interpretation for this condition and we therefore refer to this as the “distant/target” option in this experiment.

Procedure.

Subjects read the requirements for participation and compensation information on the Mechanical Turk interface. They read the instructions and informed consent information on the experiment website, which was hosted on Ibex Farm (Drummond, 2013). Instructions stated: “In

Table 1.15. Response options for Experiment 4. As in previous experiments, the comprehension question was, “Who did what to whom?”

Label	Sample response options
Distant/Target	Miss Duckie reported Mister Bear.
Local	Miss Duckie reported Mister Dog.
Dangle	Miss Duckie reported Mister Frog.
Bonkers	Mister Bear reported Miss Duckie.

this experiment, you will answer comprehension questions about 3 sentences. The whole task should take just a minute or two. When doing the experiment we ask that you stay focused and avoid distractions like multitasking. Please do not listen to music with words. Underneath each sentence there will appear four possible interpretations. Select the one that is most likely to be true based on the sentence.” They then proceeded to a practice trial, followed by a filler trial, followed by the critical trial. No feedback was given.

1.5.2 Results

Data (Figure 1.11) were analyzed using a logistic regression to model the dependent variable indicating whether subject chose the *distant/target* response (= 1) or the *local* response (= 0). The single predictor was REFERRING ELEMENT (gap, resumptive pronoun, ordinary pronoun) coded as a treatment contrast with ordinary pronoun as the reference level. Results are summarized in Table 1.16.

Table 1.16. Experiment 4 results: Multiple choice interpretation responses.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept (ORDINARY PRONOUN)	1.58	[0.95, 2.27]	> 0.99
GAP	0.14	[-0.81, 1.09]	0.61
RESUMPTIVE PRONOUN	-1.01	[-1.87, -0.2]	< 0.01

The results of this analysis showed that, contrary to our prediction, the interpretation of ordinary pronouns patterned with that of gaps, not resumptive pronouns. Specifically, we found no credible evidence that ordinary pronouns are interpreted differently from gaps. We

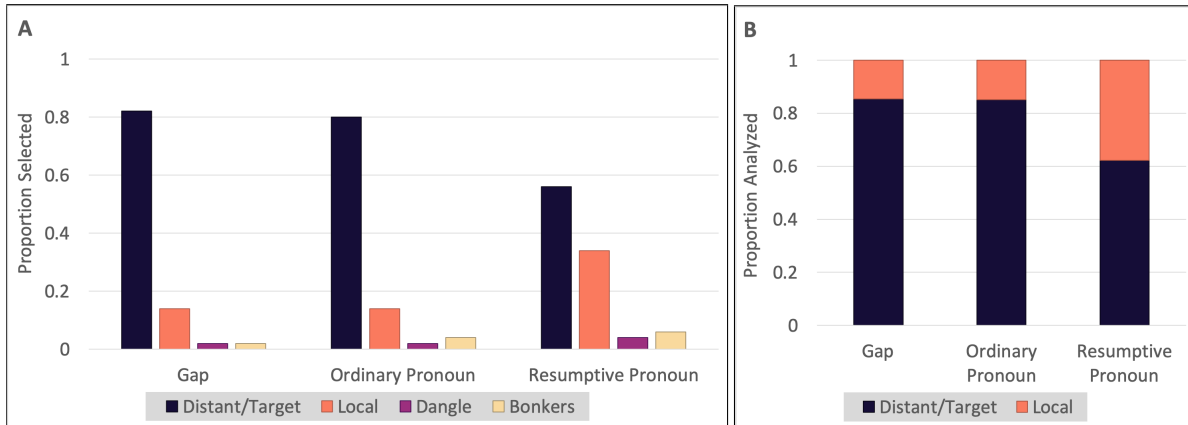


Figure 1.11. Experiment 4 results: (A) all responses and (B) just *distant/target* and *local* responses — i.e., those included in the analysis

did, however, find evidence that ordinary pronouns are processed differently from resumptive pronouns: ordinary pronouns elicited more distant/target responses than resumptive pronouns.

1.5.3 Discussion

Experiment 4 aimed to assess the hypothesis that English resumptive pronouns are in fact ordinary pronouns and not a kind of alternative gap, as is reported for languages like Hebrew and Irish. Specifically, we tested the prediction that the interpretation data for resumptive pronouns and ordinary pronouns would pattern together, to the exclusion of the gap data. Consistent with the first three experiments, resumptive pronouns resulted in decreased distant/target responses relative to gaps and increased local responses. Contrary to our predictions, however, the interpretation of ordinary pronouns patterned with that of gaps, to the exclusion of resumptive pronouns. Thus, resumptive pronouns appear to involve interpretation processes that are distinct from both gaps and ordinary pronouns.

We believe that the most likely explanation for this pattern is that resumptive pronouns are ungrammatical and confuse the comprehender. When forced to select a referent, comprehenders approach chance in selecting between the gender- and number-congruent discourse entities. Two observations from the data are consistent with this hypothesis. First, we have used the term “locality bias” to refer to the fact that resumptive pronouns result in more local interpretations

than gaps. If there were a true bias for local interpretations, however, we might expect to see a condition for which resumptive pronouns lead to more local interpretations than target interpretations.

Across four experiments, this pattern never obtained. Instead, resumptive pronouns seem to level out the rates of target and local interpretations, a pattern that is more consistent with chance performance than a locality bias. Second, the effect of islandhood on gap interpretation may serve as proof of concept. Islands induce a similar effect in the way comprehenders interpreted gap conditions. Where gaps are acceptable (i.e., in non-islands), our data show that they are understood better than where they are moderately unacceptable (weak islands) and severely unacceptable (strong islands). As target interpretations decrease with decreasing acceptability, local interpretations increase.

There is no clear reason that islands might induce local interpretations of gaps; gaps should always unambiguously refer to the filler. A more reasonable interpretation of this pattern is that islands, being ungrammatical, lead to confusion, and comprehenders' performance becomes closer to chance.

A final concern we wished to address regards generalizability. The stimuli we have used in these experiments were constructed using proper names and cleft constructions so as to avoid pragmatic content which might guide interpretation. This allowed us to isolate the contribution of parsing to the overall comprehension of resumptive pronouns. But it also rendered sentences that are subjectively odd and may not behave in the same way as more typical examples of resumption. For instance, it is not clear that English speakers would produce resumptive pronouns in these types of sentences. Indeed, in Irish, a language with grammatical resumption, resumptive pronouns are restricted in cleft constructions (McCloskey, 2011). If resumptive pronouns are not produced in the kinds of sentences we have been testing, then there is no paradox for our stimuli: the (lack of) production and the (lack of) comprehension would in fact be aligned.

Our claim that the Facilitation Hypothesis cannot explain the comprehension-production paradox relies on speakers producing resumptive pronouns in the same kinds of sentences where

resumption hinders comprehension. We therefore ran a final experiment to determine whether English speakers produce resumptive pronouns in these sentences.

1.6 Experiment 5: Production

In Experiment 5, a single-trial sentence production task, we asked whether resumptive pronouns are produced in the same types of sentences where we have shown they hinder comprehension. If we are right in claiming that resumptive pronouns are produced as the result of difficulties in production and not to aid comprehension, then we might expect to see resumptive pronouns produced in these sentences. But if the Facilitation Hypothesis is right, and resumptive pronouns are produced to facilitate comprehension, then speakers should not produce resumptive pronouns in this experiment because resumptive pronouns do not facilitate comprehension in these stimuli.

Following Morgan and Wagers (2018), we asked participants to type into a text box to complete a sentence, the beginning of which was given by a prompt (Figure 1.12). Based on Morgan and Wagers's 2018 findings, we expected to find very few resumptive pronouns in non-islands, more in weak islands, and even more in strong islands.

1.6.1 Method

Participants.

We paid 300 workers from Amazon's Mechanical Turk workforce \$0.15 (USD) each for participation. Requirements included that participants self-identified as native English speakers and that they had not previously participated in the experiment. Subjects were randomly assigned to conditions, such that we collected 100 observations per cell. A total of 372 participants were run; 72 were excluded for participating more than once (all trials from these participants were excluded). No exclusions were made on the basis of the production task responses.

Instructions
 Help us rephrase some sentences by filling in the blank. Not all sentences will have a clear right or wrong answer. Just do your best!

For example:

- **Context:**
Sue went shopping and wound up buying Little Tommy a birthday present.
- **New sentence:**
Sue bought *.(This is just an example; do not respond here.)*

Example Good responses:

- Good response: "Little Tommy a present for his birthday."
- Good response: "a present for Little Tommy's birthday."
- Good response: "Little Tommy a birthday present when she went shopping."

Example bad responses:

- Bad response: "a present."
- Bad response: "something for Tommy."

Task

Context:
Mr. Rabbit slept while Miss Piggy tickled Mr. Dino with a feather.

New sentence:
It is Mr. Dino that Mr. Rabbit slept while Miss Piggy *.*

Figure 1.12. Screenshot of Experiment 5. The trial shown here is the *strong island* condition. Participants were instructed to type in the box to complete the new sentence using the information in the context sentence.

Factors.

We manipulated one factor, ISLANDHOOD, which had three levels: *non-island*, *weak island*, and *strong island*.

Materials.

The single item set, given in Table 1.17, was derived from the Experiment 1 item set (Table 1.1). Target sentences are identical to the critical sentences in Experiment 1. Context sentences were created by removing the first clause (“It is Mr. Dino that”) from the Experiment 1 stimuli and replacing the gap/resumptive pronoun with the head noun, “Mr. Dino.” Prompts were created by removing the final verb, the gap/resumptive pronoun, and prepositional phrase from the Experiment 1 stimuli. Target responses were the same for all conditions, given in the bottom panel of Table 1.17).

Table 1.17. Experiment 5 stimuli and target responses. Participants were instructed to complete the sentence started by the prompt using the information given by the context sentence.

Clause Type	Stimulus
Non-island	
CONTEXT:	Mr. Rabbit said that Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit said that Miss Piggy...
Weak Island	
CONTEXT:	Mr. Rabbit wondered whether Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit wondered whether Miss Piggy...
Strong Island	
CONTEXT:	Mr. Rabbit slept while Miss Piggy tickled Mr. Dino with a feather.
PROMPT:	It is Mr. Dino that Mr. Rabbit slept while Miss Piggy...
Target Responses (all conditions)	
GAP:	tickled __ with a feather
RP:	tickled him with a feather

Note. Because Experiment 5 was a single-item experiment, Table 1.1 gives all stimuli used in the experiment, not just a representative item set.

Procedure.

The whole experiment appeared on the Mechanical Turk interface, including compensation information, requirements for participation, informed consent, instructions, a sample item (not involving gaps or resumptive pronouns) with sample good and bad responses, and the task itself (see Figure 1.12). The instructions and sample trial remained on the screen while participants completed the critical trial. No feedback was given.

1.6.2 Data Coding and Analysis

We coded responses in one of four categories: *gap*, *resumptive pronoun (RP)*, *name*, and *other*. Frequencies and examples of each type of response appear in Table 1.18.

Because our goal was to determine whether English speakers produce resumptive pronouns in the types of sentences we tested in Experiments 1 through 4, we were conservative in deciding what types of responses to code as “gap” or “resumptive pronoun.” Target responses appear in Table 1.17. We allowed some minor divergences from these targets which we thought were unlikely to impact the likelihood of producing a resumptive pronoun. These included differences in tense/aspect/mood (e.g., “would tickle,” “had tickled,” “was tickling”) and changes to the prepositional phrase that appeared after the gap/resumptive pronoun (e.g., if it was altered, as in “using a feather,” or altogether missing). All other changes were coded as “other” and excluded from the analysis. This included uses of a different verb (“teased” instead of “tickled”) and changes to the clause structure (e.g., “with a feather tickled him,” “helped tickle him with a feather,” “used a feather to tickle,” “took a feather and tickled him with it”).

Only trials coded as *gap* or *resumptive pronoun* were included in the analysis; all other trials were excluded leaving a total of 221 trials (84 non-islands, 75 weak islands, and 62 strong islands). The analysis differed from previous analyses in that the type of dependency – gap or resumptive pronoun – was our dependent variable (as opposed to a manipulated independent variable), and dependency type was the sole predictor.

Table 1.18. Experiment 5 coding rubric with frequency of each response type in each of the three conditions and examples.

Code	Frequency			Examples
	Non-	Weak	Strong	
Gap	74	54	27	<i>had tickled; tickled using a feather</i>
RP	10	21	35	<i>tickled him; was tickling him with a feather</i>
Name	5	6	20	<i>tickled Mr. Dino; tickled Mr. Dino with a feather</i>
Other	11	19	18	<i>used a feather to tickle him; was being tickled by a feather; kept Mr. Dino awake by tickling him with a feather; did some tickling with a feather; lightly teased with a bird's plumage; actually decided to tickle Mr. Dino</i>

1.6.3 Results

Data are shown in Figure 1.13 and results are summarized in Table 1.19.

Table 1.19. Experiment 5 results.

	$\hat{\beta}$	95%-CrI	$P(\beta > 0)$
Intercept (NONISALND)	-1.7	[-2.3, -1.2]	< 0.01
ISLANDHOOD:WEAK	0.74	[0.02, 1.5]	0.98
ISLANDHOOD:STRONG	1.9	[1.2, 2.6]	> 0.99

1.6.4 Discussion

Experiment 5 showed that resumptive pronouns are produced in the same sentences where they hinder comprehension. This contradicts the Facilitation Hypothesis, but is consistent with production models and with the idea that resumptive pronouns confuse comprehenders. Our production data are broadly consistent with those of Morgan and Wagers (2018): resumptive pronouns were produced least in the non-island condition, more in the weak island condition, and even more in the strong island condition.

It is worth noting that in the strong island condition, participants produced 56% resumptive pronouns (and 44% gaps). In strong islands in Experiment 1, participants chose the target

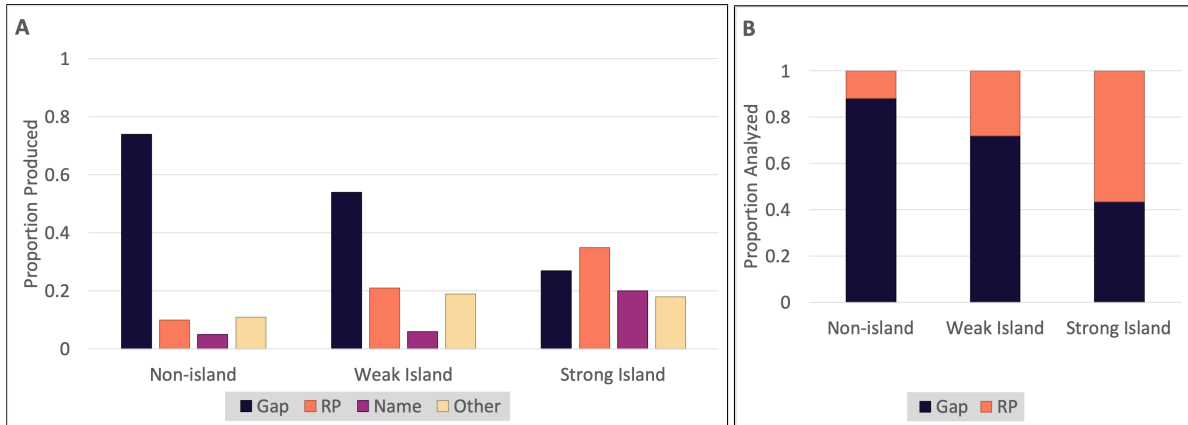


Figure 1.13. Experiment 5 results: (A) all responses and (B) just *gap* and *resumptive pronoun* responses — i.e., those included in the analysis

interpretation of this sentence 72% of the time when it appeared with a gap, but only 48% of the time when it appeared with a gap. Thus, even where resumptive pronouns are produced *more* than gaps, they still hinder comprehension.

1.7 General Discussion

This paper investigated a paradox: English speakers consistently report that resumption is unacceptable, but they nonetheless regularly and reliably produce resumptive pronouns. This tension stands to shed a sliver of light into the black box of language, and in particular, how syntax interfaces with language production and comprehension.

We investigated two different hypotheses put forth in the literature to explain the paradox. The Facilitation Hypothesis views resumption as the result of speakers trying to be helpful to their listeners by providing an explicit pronoun. Production-based accounts, on the other hand, view resumption as the result of processing gone awry during the production of particularly difficult constructions.

These two explanations make different predictions: If resumption is facilitatory for comprehenders, then people should understand sentences with resumptive pronouns better than sentences with gaps. But if resumption is the result of a mishap during production, then sentences

with the ungrammatical resumptive pronoun should be *harder* to understand.

In four comprehension experiments, we found consistent support for the production-based hypotheses. Specifically, instead of increasing the likelihood that comprehenders land on the target interpretation, resumptive pronouns were more likely than gaps to be interpreted as referring to a local distractor. This was true for offline measures as well as online measures, and even in sentences where resumptive pronouns are produced more often than gaps.

In contrast to all previous studies, in our design we carefully avoided providing pragmatic cues that might have led participants to interpret sentences by reasoning over world knowledge, bypassing the effortful task of parsing these particularly difficult structures (see Ferreira et al.'s 2002 work on “good-enough” processing). This allowed us to isolate the contribution of the syntax of resumptive dependencies to the comprehension of these structures. In all our studies, stimuli consisted of sentences describing animal characters interacting in equally implausible ways (e.g., a dog cleaning a duck with a loofa, a cat measuring a dinosaur with a ruler, etc.).

Experiment 1, a sentence-picture matching task, provides the first evidence for the resumptive pronoun penalty. The fact that we found this penalty is especially notable because comprehenders were able to look at the critical sentence and interpretation options simultaneously, and without time pressure for response. Still, accuracy rates were significantly lower for sentences with resumptive pronouns than for those with gaps.

The resumptive pronoun penalty was replicated in Experiment 2, when we increased the number of critical trials, presented sentences in a self-paced reading paradigm, and presented multiple choice interpretation options as short sentences instead of pictures. Here, we replicated Hofmeister and Norcliffe's (2013) finding that the words immediately following a resumptive pronoun were read faster than words immediately following a gap. Hofmeister and Norcliffe (2013) interpreted this decrease in reading time as evidence for facilitation. However, at least in our data, faster reading times cannot be taken to reflect facilitation, because participants' interpretations were less correct in resumptive pronoun conditions than in gap conditions.

Experiment 3, a visual world eyetracking study with auditory stimulus presentation, also

provided evidence for the resumptive pronoun penalty. Again, the interpretation of resumptive pronouns was worse than gaps, both in online (looks to animal characters) and offline measures (multiple choice interpretation questions).

Experiment 4 showed that that resumptive pronouns are not just ordinary pronouns from the perspective of the listener: While resumptive pronouns show a preference for local interpretation relative to gaps, ordinary pronouns do not.

Experiment 5 demonstrated that speakers produce resumptive pronouns in the same kinds of sentences where they hinder comprehension. This finding ensures that the resumptive pronoun penalty can be brought to bear on the comprehension-production paradox.

1.7.1 Chance Performance Supports Production-Based Accounts

The resumptive pronoun penalty is predicted under accounts of resumption as a result of production pressures. On this view, the increased rates of local interpretations do not reflect a preference for local interpretation, but instead reflect comprehenders approaching chance performance among the sensible response options. Our findings are therefore consistent with production-based accounts, but inconsistent with the Facilitation Hypothesis. We do not, however, believe that the present data are able to distinguish between the two production-based accounts we discuss in the Introduction (i.e., Asudeh, 2004, 2011 and Morgan and Wagers, 2018).

It is worth noting that the *dangle* response in our multiple choice options also represented a gender- and number-congruent referent for the pronoun, albeit one which is not mentioned in the sentence. One might then argue that, if resumptive pronouns truly led to chance performance, we should see increases in dangle interpretations as well as local interpretations relative to gap conditions. However, we designed one of our filler types in Experiment 2 and 3, Type D in Table 1.5, such that correctly responding to the interpretation question involved choosing a referent that was not mentioned in the sentence – a *dangle* interpretation. While interpretation of the other fillers showed high accuracy overall (over 80%, see Figure 1.4), participants chose the correct interpretation to Type D fillers only a third of the time. Thus, participants appear to feel

pressure to find an intra-sentential referent for the pronoun. This suggests that for critical trials, the dangle interpretation was probably not an a priori reasonable option.

Confusion and chance performance is generally consistent with what we might predict for any ungrammatical structure: If there is no grammatical representation then it cannot be fully parsed, and comprehenders must rely on extra-grammatical information for interpretation.

Indeed, we see a similar pattern associated with another type of ungrammaticality in our stimuli: islandhood. Across experiments, weak islands resulted in fewer target interpretations and more local interpretations than non-islands, and this effect was even bigger in strong islands. In the absence of pragmatic information, ungrammatical structures like resumptive pronoun dependencies and island violations result in a penalty to interpretation.

1.7.2 Other reasons to doubt the Facilitation Hypothesis

As described in the Introduction, previous experimental work has only tangentially tested the Facilitation Hypothesis, and production-based accounts of resumption suggest that resumptive pronouns may be less likely to help comprehenders than gaps. But these are not the only reasons to doubt the Facilitation Hypothesis. Here we spell out three a priori reasons that this account seems unlikely to be correct.

First, a resolution to the acceptability-production paradox must take one of three forms. It can either (i) do away with acceptability or production (or both) as metrics for grammaticality (see Shlonsky, 1992; Polinsky et al., 2013 for proposals along these lines), (ii) explain why speakers produce resumptive pronouns in spite of their being ungrammatical, or (iii) explain why comprehenders judge resumptive pronouns to be unacceptable in spite of their being grammatical.

The Facilitation Hypothesis, being a hypothesis about comprehension, does not fit any of these three types. It therefore does not fully address the paradox. The implied logic seems to be: When a speaker senses that an utterance will be confusing or difficult to process with a gap, they opt to produce a resumptive pronoun because, even though it is ungrammatical, the listener is more likely to understand the intended meaning. This would be an explanation along the lines of

(ii).

However, research on audience design indicates that speakers do *not* generally take into account the needs of their interlocutors when deciding whether to include optional function words (Ferreira and Swets, 2005; Ferreira and Dell, 2000). For instance, Ferreira and Dell (2000) show that English speakers produce the optional complementizer *that* not when it would facilitate comprehension on the part of their listener, but when they need more time to plan the following word. Often, the production system seems to be selfish, making decisions to facilitate its own goals, not those of the listener. (See Ferreira, 2019 for a review.)

Second, it is surprisingly difficult to intentionally produce ungrammatical sentences, and this difficulty is exacerbated if one tries to systematically produce a complex error, for example, one that spans multiple clauses. This anecdote suggests that it is unlikely that such a mechanism could account for the production of resumptive pronouns in English. From a formal perspective, there is no clear way that standard production models can accommodate the production of ungrammatical forms. Production is taken to be mechanistically guided by the grammar (e.g., Levelt, 1993). For example, in order to put words in the right order to form a relative clause, the system accesses an abstract syntactic representation of relative clauses and then uses the representation to guide lexical selection. But if resumptive pronouns are ungrammatical, then by definition they have no syntactic representation with which to guide production. This is not to say that the production of resumptive pronouns is impossible if they are ungrammatical (Asudeh, 2004, 2011 and Morgan and Wagers (2018) offer mechanistic accounts of how this could work, outlined above), but instead to say that current models of production mechanisms seem to preclude the possibility that speakers produce an ungrammatical string by design, for the sake of the comprehender.

Finally, the intuition that resumptive pronouns are more informative than gaps (perhaps due to their overt gender, number, and animacy cues) is probably misleading. English is a language that, for the most part, requires arguments to be pronounced. (This is unlike many other languages, including Spanish, Japanese, and Malayalam, in which nouns can be grammatically

left out.) Thus, a missing English noun can usually only be a gap, which must refer to the head noun. Resumptive pronouns, on the other hand, look just like ordinary pronouns, and can therefore plausibly refer to any number of potential referents (see 6).

Counterintuitively then, resumptive pronouns are potentially less informative than gaps. Worse yet, if there is no grammatical representation of resumptive pronouns, then there can be no corresponding requirement that the resumptive pronoun refer to the head noun. Whereas comprehenders may infer the intended link between a resumptive pronoun and the head noun, gaps should always lead to more correct interpretation because the grammar explicitly links gaps to the head noun.¹⁰

1.7.3 The relationship between comprehension and production

We started by pointing out that resumption leads to an apparent paradox in English: production and acceptability are generally reliable metrics of grammaticality, but they dissociate in the case of resumptive pronouns. Indeed, the facts about resumptive pronouns led some early resumption researchers to reject the idea that syntax is shared between production and comprehension. For example, Ferreira and Swets (2005) state, “the two systems [production and comprehension] do not consult the exact same database of grammatical rules, as indicated by the finding that the production system allows [resumptive pronouns], but the comprehension system tends to reject them.”

However, two different systems of syntactic representations for production and comprehension would pose a fundamental problem for a variety of behavioral phenomena. There would need to be a – so far unsubstantiated – mapping system linking those two systems, enabling, for

¹⁰Note that many theories hold that islands are ungrammatical, which is to say there is no way to syntactically bind a head noun to a gap in an island. If this is the case, then from the perspective of the parser, gaps and resumptive pronouns are the same: an unbound referential element. (See Phillips, 2006 for a summary of gap processing inside and outside of islands.) Gaps and resumptive pronouns still differ in their gender/number/animacy cues. If this is the only difference, then in islands, resumptive pronouns may in fact be more informative than gaps and may lead to better comprehension. But it is also possible that comprehenders use meta-syntactic knowledge about gaps, which typically may only refer to head nouns. If this is the case, then when there are multiple potential referents with the same gender/number/animacy features, gaps should still be more informative than RPs, regardless of whether they are ungrammatical in islands.

instance, structural priming across modalities (Potter and Lombardi, 1998; Bock et al., 2007; Pickering and Ferreira, 2008) and even dialogue (e.g., Pickering and Garrod, 2004).

But whether syntax is shared is not the only issue at stake. A long-standing question in the field is how much else the two modalities share. In general, models tend to assume that there is a good amount of overlap (e.g., Pickering and Garrod, 2004). Analysis-by-Synthesis theories even argue that comprehension works by synthesizing language to match the input, which is to say that a large part of comprehension *is* production (Halle and Stevens, 1959, 1962; Bever and Poeppel, 2010). If this is true, then production phenomena like resumption should also appear in comprehension. That is, comprehenders should be able to generate (and therefore interpret) strings with resumptive pronouns.

Thus, if production-based accounts of resumption are correct, then resumptive pronouns pose a challenge to theories which hold that comprehension relies on covert production. The paradoxical behavior of resumptive pronouns does not indicate that the two systems do not share grammatical representations, but it may indicate that comprehension and production are more distinct than often thought.

One possible way to rescue an Analysis by Synthesis approach would be if the confusion we see in comprehension resulted from difficulty in mapping between the syntax and semantics of these complex structures. In production, this does not lead to confusion about the message, which is a priori known, but it may explain why speakers produce resumptive pronouns instead of gaps. In comprehension, however, there may be multiple candidate messages that could have prompted the speaker to utter the string of words she did – for example, *Miss Piggy said Mr. Dino tickled Miss Cat* and *Miss Piggy said Mr. Dino tickled Miss Piggy*. The difficulty may not be in covertly producing a structure that matches what the speaker produced, but in maintaining the link between that structure and the specific message that generated it.

Another surprising way in which comprehension and production patterns dissociated in our data is that the resumptive pronoun penalty did not interact with ISLANDHOOD. We predicted that the way comprehenders process resumptive pronouns would be sensitive to syntactic context,

reflecting different histories of experience with resumptive pronouns in different structures.

For instance, in non-islands, where resumptive pronouns are relatively rare (Morgan and Wagers, 2018), it makes sense that comprehenders would struggle to interpret them. In the comprehender's experience, if a speaker had intended to refer to the head noun in this environment, they would have used a gap, but they didn't, so they must have meant something else.

In weak islands, however, where resumptive pronouns are more common (demonstrated in Experiment 5; see also Ferreira and Swets, 2005; Morgan and Wagers, 2018) and rated as more comprehensible (Beltrama and Xiang, 2016), we expected that the resumptive pronoun penalty would be attenuated compared to non-island contexts. Instead, the factors seem to have combined additively such that, independent of the fact that weak islands induced fewer target responses and more local responses, there was no consistent, credible evidence that the resumptive pronoun penalty was different in weak islands from non-islands. The same was true for strong islands.

The fact that comprehenders do not understand resumptive pronouns in islands better again suggests that they are not for the benefit of the comprehender. Producers have access to the intended message, and producing a grammatically licit string has no impact on this. From the perspective of the comprehender, though, when context cues, world knowledge, and pragmatics combine to guide interpretation, they may not always parse complex constructions. When left with nothing but the syntax, they must rely on the parse. Resumptive pronouns, as we have shown, do not provide more helpful information from a syntactic perspective than gaps.

1.7.4 Ways to salvage the Facilitation Hypothesis

Our aim here has been to understand how resumptive pronouns impact comprehension. One feature of our study was that we removed pragmatic information from our stimuli so that any differences in comprehension reflected just the contribution of parsing resumptive pronouns. But it is important to remember that comprehension is not the independent sum of parsing and reasoning. The two may interact in complex ways.

One such way would be discourse context, which we intentionally did not investigate in our studies. For instance, if there had been a discourse context in our experiments involving a robber, Mr. Bear chasing “*him*” with a knife may have made a dangle interpretation more likely than if the discourse context had involved a neighbor (depending on one’s neighbors; Kaiser et al., 2009; Koornneef and Sanders, 2013; Järvikivi et al., 2017; Williams et al., 2018). Given that neither gaps nor pronouns contain much semantic information and the two do not differ in pragmatic content in any clear way, we think it is unlikely that resumption would interact with pragmatic interpretation in such an extreme way. However, if resumptive pronouns do confuse the comprehender, as we have claimed, then it is possible that comprehenders may rely much more heavily on pragmatic information when processing resumptive pronouns, which could conceivably lead to such an interaction. This is a potentially fruitful avenue for future research.

It is also possible that some other feature of our stimuli has obscured the potential benefits of resumption for comprehension. For instance, it is possible that resumptive pronouns facilitate comprehension only in contexts where they would disambiguate between potential referents. In our stimuli, the only plausible referents for the pronoun were the head noun (*target* interpretation) and the middle subject (*local* interpretation), both of which had the same gender. Resumptive pronouns could therefore not disambiguate between the two. Detailed production data are needed to determine whether such an account holds water: If speakers produce resumptive pronouns more when the pronoun disambiguates, then perhaps the Facilitation Hypothesis is on the right track. In this case, our findings would indicate that any facilitation induced by the use of a resumptive pronoun does not come from its impact on parsing, but instead on the comprehender using extragrammatical, pragmatic, resources – namely the number, gender, and animacy cues on the pronoun – to resolve the dependency.

Perhaps the most important caveat of all is that if we wish to test the hypothesis that speakers produce resumptive pronouns to help the listener, then comprehension data – be they reading times, comprehensibility ratings, gazes, or responses to interpretation questions – will never fully suffice. Whether or not a resumptive pronoun helps or hinders comprehension is

irrelevant when the speaker may have false beliefs about resumption's effect on the comprehender. A true test of the Facilitation Hypothesis will therefore require examining whether speakers are more likely to produce a resumptive pronoun when they believe it will help the listener, regardless of how helpful that pronoun may or may not be in actuality.

1.7.5 A new perspective on the competence/performance distinction

The initial perception that resumptive pronouns comprehension and production data are at odds with one another reveals a need for a more nuanced, multilayered framework for understanding performance errors. Based on our results and previous findings, it appears that competence is the same for production and comprehension (i.e., resumptive pronouns are ungrammatical), but that production and comprehension may have their own performance failure modes that produce characteristic errors.

In the case of resumption, competent speakers of English know that it is not grammatical, and yet they produce resumptive pronouns for reasons outlined in the introduction (see Fig. 1.1). Listeners, on the other hand, must work with what is given: an ungrammatical structure containing a pronoun with no obvious referent. There is no grammatical parse available, so the listener falls back on the other tools in the comprehension toolbox. Usually, this includes reasoning over lexical, semantic, prosodic, nonlinguistic and/or contextual cues to resolve referential uncertainty or repair other errors in the signal (Levy, 2008; Park and Levy, 2011). But our experiments took most of these tools away: Apart from gender, the listener had no cues to help solve the puzzle of who the resumptive pronoun may refer to, so guesses approached chance between the two gender-congruent characters in the sentence.

In this scenario, there is a performance error in production and we show that comprehenders stumble. But consider two other cases: 1. The *missing VP effect* (Gibson and Thomas, 1999) is the reverse scenario: Comprehenders find double-center embedded RCs with a missing verb (i.e., ungrammatical) just as acceptable as their (grammatical) counterparts without any missing verbs. However, it has not been documented that speakers produce such sentences. So,

here the apparent mismatch is a consequence of a performance error just in comprehension. 2. The *depth-charge illusion* is a case where both speaker and comprehender make performance mistakes (Paape et al., 2019): The speaker produces a sentence like, “No head injury is too trivial to be ignored,” which is compositionally non-sensical. A second performance error in the comprehender creates the illusion that the sentence is well-formed.

To date, cases such as these have all been treated independently. This underscores the field’s need for a unified theoretical framework for understanding and explaining performance errors. Ideally, such a framework would account for all the cases mentioned here, and would additionally be able to predict and explain other performance errors, which may have yet to be documented.

1.7.6 Resumption: A cautionary tale for language comprehension research

Finally, our findings invite a methodological point by revealing a weakness in processing studies that measure the time course of language processing without also measuring interpretation (Romoli et al. tion). In the case of resumption, previous researchers had assumed that resumptive pronouns lead the comprehender to the same interpretation as a gap. On this assumption, it may have been reasonable to infer that resumptive pronouns are processed more economically than gaps on the basis of a reading time advantage or increased comprehensibility ratings. At least for the stimuli we tested here, however, faster reading times after resumptive pronouns cannot be interpreted as evidence for processing facilitation. We therefore offer the cautionary guideline for comprehension research: measure interpretation, too.

1.8 Conclusion

English speakers reliably produce a structure that they deem unacceptable. This is not only odd, but it poses a serious problem for standard assumptions about language and grammaticality. This acceptability-production paradox has spurred considerable curiosity. One prominent

hypothesis holds that speakers produce resumptive pronouns because they facilitate comprehension (Prince, 1990; Asudeh, 2004; Dickey, 1996; Ariel, 1999; Hofmeister and Norcliffe, 2013; Beltrama and Xiang, 2016; Erteschik-Shir, 1992; Fadlon et al., 2019).

In four comprehension experiments, however, we show that rather than facilitating comprehension, resumptive pronouns lead listeners and readers to approach chance performance in a variety of interpretation tasks. In a production experiment, we demonstrate that this is true even though speakers produce resumptive pronouns in the same sentences where they hinder comprehension. Our findings contradict the Facilitation Hypothesis, but are consistent with production theories (Asudeh, 2004, 2011; Morgan and Wagers, 2018).

We would like to conclude with another puzzle about resumptive pronouns: Languages like Cantonese (Lau, 2016), Hebrew (Ariel, 1999), Irish (McCloskey, 2002), Swedish (Erteschik-Shir, 1992), and Vata (Koopman, 1983) make productive use of gaps, ordinary pronouns, and resumptive pronouns. They do so without the puzzling pattern of acceptability and production data that we have described for English. If resumptive pronouns can be grammatical, then why does English, whose speakers readily produce resumptive pronouns, not simply grammaticize them? That is, if Hebrew speakers produce resumptive pronouns and like the way they sound, then what stops English speakers from doing the same? We suspect that the answer will shed light on something much deeper about human language than just this one quixotic syntactic structure.

1.9 Acknowledgment

Chapter 1 is a reprint of the material submitted to *Cognition*. Morgan, Adam M.; von der Malsburg, Titus; Ferreira, Victor S.; and Wittenberg, Eva. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Learning untrained syntactic structures: Behavioral evidence for abstract representations of abstract representations

2.1 Introduction

Puppies often grow up side-by-side with human infants, living in the same houses, eating off the same floors, and hearing the same language. In spite of this, the outcomes of their cognitive development – particularly with regard to language – are radically different. Within a few years, and with little explicit teaching (or possibly none at all; Ochs and Schieffelin, 1994), children learn tens of thousands of words. Puppies and other non-human animals, however, do not learn more than a few hundred words, with the notable exception of Chaser, a Border Collie who knew 1,022 words (Pilley and Hinzmann, 2013).

Words are not the extent of learning: children also come to know the patterns describing how words combine such that *‘People eat tomatoes’* describes the world, and not the plot of a campy 1970s horror movie. There is little evidence that non-human animals ever come to reliably use *syntax* in this way, although an African gray parrot named Alex who was able to form phrases like “blue peg” and “orange paper” may be an exception (Pepperberg, 1981; Kako, 1999).

But even interpreting Chaser’s and Alex’s abilities generously, Chaser was still only

capable of word learning, and Alex only of the simplest of combinatorics. These pale in comparison to the kinds of linguistic abilities that humans, even very young humans, know. To be sure, human adult grammatical competence includes words and simple grammatical patterns. For example, determiners (e.g., “the”) and nouns (“dog”) combine to form noun phrases (“the dog”), and verbs and noun phrases combine to form verb phrases (“feed the dog”). But human language also includes non-contiguous units, or *long-distance dependencies*.

Take for example the relative clause in “the dog [that I fed __].” “The dog” is the object of “fed,” but unlike in typical verb phrases, it is not adjacent to the verb. In fact, it is not even in the same clause. Speakers of English, however, are effortlessly able to understand that the two form to combine a verb phrase, despite the distance between them and without any explicit knowledge of dependencies or verb phrases. Long-distance dependencies like this come in all kinds of varieties, including questions (“Which movie did you want to see __”), *tough*-constructions (“The news was tough to swallow __”), etc. Furthermore, each of these varieties consists of a family of related structures, as in the following relative clauses:

- | | | |
|------|--|--|
| (11) | a. the dog [that __ bit me] | <i>Subject Relative Clause</i> |
| | b. the dog [that I fed __] | <i>Direct Object Relative Clause</i> |
| | c. the dog [that I threw the ball to __] | <i>Indirect Object Relative Clause</i> |

Each relative clause in (11) is named for the position of the *gap* – the spot where the modified noun, “dog,” would have appeared in an ordinary clause. That is, the subject relative clause (SRC) has a gap in subject position, the direct object relative clause (DORC) has a gap in direct object position, and the indirect object relative clause (IORC) has a gap in indirect object position. These relative clauses have different surface word orders: While SRCs like the one in (11a) have *‘that’–Verb–Direct Object* order, DORCs have *‘that’–Subject–Verb* order and IORCs have *‘that’–Subject–Verb–Direct Object–Preposition* order.

Thus, relative clauses form a family of structures with similar properties. For each type of syntactic role in a language – subject, direct object, indirect object, oblique object, etc. – the

language may have a corresponding relative clause structure. The objective of the experiments presented here is to assess the representational underpinnings of these kinds of long-distance dependencies. There are two possibilities.

The first is that each of these structures could be represented separately, at least as initially learned. Given that even very young learners of English have likely been exposed to each of these types, productive use of the various types may have been learned from these separately experienced structures, leading to distinct representations. This would be consistent with evidence suggesting that a significant portion of linguistic knowledge reflects direct experience (Tomasello, 2000b).

But it is also possible that at some (even higher) level of abstraction, relative clauses are all represented as just one structure. Such a unitary representation might take the form of a general underlying principle like, “*Clauses can modify nouns, but when doing so, the modified noun should be omitted from the clause.*” This may sound like an unnecessarily labored method of representation to posit, but indeed, almost every theory of syntax posits just such a representational apparatus (e.g., movement and operators, Chomsky, 1981, 1992; slash categories and feature percolation Pollard and Sag, 1994; Culicover and Jackendoff, 2005; etc.).

The typology of relative clause constructions provides reasons to think that both models may have a kernel of truth. Consistent with the one-representation-per-structure approach is the fact that some languages treat the various types of relative clause differently. Hebrew, for instance, is a language that forms one type of relative clause (SRCs) with gaps, as in English, but uses a different strategy to form another type of relative clause (IORCs). Facts like this indicate that at some level the various types might be represented distinctly.

On the other hand, if there were distinct representations for each type of relative clause, then one might expect some languages to have different word orders in different relative clauses (independent of whatever surface differences result from the gap left by the missing noun). Indeed, such a difference is seen between matrix and embedded clauses in languages like German and Dutch, which have verb-medial word order in main clauses but verb-final order

in embedded clauses. It stands to reason that there might be languages with a verb-final word order in subject relative clauses but a verb-initial word order in object relative clauses. To the best of our knowledge, however, languages like this are not attested. This may reflect a single underlying representation common to all relative clauses.

For the different surface forms of relative clauses to have a single underlying representation would predict certain patterns of learning behavior. Specifically, when one relative clause is learned, it would not be learned as an isolate. Instead, others may “come along for the ride,” given that the representation that is learned is not specific to a single type of relative clause, but a general representation for modifying a noun with a clause.

Here, we present four experiments in which adults learn a new artificial language via exposure to a number of sentences with relative clauses. These sentences, however, only contain a subset of the types of relative clauses given in (11). After training, participants are asked to use the new language to describe a number of pictures, some of which elicit descriptions using the *trained* relative clauses – those which participants were exposed to – and others of which elicit *untrained* ones – those which participants were not exposed to. Generalization from trained to untrained structures is assessed to infer the structure of the underlying representation of long-distance dependencies.

If each type of relative clause has a distinct representation, then learning one structure does not implicate knowledge of other types, and we should not observe generalization. But if the various types of relative clauses have a single underlying representation, then learning one type may be equivalent to learning all types. On this account, we should expect to see that participants *do* generalize to untrained structures.

The remainder of the paper is structured as follows: In the following section, we give a bit more detail about relative clauses, particularly from a typological perspective. We then discuss some previous work validating the use of artificial language learning paradigms for studying syntactic representation. After presenting four experiments which test whether learners knowledge of relative clauses is general or specific, we interpret findings and discuss possible

implications for theories of syntax and language learning.

2.1.1 Relative clauses

There is a great deal of variation in the surface structure of relative clauses across languages. For instance, as (12) shows, English relative clauses (in brackets) appear after the noun they modify, the *head noun* (in bold). Relative clauses like this are called *postnominal*.

(12) the **girl** [that _ hugged the bear] POSTNOMINAL RELATIVE CLAUSE

But in many languages, relative clauses *precede* the head noun. Such *prenominal* relative clauses are the default for at least 141 documented languages including Amharic, Basque, Chinese, Huallaga Quechua, Malayalam, and Turkish (Dryer, 2013). If English were such a language, we might express the meaning of (12) with something like:

(13) the [that _ hugged the bear] **girl** PRENOMINAL RELATIVE CLAUSE

Another source of variation in relative clauses, mentioned earlier, depends on what role the head noun plays inside the relative clause. Examples (12) and (13) are both SRCs because the head noun, “girl,” is the subject of the relative clause verb *hugged*. Inside the relative clause, the subject role is not repeated, leaving behind a *gap* (indicated with an underscore).

In principle, the gap can have any syntactic role that nouns can have. For instance, it could be in a subject, direct object, indirect object, oblique, etc. This is true for both post- and pre-nominal relative clauses. Table 2.1 schematizes these various types of relative clauses, with mock-ups of prenominal versions using English words.

In each of the four experiments presented here, participants are presented with a grammar similar to that in the right column of Table 2.1. That is, participants are told that they will be learning a language that is a composite of English words and the grammar of Chinese (Experiments 1a and 1b) or Korean (Experiments 2 and 3). They are trained on one or two types of prenominal relative clauses (for instance, only SRCs) and then tested on their knowledge of the trained type (SRCs) and an untrained type (e.g., DORCs).

Table 2.1. Relative clauses vary along two dimensions: position of the relative clause relative to the head noun (columns) and position of the gap (rows).

Gap Position	Relative Clause Position	
	Postnominal (English-type)	Prenominal (Chinese-type)
Subject	the girl [that _ hugged the bear]	the [that _ hugged the bear] girl
Direct Object	the bear [that the girl hugged _]	the [that the girl hugged _] bear
Indirect Object	the bear [that the girl gave porridge to _]	the [that the girl gave porridge to _] bear
Oblique	the bowl [that the girl served the porridge in _]	the [that the girl served the porridge in _] bowl
Object of Comparison	the girl [that the bear was furrer than _]	the [that the bear was furrer than _] girl

2.1.2 Language learning as a tool to study syntactic representation

Our objective is to use relative clause learning as a tool to probe the nature of syntactic knowledge. We are not the first to approach the question this way. Previous studies have used similar paradigms to investigate topics such as *regularization*, a phenomenon whereby people impose grammatical structure on unstructured input (Culbertson and Smolensky, 2012; Culbertson et al., 2012; Kam and Newport, 2009; Saldana et al., 2018; see also Senghas et al., 2004 for a longitudinal study of how this happened in a new natural language over several generations). Others have tested theories about specific pressures on language processing by teaching participants a grammar with, for instance, inefficient properties and showing that learners change the grammar to make it more efficient (Fedzechkina et al., 2012). (See also Culbertson and Newport, 2017; Tily et al., 2011; Christiansen, 2001; Bley-Vroman et al., 1988; Gass and Ard, 1980.)

In the present study, we aim to measure *generalization*. Generalization has long been used as a tool to study representation – in particular, to study the abstractness of linguistic representations. For instance, Berko (1958) showed that children as young as 4 who are introduced to a new word, such as *wug*, can correctly generalize to a previously unseen form

of that word, *wugs*. Similarly, Kaschak (2006) exposed adult speakers of Standard American English to the nonstandard *needs* construction, as in “the car needs washed.” After a training regimen involving exposure to several instances of this construction using the verb *needs*, Kaschak showed that participants subsequently generalized to novel instances of the construction involving different verbs (e.g., “the dog *wants* walked”) and different global syntax (a pseudocleft construction, e.g., “what the car needs is washed”).

Similarly, Culbertson and Adger (2014) report an experiment where they trained participants on a new grammar for the structure of nouns. There is a near-universal hierarchical order of elements inside determiner phrases (DPs): determiner > number > adjective > noun (Greenberg, 1963). The particular linear order can vary from language to language, but this hierarchical order is almost always maintained. For instance, in the English DP “those two big boxes,” the order is [determiner [number [adjective [noun]]]], where brackets indicate levels of the hierarchy. In Spanish the order is [determiner [number [[noun] adjective]]], which differs from English in that adjectives appear to the right of the noun, but is similar in that the same hierarchical structure can be applied. Similarly, in Arabic the order is [[number [[noun] adjective]] determiner], and in Yoruba it is [[[noun] adjective] number] determiner].

Culbertson and Adger (2014) trained adult English speakers on a novel grammar by showing them different noun phrases, each containing a noun followed by only one modifier at a time, for example, “boxes two” and, separately, “shoe big.” After training, participants were asked to produce NPs with three elements: *cars*, *red*, and *three*. Participants had learned that adjectives come after the noun and that numbers come after the noun, but they had never seen the two relative order of numbers and adjectives. They therefore had to choose the relative order of the number and the adjective.

Culbertson and Adger’s participants tended to choose the order that is consistent with the universal hierarchical structure – “big two” – even though this is different from the surface word order of English, which would be “two big.” They interpreted this finding as evidence for the hierarchical structure of representations. Similar findings have been reported in another artificial

language learning task that looked at the ordering of case and number morphemes by native speakers of English or Japanese (Saldana et al., 2019), as well as in tasks where non-signing participants use gestures to describe objects or events (Culbertson et al., 2016; Hall et al., 2013).

Four points can be gleaned from this body of work. First, at least some syntactic representations are abstract. This accounts for the productivity observed in studies like Berko (1958) and Kaschak (2006). Second, at least some syntactic representations are hierarchical, not linear, accounting participants' generalizations to word orders that are consistent with universal patterns (Culbertson and Adger, 2014; Saldana et al., 2019). Third, artificial language learning tasks present a valid way to approach questions about syntactic representation (Fedzechkina et al., 2016; Culbertson, 2012; Tily et al., 2011). Finally, findings from artificial language learning tasks reveal behaviors that cannot be traced back to pre-existing knowledge of languages that participants already speak, and therefore reflect properties of human linguistic abilities in general (Culbertson and Adger, 2014; Saldana et al., 2019; Tily et al., 2011; Culbertson, 2012).

2.1.3 The present study

Following these approaches, we present a series of experiments in which monolingual English speakers are trained on artificial languages. We aimed to measure generalization in order to better understand the makeup of the mental representation of syntactic structures. Of particular interest is which of two possible representational systems underlies speakers' knowledge of long-distance dependencies: several distinct representations or one general one.

In all four experiments, artificial languages with prenominal relative clauses were taught to participants via exposure. English words were used so as to minimize the amount of training necessary. The languages differed from English in their word orders and morphological properties, which were systematically varied to parcel out the contributions of participants' knowledge of English.

Different groups of participants were trained on different subsets of the artificial languages. For instance, some groups were trained only on prenominal SRCs, and others only on

prenominal DORCs. After training, all groups were tested on both the trained and untrained structures. If each type of relative clause is learned independently and therefore has a distinct mental representation, then participants should not be able to produce untrained structures. But if upon being exposed to one type of relative clause, participants acquire a representation of relative clauses in general, they should be able to produce relative clauses even of types they have not previously seen.

We aimed to control for a number of potential confounds. One concern stemmed from the fact that our participants were adults. Adult language learning and child language learning may be different in ways that bear on the interpretation of our task. The specific problem, according to Bley-Vroman (1989) is that there are different sources of information available to adults and infants during language acquisition. Infants have whatever learning mechanisms are in place early in life, and these may or may not be present in adults. Independently, adults have languages they already know as well as adult reasoning abilities; neither of which are available to infants. If adults in our study use either of these sources of information, it may result in syntactic representations that are different from those acquired by children in more naturalistic settings.

Experimental work provides mixed evidence. For instance, transfer effects, where properties of a speaker's native language are observable in their nonnative language, are common, indicating that prior linguistic knowledge can impact the acquisition of a new system (e.g., Gass, 1979). However, transfer effects would only pose a problem in the present study if learners' production of untrained structures could be explained in terms of English syntax. As we discuss later, this may be possible in Experiments 1a and 1b, but it is unlikely to account for any generalization in Experiments 2 and 3. Another difference that has been noted between adult and child language learning is that adults are more veridical: where children will almost always impose structure to regularize an unstructured input, adults sometimes will (Culbertson and Smolensky, 2012; Culbertson et al., 2012; Fedzechkina et al., 2012), and other times will not (Kam and Newport, 2009; Culbertson and Newport, 2017). This may pose a problem for interpreting the results of the present study if we find that learners do not generalize to

untrained structures, as it offers a potential explanation for the lack of generalization that does not necessarily have anything to do with the type of representation acquired. But if participants do generalize, then it will have manifested despite adults' tendencies to learn more veridically.

There is also evidence supporting the idea that adult language learning can be similar to child language learning. Some of this was discussed above, such as Culbertson and Adger's 2014 finding that English speakers will generalize to novel word orders when those orders are predicted by a universal hierarchical model and not by the word order of their native language. Bley-Vroman et al. (1988) discuss another study which demonstrates that adults understand constraints in their nonnative languages that they were never explicitly taught, a hallmark characteristic of natural language acquisition. Gass and Ard (1980) even argue that because adult cognition is not subject to development-related constraints, adult language learning may be a "purer" way to study whatever abstract knowledge humans have about language in general.

Because the evidence is mixed, we used an abundance of caution and took steps to prevent participants from using knowledge of English to guide any generalization to untrained structures. We did so by designing grammars that were different from English, so that participants would be unlikely to be able to use English syntax to speak the artificial language. In all experiments, these grammars contained finite prenominal relative clauses, which do not exist in English (although see Eilish, 2019). Furthermore, we varied other aspects of the grammar, such as basic word order and DP structure, from experiment to experiment such that the languages became increasingly different from English. This in turn made it decreasingly likely that participants could successfully rely on English to generate the novel structures (or that, to the extent that participants were relying on knowledge of English, trends in generalization as a function of similarity to English should be observed).

Another potential issue we aimed to avoid was participants developing and using explicit knowledge of the grammar. Syntactic knowledge is implicit knowledge, and so any production that is guided by explicit knowledge probably does not stand to shed light on the question at hand. For instance, an adult might develop an explicit strategy, such as "produce a determiner,

then a relative clause with the verb at the end, then the head noun.” If participants used such a strategy, they may have been able to produce the untrained structure without actually generating a syntactic representation.

However, explicit knowledge of complex structures like relative clauses is extremely difficult to formulate. Anecdotally, it often takes several semesters of syntax coursework for undergraduate students to understand the syntax of long-distance dependencies. As such, this was regarded as a relatively minor concern. We nonetheless took steps to determine whether participants used explicit knowledge. First, we included an “EXPLICIT” training group in Experiment 1a, in which participants received a grammar lesson on prenominal relative clauses (instead of exposure to prenominal relative clauses, as the other groups received). If explicit knowledge underlies the behavior of the implicitly trained groups, then we should expect the explicit and implicit groups to perform similarly. Second, the grammars of the artificial languages became increasingly less similar to English from experiment to experiment, which made it increasingly difficult for participants to formulate explicit rules. Finally, after the experiment all participants completed a debriefing survey. Participants were required to respond to the prompt, “Describe the sentences we trained you on. Can you explain the rules to follow to make them?” We excluded data from any participants who provided a response that coherently articulated the grammar.

A final consideration we kept in mind when designing stimuli had to do with typological validity. When considering grammatical properties across languages, a number of distributional patterns emerge. There is a divide in the field between researchers who view these as the result of processing pressures or language contact effects (Dunn et al., 2011; Evans and Levinson, 2009; MacDonald, 2013; Fitz et al., 2011) and those who think that they might reflect more rigid constraints on language (Culbertson and Adger, 2014; Gass, 1979; Gass and Ard, 1980; Bley-Vroman et al., 1988). If the latter, then grammars which are unattested may in fact be impossible to learn with the usual language processing architecture. We do not take a position on this debate, but out of an abundance of caution we attempted to design our stimuli so as to

conform to typologically attested grammars, particularly in Experiments 2 and 3.

2.2 Experiment 1a

In Experiment 1a, participants were trained on an artificial language with prenominal relative clauses. After training, their knowledge of the grammar was tested in a production task where they described pictures using the new grammar. Two groups were trained on only one type of relative clause: a SRC-ONLY group, trained only on SRCs, and a DORC-ONLY group, trained only on DORCs. These groups are collectively referred to as the ONLY groups. After training, both groups were tested on their knowledge of both prenominal SRCs and DORCs. If different types of relative clauses have distinct syntactic representations, then we expect participants not to generalize to the UNTRAINED structures. If, on the other hand, there is one general underlying representation for all types, then participants should be able to produce the UNTRAINED structures.

On the hypothesis that learning might be improved by a more diversified input grammar (consistent with *desirable difficulty* effects; Bjork, 1994), a third “BOTH” group received exposure to SRCs and DORCs. This group was tested on two TRAINED structures (and no UNTRAINED ones).

To ensure that the ONLY groups learned implicit representations, not explicit ones, participants were trained with direct exposure to sentences rather than explicit grammar lessons. However the possibility remained that such training might allow participants to develop explicit knowledge of the structure. A fourth group was therefore trained with an explicit grammar lesson so as to compare the performance of the implicitly trained groups (SRC-ONLY, DORC-ONLY, and BOTH) to performance that unambiguously reflected explicit knowledge.

Experiment 1a therefore was designed to answer the following questions: (1) Do participants trained on only one type of structure generalize to the other type? (2) If so, can this be attributed to explicit grammatical knowledge? (3) Does learning of trained structures benefit

from a more syntactically diverse input?

2.2.1 Method

Participants

Participants were continuously run until a target of 24 per group was met. A total of 112 UC San Diego undergraduates participated for course credit. Pre-screen requirements were the same for all experiments: these included that participants were over 18 years old and were native monolingual speakers of English (defined as not having learned any language but English before the age of 7). A total of 16 participants were excluded: 2 for knowing a language with finite prenominal relative clauses (such as the ones in the artificial language), and 14 due to software or experimenter errors.

Factors

The experiment had one between-subjects factor with four levels, training GROUP, and one within-subjects factor with two levels, TRIAL TYPE. Participants were randomly assigned to one of four training groups. The SRC-ONLY group received implicit training on 36 prenominal SRCs; the DORC-ONLY group received implicit training on 36 prenominal DORCs; the BOTH group received implicit training on 18 prenominal SRCs and 18 prenominal DORCs; and the EXPLICIT group received an explicit grammar lesson on prenominal relative clauses but no implicit training (other than whatever may have been gleaned from a single example of a prenominal DORC). After training, all groups were tested on 36 previously unseen items. Items appeared in one of two conditions: either TRAINED trials, where the stimulus was designed to elicit a trained structure (SRCs for the SRC-ONLY group; DORCs for the DORC-ONLY group, and SRCs or and DORCs for the BOTH GROUP) or UNTRAINED trials (DORCs for the SRC-ONLY group; SRCs for the DORC-ONLY group, and SRCs and DORCs for the EXPLICIT group). This TRIAL TYPE manipulation was counterbalanced across items such that for a given item, half of participants in each group saw it in a TRAINED trial and half saw it in an UNTRAINED trial.

Materials

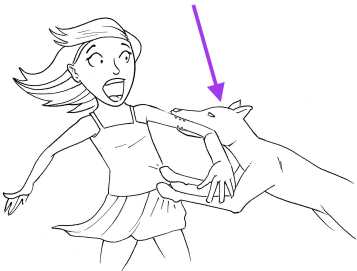
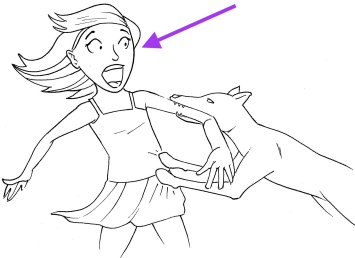
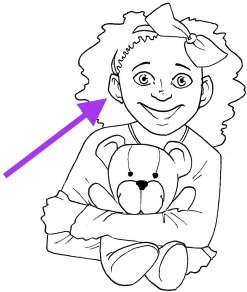
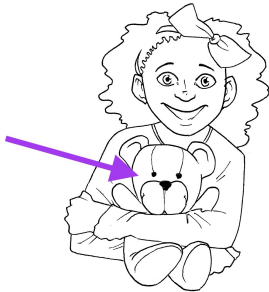
In all experiments, the artificial language used English words but non-English syntax. This greatly reduced the amount of training relative to a more typical task where both a new lexicon and grammar would have to be taught. The syntax of the relative clauses in Experiment 1a and 1b was identical to English except in that the relative clauses appeared prenominally rather than postnominally. Each sentence started with functional material that introduced a noun, for example, “Here’s the...,” or “These are some...,” or “Now we have a...” This was followed by a transitive relative clause headed by a clause-initial complementizer (“that”) and either a subject gap or a direct object gap, as in “__ hugged a bear” or “the girl hugged __.” The relative clause was in turn followed by the head noun, which for some items was a bare noun (“girl”) and for others contained a prenominal adjective (“little girl”).

It is worth noting that this grammar is odd from a typological perspective. Cross-linguistically, languages with prenominal relative clauses tend to put complementizers at the end of the clause, and while there may be a few exceptions to this (e.g., Amharic, Laz, and Tigré; see Wu, 2011 for a full discussion), to our knowledge none of these languages have clause-initial complementizers in prenominal relative clauses. We nonetheless used clause-initial complementizers because we worried that without them, DORCs would be too odd given that they would require two adjacent determiners, as in “... the [the girl hugged __] bear.” Indeed, natural languages deal with this situation in peculiar ways, as in the case of St’át’imcets, a Northern Interior Salish language which simply deletes one of the two adjacent determiners (Davis, 2010). Concerns relating to the typological validity of complementizers are addressed in subsequent experiments, where they are removed altogether.

Sample materials appear in Table 2.2. Training materials consisted of 36 pairs of images of transitive events, each with an accompanying sentence using a prenominal relative clause to describe one of the objects in the picture. Each pair of images showed the same event, but in one, an arrow pointed to the subject, and in the other an arrow pointed to the direct object. For

images with a subject arrow, the accompanying sentence used a prenominal SRC. For images with a direct object arrow, the accompanying sentence used a prenominal DORC.

Table 2.2. Sample materials from Experiments 1a and 1b.

	SRC	DORC
Train	 <p>“That’s the [that __ bit the little girl] dog.”</p>	 <p>“That’s the [that a dog bit __] little girl.”</p>
Test	 <p><i>hugged</i></p>	 <p><i>hugged</i></p>

* While these were different from the input grammar, they are counted as correct in all Exp. 1a analyses. *Note.* Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported in this table.

Testing materials consisted of 36 additional pairs of images, similarly created in pairs with arrows pointing to subjects or objects. These images were paired with verbs which were presented on the screen underneath the image during the test phase.

Procedure

All participants began by providing informed consent. They were then instructed that they would be trained on a new language which used English words but Chinese word order. The task took about 60 minutes for participants in the implicit training groups (SRC-ONLY,

DORC-ONLY, and BOTH) and under 30 minutes for participants in the EXPLICIT group.

For participants in the implicit training groups, the experiment had three phases: two training and one test. In the first training phase, participants saw a picture appear on a monitor with an arrow pointing to either the agent or patient (as determined by their group) and the experimenter read aloud the corresponding sentence. The participant was instructed that they would have to repeat the sentence soon, so to listen carefully and ask the experimenter to repeat as needed. After two trials like this, the two images from the preceding two trials would appear one at a time in random order with a prompt indicating that the participant should try to recall the sentence as the experimenter had said it. If the response was correct, they went on to the next trial. If it was incorrect, the experimenter corrected them.

Responses during training phases were considered correct if they contained a grammatically well-formed prenominal relative clause with an overt complementizer (“that”) and the correct meaning. This meant that the specific words could differ from those that the experimenter had initially used to describe the picture; lexical substitutions like “nurse” for “doctor” or “the” for “a” were not considered errors. Experimenters, who were undergraduate RAs, were told that if they were ever in doubt about whether a participant’s utterance was correct, to read the sentence again and ask the participant to repeat it verbatim.

Experimenters were instructed not to correct participants until they had made a reasonable attempt at recall (following findings from the learning literature that learning is enhanced when learners are tested rather than given information during training, e.g., Kang et al., 2013; see Roediger III and Karpicke, 2006 for a review). Experimenters did not give corrections in the form of meta-linguistic commentary (e.g., “remember to say ‘that’” or “use past tense”), which we worried might facilitate explicit learning, but to simply repeat the full sentence if the participant made any errors. If the participant was not able to say it correctly after three tries, then the experimenter could go on to the next trial so as to minimize frustration. The first training phase ended after all 36 training images had been presented and recalled in this fashion.

In the second training phase, participants saw the same 36 training images again, pre-

sented one at a time in random order and with the verb written underneath. They were instructed to try to recall the sentences as they had in the previous phase. Again, experimenters corrected them as needed by reading aloud the full sentence, but only after they had attempted to recall the sentence.

Instead of 36 example sentences, the EXPLICIT training group received a formal grammar lesson accompanied by a single example of one prenominal DORC and its postnominal translation (as in Table 2.1). Experimenters let participants read these examples on the screen and then gave a more detailed explanation. They explained that relative clauses are sentences, like “The girl hugged the bear,” which modify nouns, like “girl” or “bear.” Using the on-screen postnominal (English-type) relative clause, they pointed out that the noun being modified is not repeated inside the relative clause. They then explained that while in English, relative clauses appear after the noun they modify, in other languages like Chinese, the relative clause appears before the modified noun. During the course of these instructions, they read the prenominal relative clause aloud exactly once. Participants were instructed to ask as many questions as they wanted before they would be asked to produce prenominal relative clauses on their own without receiving help.

After training, all groups went on to the same test phase. Participants were instructed that they would describe 36 brand new images using their newly acquired grammar. They were told that the experimenters could not provide feedback during this phase, but may occasionally ask them to repeat the full sentence fluently to facilitate later transcription of the audio recording. All participants in the implicit training groups saw 18 images that elicited SRCs and 18 that elicited DORCs. For the BOTH and EXPLICIT groups, these appeared in random order, while for the SRC-ONLY and DORC-ONLY groups the order was pseudo-randomized such that the first four trials always elicited TRAINED structures. This was intended to facilitate transition to the test phase before (implicitly) asking them to generalize to the untrained structure.

Data coding and analysis

After the experimental session, research assistants manually transcribed and coded responses. Data from 22 trials were excluded because the trial was inadvertently skipped, a relative clause was produced with an intransitive verb,¹ or an incorrect response was due to experimenter error (e.g., if the experimenter did not correct a misinterpretation of the depicted event). The remaining responses were coded for structure type and for what errors they contained (if any). Structure types could be SRC, DORC, or uncodable (any utterance that could not clearly be labeled as a SRC or a DORC or that contained three or more errors). The most common errors were given specific labels, shown in Table 2.3. Responses were coded as correct if they were grammatically well-formed instances of the elicited structure and if they conveyed the correct meaning. To be considered grammatically correct, the particular words did not matter, but the response had to be a full sentence with a prenominal finite transitive relative clause. Due to the high number of instances where a relative pronoun (e.g., “who”) was used rather than the overt complementizer (“that”), or where neither was used, we counted all three of these alternatives as correct in all analyses.

For all experiments, logistic mixed effects regressions were used to model responses as a function of GROUP and TRIAL TYPE (R Core Team, 2018; Bates et al., 2015); both factors were treatment coded. All models contained random intercepts for participants and items, and all fixed effects were allowed to vary by random factors if the effect varied within the factor. Following Barr et al. (2013), we report the maximal random effects structure which allowed the model to converge: random effects correlations and then random slopes were removed from the full model one by one in order from least variance accounted for to most until the model converged. For all analyses, we report the model output, and for significant effects of theoretical interest, we also report the results of model comparisons to determine whether each fixed effect contributed

¹Fox (1987) has argued that intransitive subjects may be more similar to direct objects than to transitive subjects when it comes to relativization from the perspective of processing and typology; we therefore excluded the few trials where participants produced intransitive active-voice SRCs so as not to inadvertently give one group an advantage.

Table 2.3. Percentage of the most common errors in Experiment 1a, by condition.

Error	BOTH Trained	EXPLICIT Untrained	SRC-ONLY		DORC-ONLY		Overall
			Tr.	Untr.	Tr.	Untr.	
Missing head determiner (“ <i>Here’s that hugged the bear girl.</i> ”)	11.92	0.81	15.74	15.04	21.76	18.52	12.07
Relative pronoun* (“ <i>Here’s the who hugged the bear girl.</i> ”)	10.07	2.08	11.11	12.50	11.34	12.27	8.94
English (“ <i>Here’s the girl that hugged the bear.</i> ”)	0.69	29.63	0.93	1.16	4.17	3.93	8.85
Wrong type (e.g., well-formed SRC in response to a DORC-elicitation)	4.05	1.85	0	28.7	1.39	1.13	6.65
Missing determiner inside RC (“ <i>Here’s the that hugged bear girl.</i> ”)	4.75	1.04	2.08	0.93	6.94	8.10	3.70
Number agreement error (“ <i>Here’s the that the grandma baked cookies.</i> ”)	1.74	1.97	1.39	3.47	9.95	6.02	3.57
No complementizer* (“ <i>Here’s the hugged the bear girl.</i> ”)	2.55	5.32	3.70	1.39	3.24	2.78	3.36
Head determiner after RC (“ <i>Here’s that hugged the bear the girl.</i> ”)	0.69	1.50	4.17	3.70	4.63	4.63	2.69
Repeated head determiner (“ <i>Here’s the that hugged the bear the girl.</i> ”)	1.85	4.05	1.85	1.39	2.31	3.24	2.58

* While these were different from the input grammar, they are counted as correct in all Exp. 1a analyses. *Note.* Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported in this table as it was often too difficult to determine which particular errors led to the response.

significantly to model fit.

To address the question of whether participants generalize to untrained structures, it is important to establish that performance on untrained structures is significantly higher than what would be expected by chance. However, chance performance is difficult to estimate in this task. A clearly too-conservative estimate would be the chances of arriving at a correct structure by randomly ordering the words in an utterance. That is, for a 7-word utterance like “Here’s the that hugged the bear girl,” any particular string has a $\frac{1}{7!} = 0.02\%$ chance of arising by randomly ordering the words.

Rather than attempting to estimate chance directly, for each experiment we instead perform an analysis where we separate participants into two subsets: those who produced more than 50% of the trained structures correctly and those who produced fewer than 50% correctly. If production of untrained structures simply reflects chance performance, then we should expect the amount of generalization between these two groups not to differ. However, if generalization reflects learning of a generalized representation for relative clauses, then we should expect that the subset that learns the representation better will produce more of the untrained structures. These analyses are mixed-effects logistic regressions formulated as above, but with fixed-effects terms for GROUP and SUBSET – whether they produced the trained structure correctly on more or less than 50% of trials.²

2.2.2 Results

Data are shown in Figure 2.1 and model results appear in Table 2.4. Our first a priori questions was: Do the ONLY groups generalize to the untrained structures? That is, does the SRC-ONLY only group learn DORCs and does the DORC-ONLY learn SRCs? We addressed this with Model 1, a 2×2 model of response as a function of TRAINED vs. UNTRAINED structure and GROUP: SRC-ONLY and DORC-ONLY. Data from the BOTH and EXPLICIT groups were excluded from this analysis. This model converged with the full random effects structure without random

²Thanks to Eva Wittenberg for suggesting this clever approach.

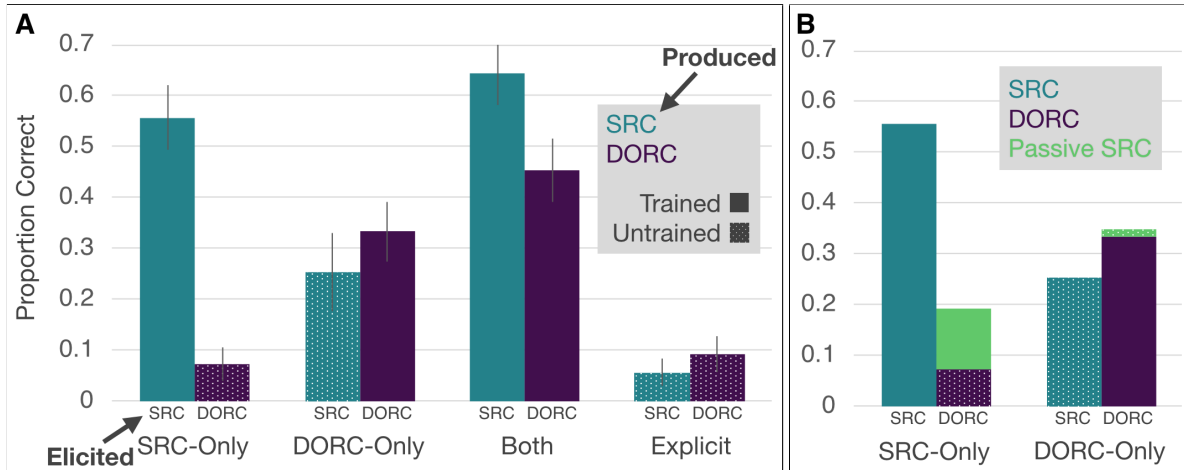


Figure 2.1. Experiment 1a results: (A) All correct responses and standard errors as a function of group and elicited structure (SRC or DORC). Colors indicate the type of structure produced and dots indicate untrained structures. (B) The ONLY groups again, but including well-formed passive SRC productions.

effects correlations.

There was a main effect of GROUP (confirmed by model comparison; $\chi^2(1) = 6.588$, $p = .010$) whereby the DORC-ONLY group performed less well than the SRC-ONLY group. There was a main effect of TRIAL TYPE ($\chi^2(1) = 41.094$, $p < .001$), whereby participants were less correct on untrained trials than on trained trials. Finally, there was an interaction ($\chi^2(1) = 7.560$, $p = .006$) reflecting the fact that the DORC-ONLY group produced more untrained structures than the SRC-ONLY group.

To determine whether generalization was above chance, we looked for a relationship between learning of trained structures and generalization to untrained structures (Figure 2.2). We performed a subset analysis, comparing the amount of generalization by participants in ONLY groups who produced more than 50% of trained structures correctly to those who produced fewer than 50% correctly. The model converged with the full random effects structure. The results of this model, Model 1.2 in Table 2.4, show that the 21 higher-performing participants produced significantly more untrained structures (10% for the SRC-ONLY group and 67% for the DORC-ONLY group) than the 27 lower-performing participants (2% for the SRC-ONLY group and 11% for the DORC-ONLY group; model comparison confirmed that the effect contributes

Table 2.4. Experiment 1a results.

	Model results			
	β	z	p	
<i>Model 1: ONLY groups</i> \times TRIAL TYPE				
Intercept	0.276	0.785	.432	
GROUP: DORC-ONLY	-1.305	-2.626	.009	**
TRIAL TYPE: UNTRAINED	-4.957	-6.599	< .001	***
Interaction	2.874	3.103	< .001	***
<i>Model 1.2: ONLY groups' UNTRAINED trials, split by performance on TRAINED</i>				
Intercept	-10.182	-4.237	< .001	***
GROUP: DORC-ONLY	4.825	2.039	.041	*
SUBSET: $\leq 50\%$	5.719	2.402	.016	*
Interaction	0.623	0.220	.826	
<i>Model 1.3: A re-run of Model 1, but with passive SRCs coded as correct</i>				
SRC-ONLY, TRAINED (intercept)	0.280	0.780	.435	
DORC-ONLY, TRAINED	-1.231	-2.427	.015	*
SRC-ONLY, UNTRAINED	-2.453	-6.040	< .001	*
DORC-ONLY, UNTRAINED (interaction)	0.828	1.340	.180	
<i>Model 2: Just untrained trials (EXPLICIT and ONLY groups)</i>				
Intercept	-5.666	-5.850	< .001	***
GROUP: ONLY	2.543	2.274	.023	*
TRIAL TYPE: DORC	.0170	0.235	.814	
Interaction	-2.571	-1.983	.047	*
<i>Model 3: Just trained trials (BOTH and ONLY groups)</i>				
Intercept	0.904	2.562	.010	*
GROUP: ONLY	-0.621	-1.284	.199	
TRIAL TYPE: DORC	-1.228	-3.681	< .001	***
Interaction	-0.213	-0.331	.741	

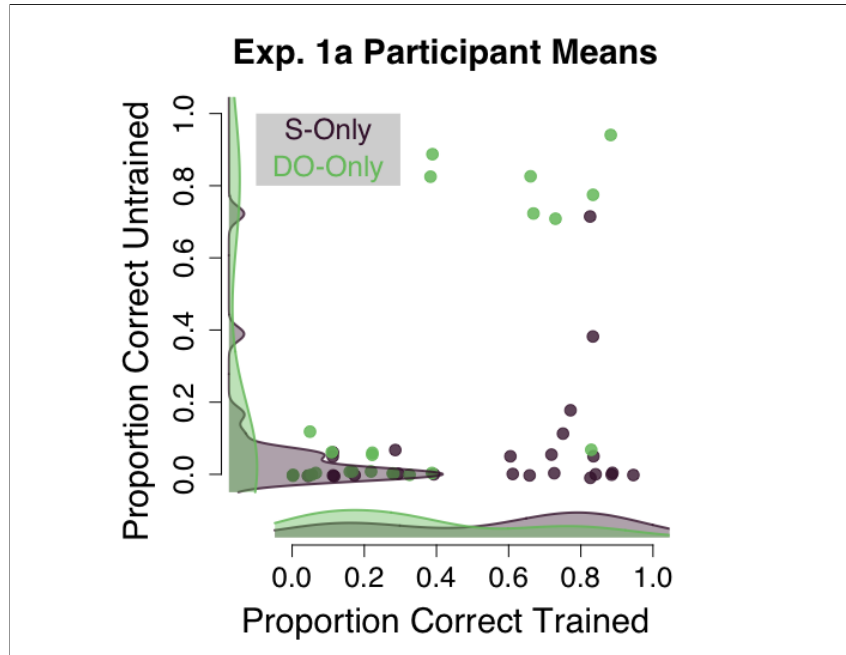


Figure 2.2. Experiment 1a: The relationship between learning trained structures (horizontal axis) and generalization to untrained structures (vertical). Individual participants (dots) are jittered with standard deviation of .005. Density plots appear along axes. Participants who learned the trained structures better also generalized more, suggesting that generalization did not reflect chance performance.

to model fit: $\chi^2(1) = 4.860, p = .027$). This indicates that generalization, at least among the high-performers, does not reflect chance.

The results of Model 1 indicate that the SRC-ONLY group did not generalize to untrained structures as much as the DORC-ONLY group. But as Figure 2.1b shows, the SRC-ONLY group produced an unexpected number of passive structures in DORC-TRIAL trials. Passivizing a DORC results in a SRC with the same meaning as the original DORC. That is, the meaning of the DORC in “the bear [that the girl hugged __]” can be expressed just as well with the passive SRC, “the bear [that __ was hugged (by the girl)].” The SRC-ONLY participants discovered a way to respond correctly on DORC-eliciting trials without having to generalize to the untrained structure. Thus it is not clear whether the SRC-ONLY group did not generalize to DORCs because they did not know how, or because there was an easier strategy – passivization – available to them.

To determine whether the asymmetry in generalization indexed by the interaction term in Model 1 might have disappeared if SRC-ONLY participants had successfully produced DORCs on every trial in which they produced a passive SRC, we re-ran Model 1 on updated data in which all well-formed passive SRCs produced in DORC-TRIAL trials were coded as “correct.” This model, Model 1.3 in Table 2.4, converged with a full random effects structure but without random effects correlations. The results still show a clear difference between trained and untrained structures, but the interaction representing asymmetrical generalization was no longer significant.

Our second question was whether generalization to untrained structures might reflect use of explicit knowledge of the grammar during production. This question was addressed by Model 2, which compared the ONLY groups’ performance on untrained structures to the EXPLICIT group. Because there were no TRAINED trials in this analysis, the TRIAL TYPE factor here was coded as SRC-eliciting and DORC-eliciting. No data from the BOTH group nor the TRAINED structures for the ONLY groups were included. The model converged with the full random effects structure but without random effects correlations.

Results of Model 2 showed that the explicit group’s production of SRCs and DORCs did not significantly differ (no main effect of TRIAL TYPE $\chi^2(1) = 0.052, p = .820$). The ONLY groups produced significantly more UNTRAINED structures than did the EXPLICIT group (a main effect of GROUP; $\chi^2(1) = 4.935, p = .026$). There was a significant interaction, reflecting the fact that the DORC-ONLY group generalized more than the SRC-ONLY group, however this should be interpreted with caution as model comparison revealed that it only marginally contributed to model fit ($\chi^2(1) = 3.674, p = .055$).

Our third question was whether participants’ learning of trained structures would benefit from a more syntactically diverse input – specifically, if the BOTH group would learn SRCs better than the SRC-ONLY group and DORCs better than the DORC-ONLY group. Model 3 converged with the full random effects structure but without random effects correlations. Results did not provide evidence for the hypothesis: there was no main effect of GROUP ($\chi^2(1) = 1.629, p = .202$), although there was a main effect of TRIAL TYPE, reflecting the fact that, across groups,

DORCs were learned less well than SRCs ($\chi^2(1) = 11.552, p < .001$). The interaction was also not significant ($\chi^2(1) = 0.107, p = .743$).

2.2.3 Discussion

Experiment 1a showed that learners are indeed capable of producing previously unseen structures: participants in the DORC-ONLY group generalized to untrained structures, although participants in the SRC-ONLY group appeared not to. This finding is supported by the fact that participants who learned trained structures better also generalized to untrained structures more, suggesting that generalization does not reflect chance. Furthermore, given that EXPLICIT participants received explicit training on prenominal relative clauses, their performance may be viewed as a liberal estimate of chance. The DORC-ONLY group performed significantly better than the EXPLICIT group, further supporting the idea that this group successfully generalized. The SRC-ONLY group, on the other hand, performed comparably to the EXPLICIT group on untrained structures.

While the results of Model 3 did not show statistical evidence for a benefit associated with a syntactically diverse input, it is worth noting that the BOTH group performed numerically better on trained structures than either of the ONLY groups, and this was despite receiving only half the exposure to each trained structure. This suggests a possible benefit for diverse inputs that went undetected due to a lack of statistical power.

Finally, a post-hoc analysis (Model 1.3) revealed a potential explanation for why the SRC-ONLY group generalized less than the DORC-ONLY group. On several UNTRAINED trials, SRC-ONLY participants produced a passive SRC that both had the same meaning as the target DORC sentence and was of the TRAINED type (Figure 2.1b). It may have been easier for SRC-ONLY participants to produce passive SRCs than to generalize to DORCs. This lack of generalization may therefore reflect participants using a simpler strategy than generalization to respond to UNTRAINED trials. Experiment 1b was designed to remove this confound.

2.3 Experiment 1b

Experiment 1b was a partial replication of Experiment 1a. Two groups were trained on prenominal relative clauses: a SRC-ONLY group received training only on SRCs and a DORC-ONLY group received training on DORCs. The experiment differed from Experiment 1a in that the verbs given in the test phase were presented in unambiguously active-voice forms (e.g., “was hugging”) rather than in the simple past (e.g., “hugged”), which is often ambiguous between simple past and the participle form used in passives. Participants were instructed to use the verbs as written. If the asymmetry in generalization observed in Experiment 1a reflects SRC-ONLY participants not generalizing because they were not able to, then we should continue to see the asymmetry in this study. If, however, participants did not generalize simply because an easier response strategy was available to them, then we should no longer see an asymmetry in generalization.

2.3.1 Method

Participants

Participants were continuously run until the target of 24 per group was reached. A total of 52 UC San Diego undergraduates participated for course credit. Four participants were excluded: two for natively speaking a language other than English, one due to experimenter error, and one for producing a high number of passive SRCs during the test phase.

Factors

Two factors were manipulated: TRIAL TYPE, which was a within-subjects factor with two levels, TRAINED and UNTRAINED; and training GROUP, a between-subjects factor with two levels, SRC-ONLY and DORC-ONLY.

Materials

Materials were the same as those in Experiment 1a except for the verbs provided during the test phase. To prevent participants from producing passive structures, verbs were presented in unambiguously active-voice forms (e.g., “was hugging”).

Procedure

The procedure was identical to that of Experiment 1a except that during the test phase, if a participant produced the verb in a different form than that on the screen, experimenters asked participants to try again, using the verb as written. No feedback was given on the basis of the well-formedness of the response.

2.3.2 Results

Results are shown in Figure 2.3 and summarized in Table 2.5; the final model converged with the full random effects structure after random effects correlations were removed. While DORC-ONLY participants still produced numerically fewer trained structures than SRC-ONLY participants, this main effect was no longer significant. The SRC-ONLY group produced significantly fewer untrained structures than trained ($\chi^2(1) = 21.287, p < .001$). Critically, the interaction representing the asymmetry in generalization was not significant as it had been in Experiment 1a ($\chi^2(1) = 0.373, p = .542$).

To test whether generalization in this experiment reflected chance production, we again compared participants who produced the TRAINED structure correctly more than 50% of the time to those who did not (Model 2; Figure 2.4 shows the relationship between learning of trained and untrained structures across participants). The model converged with the full random effects structure. A significant main effect of SUBSET confirmed that the 24 higher-performing participants generalized to the untrained structure (55% for the SRC-ONLY group and 63% for the DORC-ONLY group) more than the 24 low-performing participants (14% for SRC-ONLY and 12% for DORC-ONLY; model comparison confirmed that SUBSET significantly contributed to

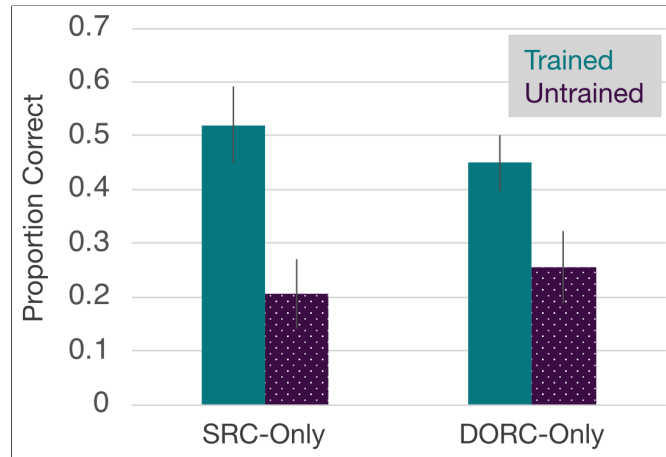


Figure 2.3. Proportion well-formed responses of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 1b.

Table 2.5. Experiment 1b results.

	Model results			
	β	z	p	
<i>Model 1: ONLY groups</i> × TRIAL TYPE				
Intercept	0.050	0.114	.909	
GROUP: DORC-ONLY	-0.472	-0.778	.436	***
TRIAL TYPE: UNTRAINED	-2.690	-4.812	< .001	
Interaction	0.477	0.615	.538	
<i>Model 2: ONLY groups' UNTRAINED trials, split by performance on TRAINED</i>				
Intercept	-4.594	-3.552	< .001	***
GROUP: DORC-ONLY	-0.077	-0.052	.959	
SUBSET: ζ 50%	3.040	2.038	.033	*
Interaction	1.520	0.867	.356	

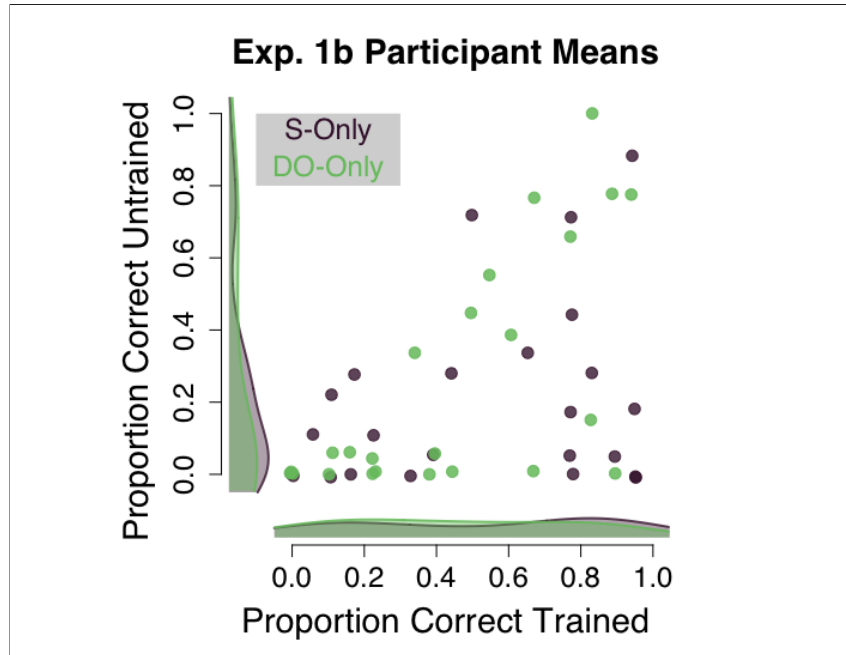


Figure 2.4. Experiment 1b: The relationship between learning trained structures and generalization to untrained structures.

model fit: $\chi^2(1) = 6.255, p = .012$). This finding did not vary by group, and again indicates that generalization, at least among the high-performers, reflects above-chance performance.

2.3.3 Discussion

Experiment 1b removed a confound from Experiment 1a – namely, that participants in the SRC-ONLY group were able to produce either a trained structure or an untrained structure to correctly respond to DORC-TRIAL trials. We reasoned that if the asymmetry in generalization observed in Experiment 1a reflected this confound, it would disappear in Experiment 1b where participants were prevented from producing passive SRCs in response to DORC-eliciting pictures. Indeed, while the SRC-ONLY group produced numerically fewer DORCs (20.6%) than the DORC-ONLY group produced SRCs (25.5%), this difference was no longer significant.

Experiments 1a and 1b therefore demonstrate that participants trained on only one type of relative clause are able to generalize to untrained types. However, the fact that participants in Experiment 1a produced English passive structures inside prenominal relative clauses suggests a

heavy reliance on knowledge of English. While we had attempted to prevent participants from using knowledge of English by teaching prenominal relative clauses, this may not have gone far enough. Experiment 2 was designed to address this by employing a new grammar that is even more different from English. It also uses prenominal relative clauses, but these clauses have a verb-final word order.

Experiment 2 was also designed to increase statistical power. This followed from the observation that some features of the data differed from Experiment 1a to 1b in unexpected ways. For example, DORC-ONLY participants produced only 33.3% DORCs in Experiment 1a, but 44.9% in Experiment 1b. We therefore aimed to increase power in two ways. First, we increased the number of participants from 24 per group to 36. Second, following the Experiment 1a data indicating that training on syntactically diverse inputs might improve learning (i.e., the BOTH group numerically outperformed ONLY groups on TRAINED structures, despite half the training), we trained participants on two types of relative clauses and tested on a third. To do so, we used dative constructions, from which three types of relative clauses can be formed: SRCs, DORCs, and IORCs.

In addition to being less similar to English, the Experiment 2 grammar was designed to be more typologically typical than that of Experiments 1a and 1b. The clause-initial complementizer *that* was removed, as these are unattested among languages with prenominal relative clauses. The verb-final word order also makes the new grammar more typologically typical: roughly half of languages with verb-final dominant word order have prenominal relative clauses, while only five languages with verb-medial word order (as in English) have prenominal relative clauses (and all of these are dialects of Chinese or languages in close geographic proximity to Chinese-speaking populations; Comrie, 2008). If participants generalize in this grammar, it is unlikely to reflect the use of knowledge of English.

2.4 Experiment 2

In Experiment 2, three groups of participants were trained on subsets of an artificial language that had prenominal relative clauses with verb-final word order. Participants received exposure to combinations of two types of relative clauses: SRCs and DORCs, DORCs and IORCs, or SRCs and IORCs. After training, participants were tested on their knowledge of all three types.

The syntax of the relative clauses in Experiments 1a and 1b had considerable overlap with the syntax of English relative clauses. Generalization to the untrained structure may therefore have reflected reliance on knowledge of English and not a new representation for prenominal relative clauses. Because the relative clause-internal word order in Experiment 2 is different from English, it should be harder for participants to rely on English. Here, then, we again ask whether participants will be able to produce structures that they have not been directly exposed to.

2.4.1 Method

Participants

Participants were continuously run until the target of 36 per group was reached. A total of 136 UC San Diego undergraduates participated for course credit. There were 28 exclusions: 5 for either being native speakers of a non-English language or for having learned Japanese (a language with finite prenominal relative clauses and verb-final word order), 7 due to experimenter error, and an additional 16 who experimenters reported were unable to learn the trained structures during the training phases. No exclusions were made on the basis of performance in the test phase.

Factors

Two factors were manipulated: TRIAL TYPE, which was a within-subjects factor with two levels: TRAINED and UNTRAINED; and GROUP, a between-subjects factor with three levels: SRC+DORC, SRC+IORC, and DORC+IORC.

Procedure

The procedure differed from that of Experiments 1a and 1b in two ways. First, to incrementally introduce novel grammatical properties to participants, we added a training phase to the beginning of the experiment in which participants learned to produce monoclausal sentences with the verb-final word order; 14 of these were monotransitive (i.e., including a subject and a direct object) and 14 were ditransitive (i.e., including a subject, a direct object, and an indirect object). This phase proceeded in the same way as the first training phase in previous experiments: experimenters described two pictures in a row, and then participants saw the two pictures in random order and recalled the descriptions, receiving help in the form of repeated full sentences as needed.

The second difference was in the test phase. Each trial consisted of three parts. First, an image of a ditransitive event appeared on the screen with no arrows. The participant was asked to produce a monoclausal sentence like the ones learned in the first training phase. Experimenters were instructed to help participants as necessary, ensuring that they had produced a monoclausal sentence with the correct meaning and with *Subject–Indirect Object–Direct Object–Verb* word order before proceeding.

Next, participants saw the same image, but with an arrow pointing to either the subject, direct object, or indirect object. Participants were instructed to produce an English translation of the relative clause sentence they would be asked to produce in the next part of the trial, for instance, “Those are the cookies that the grandmother baked for the children.” Experimenters were again instructed to help as needed, and not to go on until the participant had produced a grammatical English relative clause describing the scene. By first asking participants to produce the monoclausal base sentence and then the English relative clause, we ensured that if participants did not produce a well-formed relative clause, it could not be attributed to a misunderstanding of the event or not knowing the base structure of the clause.

Finally, participants saw the same picture with the same arrow a second time and were

asked to produce a sentence in their new grammar describing the person/animal/object the arrow pointed to. Experimenters gave no feedback during this part of the trial. The whole experiment lasted roughly two hours.

Materials

To reduce the length of the study, the number of items was reduced in training phases from 36 to 28. Participants were again tested on 36 new items in the test phase: 12 SRC-eliciting images, 12 DORC-eliciting images, and 12 IORC-eliciting images.

The artificial language in Experiment 2 used English words and had prenominal relative clauses and a verb-final word order. As in Experiments 1a and 1b, the internal structure of nominal elements (including DPs, NPs, and PPs) and all morphology (including verb tense and agreement) were identical to English. Training items contained several plural nouns in various syntactic positions and instances of verbs agreeing with singular and plural subjects to ensure that participants had cues to this effect.

Stimuli in the first training phase consisted of pictures of transitive and ditransitive events accompanied by simple monoclausal descriptions, as in Table 2.6. Word order for transitive events was always *Subject, Direct Object, Verb*. For ditransitive events, it was *Subject–Indirect Object–Direct Object–Verb*. Indirect objects were realized as prepositional objects with either the preposition *to* or *for*, as determined by the verb’s preference in English (e.g., “give cookies *to*” but “bake cookies *for*”).

Stimuli in the second and third training phases consisted of pictures of ditransitive events with an arrow pointing to the subject, direct object, or indirect object. Pictures were paired with sentences with prenominal relative clauses, as in Table 2.7. Indirect object relative clauses were realized with a gap inside the prepositional phrase (so-called *adposition stranding*). In the test phase, stimuli consisted of images paired with verbs; sentences with prenominal relative clauses were the target elicitations.

Table 2.6. Experiment 2: sample stimulus items from Training Phase 1.

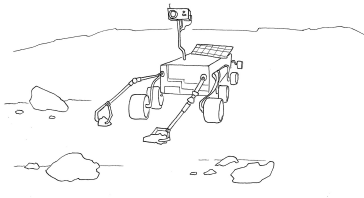

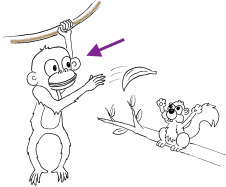
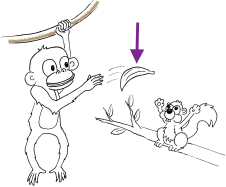
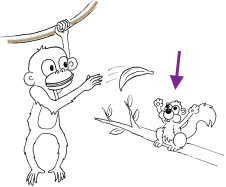



Monotransitive	Ditransitive
 <p>“[A robot] [moon rocks] collects.”</p>	 <p>“[The man] [for the woman] [a rose] bought.”</p>

Table 2.7. Experiment 2: sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.

	SRC	DORC	IORC
Train	 <p>“That’s the [__ to the squirrel the banana threw] monkey.”</p>	 <p>“That’s the [the monkey to the squirrel __ threw] banana.”</p>	 <p>“That’s the [the monkey to __ the banana threw] squirrel.”</p>
Test	 <p>bakes</p>	 <p>bakes</p>	 <p>bakes</p>
Target	<p><i>That’s the [__ for the kids cookies bakes] grandma.</i></p>	<p><i>Those are [the grandma for the kids __ bakes] cookies.</i></p>	<p><i>Those are the [the grandma (for) __ cookies bakes] kids.</i></p>

Data coding and analysis

To be coded as correct, a response had to be a well-formed prenominal relative clause of the elicited type (e.g., a SRC for an image with an arrow pointing to the subject) with the correct meaning. Specific requirements included that the response contained a matrix clause with a subject, verb, and determiner that agreed in number with the head noun (e.g., “These_{pl} are_{pl} the_{pl} [...] cookies_{pl}”); the internal structure of all nominal elements was correct in English (e.g., “the elderly couple”); the prenominal relative clauses were finite and had *Subject–Indirect Object–Direct Object–Verb* word order (allowing for the gap position); and verbs agreed in number with subjects (e.g., “. . . the [the grandma_{sg} for the kids __ bakes_{sg}] cookies”). The most common errors are summarized in Table 2.8.

Table 2.8. Percentage of the most common errors in Experiment 2, by condition.

	SRC+DORC		SRC+IORC		DORC+IORC		Overall
	Trained	Untrained	Trained	Untrained	Trained	Untrained	
Constituent order error (“. . . <i>the [the banana to the squirrel threw] monkey.</i> ”)	17.13	23.15	12.50	40.24	3.36	40.97	18.90
Missing preposition* in IORC (“. . . <i>the [the squirrel the banana threw] monkey.</i> ”)	5.32	63.43	8.10	6.43	2.27	3.24	11.73
Repeated head determiner (“. . . <i>the [the squirrel the banana threw] the monkey.</i> ”)	9.38	14.35	7.26	11.19	5.44	5.56	8.36
Missing head determiner (<i>That’s [the squirrel the banana threw] monkey.</i>)	7.52	8.56	6.19	6.19	5.90	3.47	6.39
Number agreement error (“ <i>Here’s a [the grandma for the kids baked] cookies.</i> ”)	3.24	4.86	3.57	3.57	7.29	3.01	4.41
Wrong type (e.g., well-formed SRC in response to a DORC-elicitation)	2.20	2.31	2.74	7.38	2.66	12.04	4.10

* While these were different from the input grammar, they are counted as correct in all Exp. 2 analyses. *Note.* Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported.

The high number of missing prepositions in IORCs may reflect a reasonable strategy for relativizing a prepositional argument. Indeed, adposition stranding is exceedingly rare cross-linguistically. (Hungarian is, to our knowledge, the only language outside the Germanic family

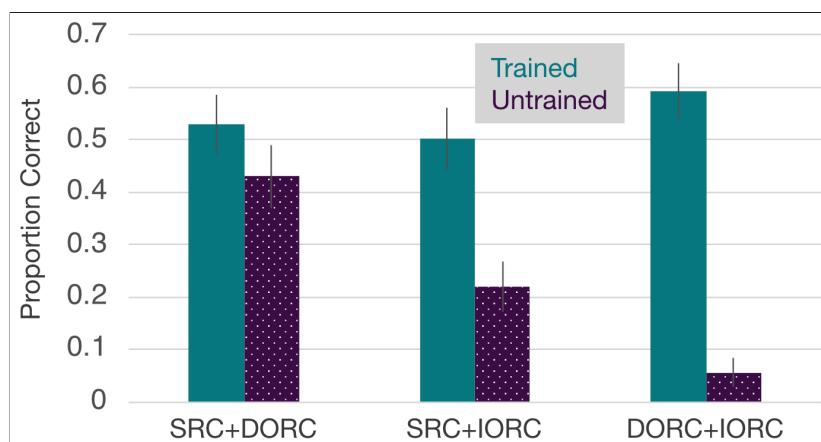


Figure 2.5. Proportion of well-formed productions of the elicited structure as a function of GROUP and TRIAL TYPE in Experiment 2.

that allows this; Marácz, 1984.) A more common strategy is to simply delete the preposition, something which is done in a number of prenominal relative clause languages across families (e.g., Akhvakh, Evenki, Korean, Malayalam, Conchucos Quechua, etc.; Wu, 2011). We therefore coded IORCs as correct with either missing noun phrases or missing prepositional phrases.

2.4.2 Results

Results are depicted in Figure 2.5 and summarized in Table 2.9. The converging model contained random intercepts for participants and items and a random slope for TRIAL TYPE that varied within participants. There were no significant differences in how well the training groups learned the TRAINED structures. Untrained structures, however, were produced significantly less well than trained structures ($\chi^2(1) = 5.692, p = .017$). The SRC+DORC group produced more UNTRAINED structures than either of the other two groups (model comparison confirmed that the interaction term contributed to overall model fit: $\chi^2(1) = 56.349, p < .001$). To determine whether the SRC+IORC group also produced more UNTRAINED structures than the DORC+IORC group, we performed one additional pairwise comparison with a two-tailed test of equal proportions; this test confirmed the difference ($\chi^2(1) = 47.064, \text{Bonferroni-corrected-} p < .001$).

Table 2.9. Experiment 2 results.

	Model results			
	β	z	p	
<i>Model 1: GROUP × TRIAL TYPE</i>				
Intercept	0.088	0.229	.819	
GROUP: SRC+IORC	−0.249	−0.460	.646	
GROUP: DORC+IORC	0.214	0.460	.689	
TRIAL TYPE: UNTRAINED	−0.810	−2.406	.016	*
Interaction: SRC+IORC, UNTRAINED	−1.694	−3.334	< .001	***
Interaction: DORC+IORC, UNTRAINED	−4.606	−7.418	< .001	***
<i>Model 2: UNTRAINED trials, split by performance on TRAINED</i>				
Intercept	−2.853	−4.775	< .001	***
GROUP: SRC+IORC	−1.752	−1.854	.064	.
GROUP: DORC+IORC	−21.12	−0.001	.999	
SUBSET: ≥ 50	3.852	5.065	< .001	***
Interaction: SRC+IORC, UNTRAINED	0.033	0.028	.978	
Interaction: DORC+IORC, UNTRAINED	15.83	0.001	.999	

To determine whether production of untrained structures reflected chance performance, we again compared participants who produced more than 50% of trained structures to those who produced fewer (this relationship is shown in Figure 2.6). The converging model contained random intercepts for participants and items and a random slope for group that was allowed to vary within participants. A significant effect of SUBSET confirmed that the 64 higher-performing participants generalized more (SRC+DORC: 18%, SRC+IORC: 16%, DORC+IORC: 6%) than the 43 lower-performing participants (SRC+DORC: 75%, SRC+IORC: 67%, DORC+IORC: 53%; model comparison: $\chi^2(1) = 22.732, p < .001$). There was no statistical evidence that this effect varied by group. This again indicates that generalization reflects above-chance performance.

2.4.3 Discussion

The grammar of the artificial language in Experiment 2 differed from English both in the prenominal position of the relative clause and in the basic word order. As such, participants were unlikely to be able to use knowledge of English to generalize to UNTRAINED structures. We nevertheless replicated the finding from Experiments 1a and 1b that participants produced

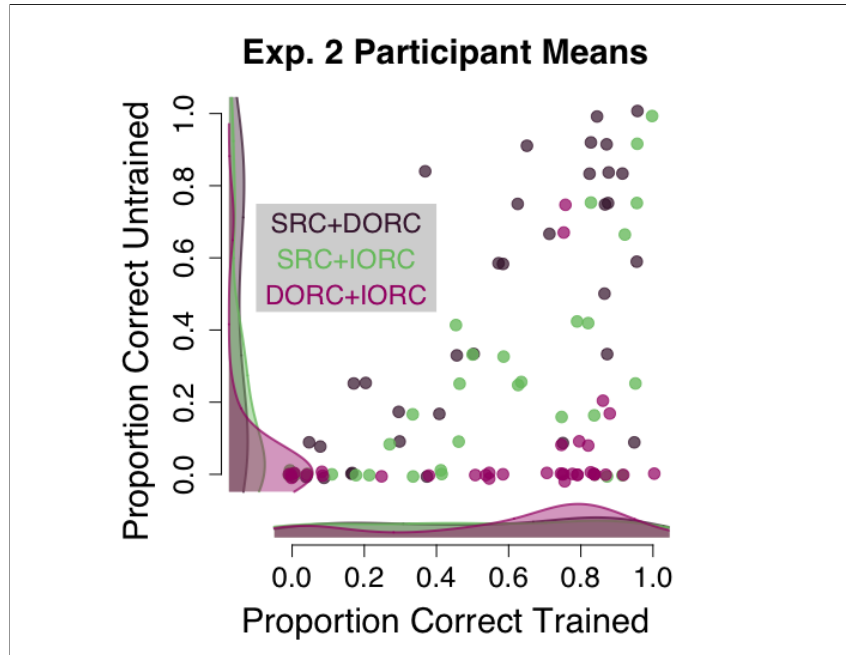


Figure 2.6. Experiment 2: The relationship between learning trained structures and generalization to untrained structures.

structures they had not been directly exposed to. This supports the idea of a single, general representation of relative clauses.

The pattern of generalization in Experiment 2, however, was unexpected. Participants in the SRC+DORC group generalized to IORCs and participants in the SRC+IORC group generalized to DORCs, but very few participants in the DORC+IORC generalized to SRCs. To determine whether this might reflect some specific feature of the grammar in Experiment 2, Experiment 3 was designed to replicate Experiment 2 using a different grammar. To simplify the design, we did not include an SRC+IORC group.

2.5 Experiment 3

Experiment 3 aimed to replicate the unexpected pattern of generalization displayed by the SRC+DORC and DORC+IORC groups in Experiment 2. The grammar was a modified version of the Experiment 2 grammar using bare nouns (i.e., with no determiners) and case marking, or the use of suffixes to specify whether a noun is a subject, direct object, indirect object, oblique

object, etc. (as in, e.g., Korean). In a continued effort to adhere to typological typicality and to reduce reliance on knowledge of English.

2.5.1 Method

Participants

Participants were continuously run until a pre-determined stop date or the target of 36 per group was reached. Because the stop date was reached prior to reaching the target number, a total of 45 participants were run. Data from 3 participants were excluded: 1 for early childhood exposure to Chinese (which has prenominal relative clauses), 1 for attempting to explicitly describe the grammar aloud throughout the training phases, and 1 who was unable to learn the trained structures. No exclusions were made on the basis of performance in the test phase. The final data set included 21 participants in each group.

Factors

Two factors were manipulated: TRIAL TYPE, which was a within-subjects factor with two levels, TRAINED and UNTRAINED; and training GROUP, a between-subjects factor with two levels: SRC+DORC and DORC+IORC.

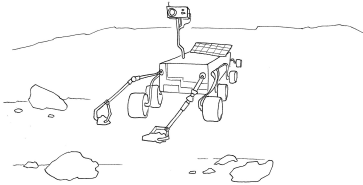

Materials

In response to experimenters reporting that participants appeared to fatigue toward the end of the training phases, we again shortened the experiment by reducing the number of training items from 28 to 20: 10 of each of the trained structures. Participants were again tested on 36 new items, 12 of which elicited SRCs, 12 DORCs, and 12 IORCs.

The grammar of the artificial language in Experiment 3 was a modified version of the Experiment 2 grammar. It had no determiners, no prepositions, and included nominative, accusative, and dative case marking suffixes (although the head noun was not marked for case, consistent with the way case-marking languages like Korean and Japanese treat the object of the verb *be*). These changes removed many of the remaining components of English word order

from the relative clause. The grammar was more typologically typical in that it included case marking and excluded prepositions, both of which are common features of verb-final languages. Sample items appear in Tables 2.6 and 2.7.

Table 2.10. Experiment 3: Sample stimulus items from Training Phase 1.

Monotransitive	Ditransitive
	
<p>“Robot-uh rocks-en collects.”</p>	<p>“Man-uh woman-ik rose-en buys.”</p>

Procedure

The procedure was identical to that of Experiment 2. The whole experiment took roughly 90 minutes.

Data coding

Responses were coded as correct if they contained a prenominal relative clause that conformed to the input grammar. Specific criteria included that nouns inside the relative clause had the correct case markers and appeared in the correct order: *Subject–Indirect Object–Direct Object–Verb* (modulo the missing noun). The head noun was not allowed to have case marking. The most frequent errors are reported in Table 2.12.

2.5.2 Results

Data are shown in Figure 2.7 and models are reported in Table 2.13. The main analysis, Model 1, converged with random intercepts for items and participants, no random effects correlations, a random slope for TRIAL TYPE that varied within participants, and random slopes

Table 2.11. Experiment 3: Sample relative clause stimulus items from Training Phases 2 and 3 and the Test Phase.

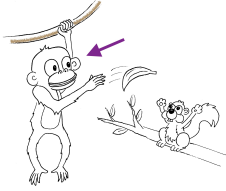
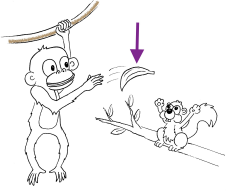
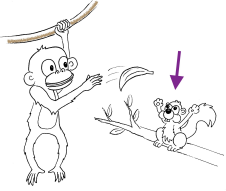

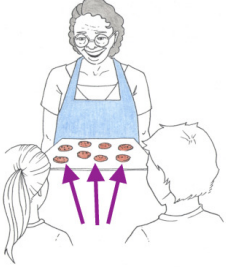

	SRC	DORC	IORC
Train	 <p>“Here’s this [squirrel-ik banana-en throws] monkey.”</p>	 <p>“Here’s this [monkey-uh squirrel-ik __ throws] banana.”</p>	 <p>“Here’s this [monkey-uh __ banana-en throws] squirrel.”</p>
Test	 <p>bakes</p>	 <p>bakes</p>	 <p>bakes</p>
Target	<p><i>Here’s this [__ kids-ik cookies-en bakes] grandma.</i></p>	<p><i>Here are these [grandma-uh kids-ik __ bakes] cookies.</i></p>	<p><i>Here are these [grandma-uh __ cookies-en bakes] kids.</i></p>

Table 2.12. Percentage of the most common errors in Experiment 3, by condition.

	SRC+DORC		DORC+IORC		Overall
	Trained	Untrained	Trained	Untrained	
Constituent order error (“... <i>this [banana-en squirrel-ik throws] monkey.</i> ”)	36.51	29.71	23.41	27.38	29.50
Wrong case marker (“... <i>this [squirrel-uh banana-en throws] monkey.</i> ”)	7.54	17.46	3.17	15.87	9.13
Missing case marker (“... <i>this [squirrel-ik banana throws] monkey.</i> ”)	12.70	16.67	2.58	5.95	8.86
Case marker on head noun (“... <i>this [squirrel-ik banana-en throws] monkey-uh.</i> ”)	6.94	10.32	3.17	10.32	6.81
Number agreement error (“ <i>Here’s this [grandma-uh kids-ik bakes] cookies.</i> ”)	2.18	6.35	4.17	1.19	3.37

Note. Errors were not mutually exclusive. Errors from responses that had three or more errors are not reported in this table.

for TRIAL TYPE and the TRIAL TYPE \times GROUP interaction that varied within items. Neither the main effect of GROUP nor the main effect of TRIAL TYPE were significant, but their interaction was ($\chi^2(1) = 7.173, p = .007$), reflecting the fact that the SRC+DORC group generalized more than the DORC+IORC group.

Table 2.13. Experiment 3 results.

	Model results			
	β	z	p	
<i>Model 1: GROUP \times TRIAL TYPE</i>				
Intercept	-0.367	-0.513	.609	
GROUP: DORC+IORC	0.280	0.281	.779	
TRIAL TYPE: UNTRAINED	-1.238	-1.435	.151	
Interaction	-3.295	-2.624	.009	**
<i>Model 2: UNTRAINED trials, split by performance on TRAINED</i>				
Intercept	-6.681	-3.268	.001	**
GROUP: DORC+IORC	-1.300	-0.598	.550	
SUBSET: $\leq 50\%$	9.421	2.963	.003	**
Interaction	-5.979	-1.309	.191	

The relationship between learning and generalization across participants is shown in Figure 2.8. We compared the amount of generalization among the 20 learners who produced

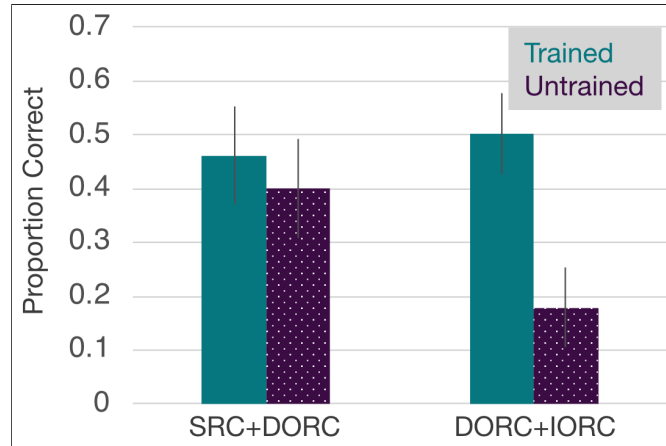


Figure 2.7. Proportion of well-formed productions of the elicited type as a function of GROUP and TRIAL TYPE in Experiment 3.

more than 50% correct trained structures to the 22 who did not. The model, Model 2 in Table 2.13, converged with a full random effects structure but no random effects correlations. Higher-performing participants generalized more (81% in the SRC+DORC group and 64% in the DORC+IORC group) than lower-performing participants (11% in the SRC+DORC group and 17% in the DORC+IORC group; $\chi^2(1) = 12.78, p < .001$), again indicating that generalization reflects above-chance performance.

2.5.3 Discussion

Experiment 3 replicated the findings of previous experiments in that participants produced the UNTRAINED structures. It furthermore replicated the specific pattern of generalization observed in Experiment 2, whereby SRC+DORC participants generalized to IORCs more than DORC+IORC participants generalized to SRCs. Across groups, generalization was higher for participants that learned the trained structure better, indicating that generalization, at least among the highest performing participants, does not reflect chance performance.

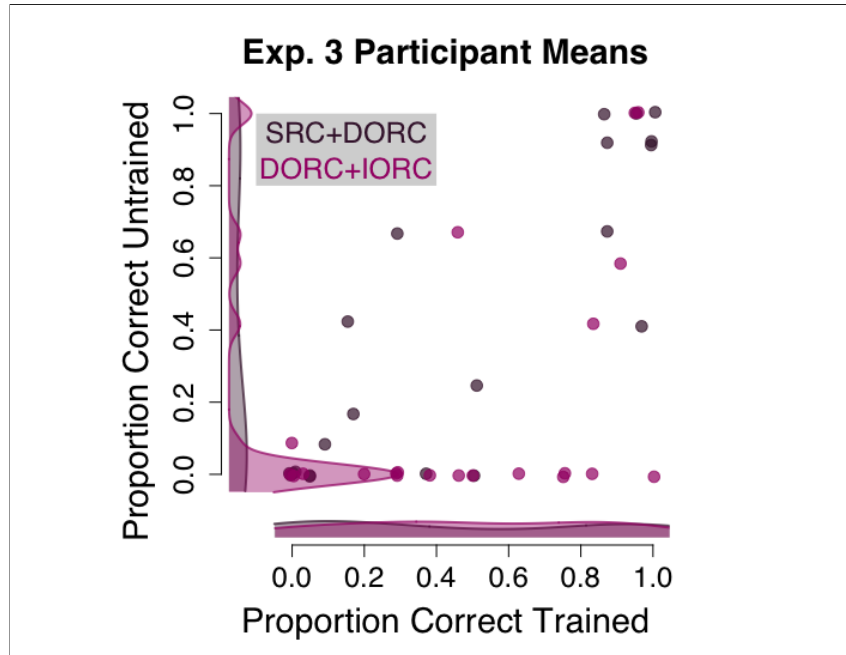


Figure 2.8. Experiment 3: The relationship between learning trained structures and generalization to untrained structures.

2.6 General Discussion

Our aim here has been to discern between two proposals for the representation of long-distance dependencies. To do so, we asked whether participants trained on one or two types of relative clause would also learn other types, despite not having had direct exposure to them. If not, it may be an indicator that each type of relative clause is represented independently. But if so, it would support a model of relative clauses as having the same underlying representation: something like a general principle according to which the head noun is not repeated inside the relative clause.

In four experiments, participants were implicitly trained on a new grammar for relative clauses. Experiments 1a and 1b provided the first evidence that learners of one type of relative clause also learn other types. In Experiment 1a, one group received an explicit grammar lesson rather than implicit training. At test, this group performed worse than the implicit groups, indicating that the implicit groups' productions were not guided by explicit representations.

However, the conclusion that participants learned untrained structures is called into question by the fact that many passive relative clauses were produced, perhaps indicating that participants used knowledge of English syntax to perform the task.

Experiments 2 and 3 resolved this potential confound by teaching grammars that were even more different from English so as to remove the possibility that generalization might reflect the use of English knowledge. Both experiments showed evidence for generalization. This finding implies that when the syntax of a relative clause is learned, it is not a representation of a specific structure, but a more abstract representation of relative clauses in general.

If knowledge of one relative clause is knowledge of all relative clauses, then we might have expected performance to be equally good on untrained structures and trained structures. However, across experiments, participants consistently performed better on trained structures than untrained ones. This indicates that these structures were still in some sense novel to participants.

There are a number of possibilities for why this might be. One is that it reflects a deficit in performance and not in competence. That is, even if participants have a general representation for relative clauses, they may need to practice mapping from particular types of semantic representations to the syntactic representation. For example, participants trained on SRCs receive a good deal of practice relativizing agents. But relativizing patients, as in many of our DORC stimuli, may also require practice, which SRC-ONLY participants received far less of.

Another intriguing possibility is that relative clauses may be simultaneously represented in multiple ways. That is, it is possible that learners acquire both a general representation, as we have argued on the basis of generalization, and a specific representation that directly reflects the particular surface word order of each type.

It has long been noted that general and specific representations are not mutually exclusive (Goldberg, 1995). Idioms, for instance, must have both compositional and noncompositional representations. Rachel can “pull Dan’s leg” – or joke around with Dan – and Dan can “pull Rachel’s leg” – Dan can joke around with Rachel. If these had no compositional structure, then

each would have to be learned as a separate idiom to be correctly interpreted. But if they had no noncompositional representation, then they would be about two people tugging on limbs. It is possible that the difficulty associated with producing untrained structures may reflect the difficulty associated with deriving a specific representation from the general representation. (This may not in fact be different from the performance/competence possibility mentioned above.)

Whatever the reason for the lower performance on untrained structures, there is still a need for a general representation of relative clauses to account for production of untrained structures in the present study. The idea of such a representation is not new. Indeed, Chomsky (1959) cited the need for general representations of long-distance dependencies in his famous rebuke of behaviorism, and several theoretical frameworks attempting to formalize this idea have since been developed. For example, *derivational* models account for long-distance dependencies with operations that “move” or “delete” noun phrases (e.g., Chomsky, 1981, 1992), while more surface-oriented models include phonologically null lexical elements called “slash-categories” that share features with disparate parts of the syntactic representation (e.g., Pollard and Sag, 1994; Culicover and Jackendoff, 2005).

To the best of our knowledge, all prior evidence for a general representation of long-distance dependencies has been inferred from linguistic observations such as the similarities between the various types of English relative clauses. Such data have the benefit of reflecting knowledge obtained in the most naturalistic ways possible: they model the grammars of native speakers. But it is also impossible to disentangle various influences on language acquisition in such subjects. As noted earlier, even very young children have probably been exposed to all types of relative clauses in their language. As such, it cannot be determined whether productive use of a given structure reflects generalization on the one hand, or reliance on direct experience on the other (although see Montag and MacDonald, 2015 for compelling evidence favoring direct experience)

The present study supplements the linguistic data with experimental data. The experiments having been carefully controlled, confounds such as the potential for exposure to other

types of relative clauses are minimized and the observed generalization to untrained types constitutes strong evidence for the existence of a general underlying representation.

2.6.1 Implications for theories of language acquisition

The present findings pose a problem for some prominent theories of language acquisition. In particular, *experience-based* (or sometimes *usage-based*) theories of acquisition explain learning as a process of first memorizing particular *items* (or *exemplars*), and then abstracting over these to arrive at a system of structural representations (e.g., Tomasello, 2000a, 2009; McCauley and Christiansen, 2011; Christiansen and Chater, 2015). These theories are robust to a number of observations that the traditional picture of a purely abstract system struggles to accommodate. For example, to account for idiosyncratic argument structures – why one must use a preposition with the verb *dine*, as in “dine *on* steak,” but not with the semantically similar verbs *eat* or *devour* – some item-based knowledge is clearly required.

However, if the acquisition of a representation reflects a process of abstraction over exemplars, it should not be possible to acquire a structure without direct exposure to it. But that is exactly what participants in our study did. One might argue (as we have) that if all relative clauses are represented with a single, general representation, then exposure to one type is exposure to all types, and the UNTRAINED structures were therefore not, in fact, untrained. This may certainly be the case, but it is not clear that experience-based theories can account for the acquisition of such a general representation.

The proposed learning mechanism in experience-based theories results in surface-oriented syntactic representations. That is, if a learning mechanism derives syntactic representations by finding similarities among the surface word orders of exemplars, abstracting away from specific words, then the resulting syntactic representations should map straightforwardly onto surface word order. In such an approach, it might be possible to represent a relative clause like “the dog [that I fed]” with a Phrase Structure-like rule of the form:

(14) *Relative Clause* → “*that*” *Subject Verb*.

But note that this rule only can only account for DORCs (e.g., “the bear [that the girl hugged __]”). Knowledge of SRCs (e.g., “the girl [that __ hugged the bear]”) would require another rule, something like:

(15) *Subject Relative Clause* → “that” *Verb Direct-Object*.

A family of structural representations would therefore be required to account for all the various forms of relative clauses. But this is inconsistent with the idea of a single, general representation, which is necessary to account for the type of generalization we have demonstrated.

While there are some attempts at modeling relative clause acquisition in experience-based theories, none resolve the issue outlined here. In some cases, these models amount to little more than arguments that children’s relative clauses are not as complex as adult relative clauses (e.g., Tomasello, 2009). Evidence suggests that this is true, but this observation alone is not explanatory. In a more elaborated account, Fitz et al. (2011) describe a computational model that learns to produce relative clauses. But this model starts out with a hardwired mechanism specifically for tracking the role of the relativized noun inside and outside the relative clause, thereby assuming much of the knowledge a priori rather than learning it.

To be clear, we do not wish to argue that a general representation for relative clauses is necessary for acquiring relative clauses. Indeed, we believe that it is likely that specific representations are learned from exposure, and furthermore that these representations are probably easier to use than general ones (as indexed by the lower performance on untrained structures across the present experiments). But our data do indicate that in addition to these representations, a general representation also exists. Theories of acquisition must be able to account for this type of representation.

One possible way forward for experience-based theories may be to stipulate that the abstraction mechanism that derives hierarchical structures is more general than currently characterized. Rather than only abstracting over strings of words, it may also abstract over other linguistic representations like the representations in (14) and (15). The general representation

of relative clauses that we have provided evidence for may well be derived from a process that starts with noting the similarity among many individual types of relative clauses.

How such a theory might explain the generation of the particular representational apparatus that allows such representations is not clear. But this is not a new problem for these theories. Such questions must also be answered for the generation of representations of hierarchical structures. (Indeed, this is a subset of a larger problem in cognitive science, where it remains unclear how any knowledge that is not straightforwardly innate or perceptual can be generated; Fodor, 1975.)

2.6.2 Asymmetries in generalization

One puzzling feature of our data was that in two experiments, participants trained on DORCs and IORCs did not generalize to SRCs as much as participants trained on SRCs and DORCs generalized to IORCs. This is particularly surprising in light of a typological contingency hierarchy, the Accessibility Hierarchy, according to which the existence of DORCs and IORCs in a language implicates the existence of SRCs, but the existence of SRCs and DORCs does not necessarily implicate the existence of IORCs (Keenan and Comrie, 1977). That is, there are no languages that have DORCs and IORCs but do not have SRCs, but there are languages (e.g., Hebrew and Standard Arabic) which have SRCs and DORCs, but not IORCs.

This is the opposite of the pattern that we observe. Interestingly, a finding similar to ours is also reported by Yip and Matthews (2007), who documented the acquisition of relative clauses in Cantonese-English bilingual children. Curiously, the first relative clauses these children acquired in *both* languages were *prenominal* DORCs.

The Accessibility Hierarchy has long been thought to reflect processing difficulty (Keenan and Comrie, 1977; Keenan and Hawkins, 1987). That is, SRCs are easier to process than DORCs and IORCs (Keenan and Hawkins, 1987; Diessel and Tomasello, 2005; Cook, 1975; Clemens et al., 2015; Wagers et al., 2018; Hatch, 1971; Kwon et al., 2010), suggesting that the hierarchy may reflect the relative difficulty of the various types of relative clauses played out over the

course of language evolution. This would be consistent with findings from the statistical learning literature that suggest that easier patterns come along for the ride when learners are trained on harder patterns (Thompson and Newport, 2007).

However, this cannot explain the pattern we observe in Experiments 2 and 3. Indeed, it predicts the opposite of what we observed. But it may be possible for difficulty to have different effects depending on the particular task. For instance, diachronically, it may lead to attrition of more difficult structures in a language. In synchronic language learning, however, it is possible that the increased difficulty of the DORC+IORC groups' training materials led to a stronger focus on acquiring specific representations, but ultimately a weaker general representation of relative clauses (if in fact these two types of representation coexist).

2.7 Conclusion

We began by pointing out that there are two ways in which the syntax of long-distance dependencies might be represented: either as a family of independent representations with similar properties, or as a single general representation. While this latter possibility may seem like an unnecessarily baroque type of representation to posit, it is in fact the way most theoretical accounts of syntax model long-distance dependencies.

Here, we present four experiments which carefully control exposure such that learners have direct experience with some types of relative clauses, but not with others. In spite of this, they are able to produce structures that they have never been exposed to, indicating that they have access to a general representation.

A potentially beneficial avenue for future research might be to better identify exactly how learners' knowledge of English influenced their performance in the tasks reported here. There are many possible approaches to this, but we suspect that a particularly fruitful approach will be to repeat the present task with speakers of languages like Palestinian Arabic, which does not have DORCs or IORCs (Shlonsky, 1992), or with speakers of languages like Diyari or Bambara,

whose relative clauses do not have gaps at all (Dryer, 2013). Of course, the truest test may come from work with children, although it is hard to imagine how one might reasonably (and ethically) perform such an experiment.

Chapter 3

Individual differences reveal gradience in syntactic representations

3.1 Introduction

Humans have a knack for abstraction: we can discuss philosophical questions like whether ‘tis nobler “to suffer the slings and arrows of outrageous misfortune, or to take arms against a sea of troubles” without needing to specify exactly what troubles or which arms (Shakespeare, 1996). We can discern a landscape in Helen Frankenthaler’s 1952 abstract painting “Mountains and Sea” (or so some say). And we can string words like *dog* and *bite* and *mailman* together using rules that refer not to specific words, but to abstractions like nouns and verbs.

Abstraction is also at the heart of symbolic systems. For instance, working memory has been modeled as three to four abstract “slots” (Luck and Vogel, 1997). Phonological rules govern the combinatorics of phonemes, such that abstract features like nasalization or voicing can spread from phoneme to phoneme. And syntax governs the combinatorics of words such that “Colorless green ideas sleep furiously” is a grammatical sentence, but “Sleep parrots bright green soundly,” is not – even if more readily lent to interpretation (Chomsky and Lightfoot, 2002).

Recently, however, many subfields include approaches that move away from purely abstract characterizations. Evidence has surfaced that representations which were once thought to be fixed, immutable, and purely abstract seem to be just the opposite: malleable and defined in part by experience. For instance, after exposure to altered phonemes, like an /s/ with a

slightly below normal peak frequency, participants' productions of /s/ become lower, as does their perceptual boundary between /s/ and /ʃ/ (Norris et al., 2003). Similarly, words that are more frequent can be accessed more quickly, and infrequent syntactic structures become more acceptable after repeated exposure (Snyder, 2000).

In response to these findings, models in some fields represent linguistic features in a continuous space, capturing gradient properties of what were initially assumed to be categorical symbols. Pierrehumbert (2001), for example, argues that phonetic representations are *item-based* – dependent on an exemplar space that is updated with each new exposure.

The generalization that emerges is that symbols can (and often do) have a gradient nature, but the machinery that coordinates them is still generally thought to be abstract. But if we divide symbolic systems into symbols and systems, it is not clear where systems like syntax and phonology belong.

On one hand, syntax seems more like the abstract machinery: it is used to coordinate symbols like nouns and verbs to make bigger symbols like phrases, and then to coordinate those into bigger symbols like sentences. Even the language that is used to describe syntax – analogies like “frame” and “scaffolding” and “argument slot” – invoke the idea of abstraction. From this perspective, it seems much more like system than symbol.

But while this characterization successfully accounts for many behavioral phenomena (e.g., linguistic productivity), there are a number of cases where a purely abstract model of syntax cannot account for behavior. This is clear to anyone who has attempted to learn a foreign language: idiosyncrasies abound which require item-specific knowledge. For example, one can “give a toy to a boy” (the *prepositional dative* construction) or “give a boy a toy” (the *double object* construction), but while you can “donate a koi to a goy,” you cannot “donate a goy a koi.”

Idioms are similarly problematic. For example, *letting the cat out of the bag* does not have anything to do with cats or bags. Specific knowledge about particular strings of words is sometimes necessary for comprehension. Relying on syntax to understand this sentence will lead to a very different interpretation.

A number of other properties of syntax suggest a symbolic nature. Whole argument structures can be primed with subliminal presentation of a single verb (Trueswell and Kim, 1998); syntactic representations can be coordinated (“[[the cat] *and* [the dog]] chased...”); and, at least according to some theories, they can even be systematically manipulated with operations to derive strings like “the mailman that the dog chased” from “the dog chased the mailman.”

Syntax thus has a sort of Heisenbergian dual nature, behaving like an architecture from some vantage points but like a symbol from others. While some degree of abstraction is necessary for productivity, some degree of specificity is also clearly necessary. But if there is a part of the syntactic representation that is more like a symbol, then perhaps syntax, like phonemes and words, exists in an exemplar space. Our goals here are to determine whether there are in fact gradient properties of syntactic representations, to develop a method for detecting and quantifying them, and to determine whether this gradience reflects differences in experience.

3.1.1 Individual differences

Gradient differences in syntactic representations represent a departure from the standard way of thinking about syntax. It is often assumed that native speakers of the same language have the same grammars. That is, in spite of individuals’ varied experiences with these structures, the representations should not vary. (See Dabrowska, 2012 for a number of specific instances, particularly in the developmental literature, where this assumption is made explicit.)

Indeed, it is hard to even imagine an alternative. How could a rule like “object before verb” possibly be continuous? The object is either before or after. A verb either agrees with a number feature or it does not. While phonemes can be fully represented in a continuous space, syntactic representations are more categorical.

Some research shows that such categorical individual differences do exist. Dabrowska (2008) reports a task where Polish-speaking teenagers were asked to use nonce words in contexts requiring them to add a genitive suffix. In Polish, however, which particular genitive suffix to use is conditioned on a complex set of grammatical and semantic properties. Dabrowska showed

that participants' choice between two suffixes formed a bimodal distribution. Participants in her study, despite being native Polish speakers had acquired two different rules: for some, gender was the only determining factor, but for others, gender and animacy interacted.

But as for gradient differences, the evidence is suggestive, but not quite as convincing. This is largely due to methodological problems: individual differences are notoriously difficult to measure. Individual differences in reading times in a language comprehension task, for instance, have been claimed to reflect individual differences in processing (Kidd et al., 2018). But noise being present in any behavioral measure, it is important to know whether such findings are reliable.

Even when reliable individual differences are identified, it is often not possible to attribute them to any particular source. For instance, reading time differences could reflect differences in language processing, as Kidd et al. argue; but they could also derive from differences in visual processing, attention, or even different computer hardware (Enochson and Culbertson, 2015).

Chipere (2001) and Dabrowska and Street (2006) both report experimental findings indicating that while native English speakers accurately comprehend active-voice sentences across education levels, a group with fewer years of formal education comprehend passive-voice sentences less well than a group with more formal education. Dabrowska and Street take the ceiling performance in both groups on active sentences as evidence against an explanation relying on processing differences. Instead, they attribute their findings to a less “entrenched passive schema,” the result of less experience with passive structures.

However, Chipere notes that participants in the low academic achievement group reported an “aha” moment when the correct way to interpret passive structures was explained to them, and their performance in a subsequent comprehension task was no longer different from the high academic achievement group. This suggests that if the difference between groups was in the syntactic representation of passives, then the groups went from not having a representation prior to explicit training to having a representation afterwards. This seems extremely unlikely.

Regardless, an “aha” moment seems to reflect a categorical change, not a gradient one

like we aim to investigate. Furthermore, Dabrowska and Street report that within their low academic achievement group, 10 of the 32 participants were at ceiling on comprehension of passives, 14 were above chance, and 8 were at or below chance. This points toward a potential attentional explanation. In particular, if the 8 participants who were at chance were removed from the analysis, it is possible that the low and high academic achievement groups would no longer significantly differ.

As with Kidd et al.'s 2018 study, the difficulty here is in interpreting individual differences from a single task. One way of increasing validity when looking at data from a single task is with a *split-half* analysis. For instance, a common way to measure syntactic representations is with acceptability judgment tasks in which participants rate the subjective acceptability of sentences on a Likert-like scale. Grammatical sentences tend to be rated high, and ungrammatical sentences low. To determine whether there are reliable individual differences in acceptability judgment data, one could split each participant's data into two sets of ratings: those from odd-numbered trials and those from even-numbered trials. If the mean of participants' scores on odd trials correlate with their means on even trials, then there are reliable individual differences.

Split-half analyses are therefore a useful tool for determining whether individual differences reflect noise or something more systematic. But because there are a number of task-specific sources of variance that are the same from one half of the task to the next, these analyses do not shed light on which particular sources account for the differences. For acceptability ratings, individual differences may reflect different syntactic representations among participants. But they may also reflect differences in any other process involved in acceptability judgment, such as lexical access, recognition memory, parsing strategies, perceptual restoration, or differences in the way participants map from a subjective acceptability scale to a Likert scale – one person's 4 may be another's 5.

Here we aimed to more precisely isolate syntactic representations by comparing data from two behaviors that share syntactic representations, but little else. Specifically, we measured participants' acceptability judgments and production choices for four structures. This approach

eliminates many unintended sources of intra-individual reliability, such as idiosyncratic strategies for mapping from acceptability to a Likert scale. Such effects are only present in one of the two measures, so they cannot be the cause of any observed correlations. Processes and representations that are not shared by both production and acceptability judgment behaviors will not contribute to the reliability metric.

Processes and representations that *are* shared, however, will be revealed by correlations between the two behavioral measures. There may be a few such things in addition to syntactic representations, including lexical access, frequency, and the weight given to pragmatic content. We consider such alternatives in the General Discussion.

3.1.2 The present study

Here we aim to understand whether syntactic representations, like phonetic representations, exist in an exemplar space. To do so, we compared how often people produce a particular structure to how much they prefer that structure. Specifically, we constructed stimuli using *structural alternates*: pairs of structures which convey the same meaning.¹ For instance, *dative* constructions can be formulated in two ways. One could use the *double object* (DO) dative construction: “She gave the teacher an apple,” an utterance with two NPs after the verb. But the speaker could have also used the *prepositional dative* (PD) construction: “She gave an apple to the teacher.” This utterance has an NP followed by a PP after the verb. In spite of having different structures, these two utterances have the same meaning (at least at some level of coarseness).

We asked participants to produce sentences like these, which implicitly required them to choose one of the two structural alternates. We also asked participants to rate the acceptability of sentences, some of which had the DO structure, and others the PD structure. We defined a

¹It is unlikely that different structures ever convey *exactly* the same meaning. For example, the DO dative alternate is said to convey that the theme was successfully transferred from the agent to the goal, and not simply that the action was begun and possibly completed. Similar differences may also exist between the other three alternates. Such differences tend to be small and therefore seem unlikely to play a large role in the present task, but it is in principle possible that individual differences may to some degree reflect individual differences in these form-to-meaning mappings, and not in the structural representations themselves.

participant's *preference* for a structure as the their mean acceptability rating for that structure minus their mean acceptability rating for the alternate structure.

We then compared how often participants chose to produce a particular alternate and how much they preferred that alternate. That is, we asked whether the same speakers who produce the DO structure more often than other speakers also prefer the DO structure more than other speakers. If so, it suggests that some property of the structural representation, which we will refer to as its *strength*, varies from speaker to speaker. Speakers with stronger representations of the DO structure judge DO constructions as more acceptable and produce them more often.

We used four types of structural alternates in this study: DATIVE, LOCATIVE, WH-ISLAND, and AGREEMENT, summarized in Table 3.1. Each type has different properties, which was intended as a way for us to disentangle individual differences resulting from grammatical and extra-grammatical properties. Here we will briefly discuss these properties and our specific predictions for each structure.

In dative constructions, the verb takes three arguments: an agent (the person who gives; usually the subject), a theme (the thing being given; usually the direct object), and a goal (the person who the theme is being given to; usually the indirect object). The theme is always an NP, but the goal can either be an NP that precedes the theme or a PP that follows it. This is a true syntactic difference, involving different word orders and different parts of speech. If we observe individual differences for datives, the differences likely reflect differences in the strength of the structural representations.

In locative constructions, the verb takes an agent, a theme, and a location argument. While the theme and location may appear in either order after the verb, the first argument is always an NP and the second is always a PP. When the goal appears in the PP, the preposition is usually *on*, and this alternate is called an “on locative.” When the theme appears in the PP, the preposition is usually *with*, resulting in a “with locative.” Given that both alternates have the same order of syntactic constituents – an NP followed by a PP – the underlying structural representation may be the same (although this depends on the particular theoretical framework). What certainly

Table 3.1. Examples of each of the four structural alternates we used in our stimuli.

STRUCTURE CLASS	
<i>structural alternate</i>	Example
DATIVE	
<i>PD</i>	We all believed that the coach taught the plan to the players.
<i>DO</i>	We all believed that the coach taught the players the plan.
LOCATIVE	
<i>with</i>	Michael groaned watching the chef drizzle the cake with chocolate.
<i>on</i>	Michael groaned watching the chef drizzle chocolate on the cake.
WH-ISLAND	
<i>gap</i>	There goes the little girl that the teacher understood why I scolded __ after school.
<i>pronoun</i>	There goes the little girl that the teacher understood why I scolded her after school.
AGREEMENT	
<i>singular</i>	The farmer was frustrated; the gate to the pastures was gradually falling down due to disrepair.
<i>plural</i>	The farmer was frustrated; the gate to the pastures were gradually falling down due to disrepair.

Note. The sentences above are also sample stimuli from the Experiment 1 acceptability phase, which are the same as the target sentences from the Experiment 1 production stimuli given in Table 3.2.

differs is the mapping from thematic roles to syntactic roles. If we observe individual differences for locatives, they may reflect either differences in the structural representation or differences in the thematic-syntactic role mapping.

Wh-island structures may be formulated with *gaps*, indicated with an underscore in examples, or with a special kind of pronoun called a *resumptive pronoun*. Both of these structures show relatively low acceptability in English, although they have both been shown to be produced (Ferreira and Swets, 2005). According to Morgan and Wagers (2018), English speakers have a grammatical representation of gaps, but not of resumptive pronouns, which they take to be the result of a breakdown in production processes. If this is true, then individual differences in syntactic representations should only be detectable in GAP structures, and so production should be predicted by the acceptability of the GAP alternate, not the difference in acceptability between gaps and resumptive pronouns, as shown by Morgan and Wagers. We therefore analyzed PRONOUN production as a function of GAP acceptability.

Finally, we included the agreement stimuli as proof of concept for our method of detecting individual differences. These *agreement attraction* errors have been elicited in experimental settings dating back to the 1990s (Bock and Miller, 1991), and were anecdotally reported throughout the literature prior to that. As with wh-islands, the two agreement alternates differ in their grammatical status. The *singular* alternate is grammatical because the verb agrees in number with the head of the subject NP (“gate” in Table 3.1). The *plural* alternate is ungrammatical because the verb agrees with a plural noun that intervenes between the head NP and the verb. Again, then, we should expect any individual differences in the representation to be reflected in a relationship between production and SINGULAR acceptability, not preference. However, whereas wh-island structures are very rare (Ferreira and Swets, 2005, estimate that English speakers hear roughly one per day), subject-verb agreement is very common. Any individual differences in the representations of the SINGULAR alternates should be very small, and likely undetectable. We therefore did not expect to see evidence for individual differences for these stimuli.

Experiment 1 tested all four of these structures. The results established that there are

in fact individual differences in syntactic representations. Experiment 2 aimed to replicate the finding of Experiment 1 with a slightly modified experimental design, and to test the hypothesis that individual differences reflect varied experience with syntactic structures. We conclude by discussing implications for syntax, as well as for the representation of symbolic systems in general.

3.2 Experiment 1

In Experiment 1 we asked whether there are individual differences in participants' representations of syntactic structures. Participants produced sentences that required them to choose one of two possible structures, and then provided acceptability ratings for sentences with each of these eight structures. We then compared individuals' preference for a given structure to the frequency with which they produced it. If the strength of individuals' representations of syntactic structures varies from person to person, then we should see a significant relationship between the production and acceptability data. We also included a verbal working memory task in order to test the hypothesis that individual differences for some structures might reflect differences in general cognitive abilities as opposed to in syntactic representation.

3.2.1 Method

Participants

350 UC San Diego undergraduates participated for course credit. This number was chosen in hopes that it would be large enough to detect small effects in a dataset where each participant contributes only one data point.

A total of 55 participants were excluded. Pre-screening instructions indicated that participants should be native monolingual speakers of English, having learned English and no other languages before the age of 7. Based on responses to a linguistic background questionnaire, we excluded 32 participants for not meeting this criterion. As an attention check, we included fillers that had unambiguously high acceptability (e.g., "Everyone takes advantage of the few



Figure 3.1. Screenshot of a production trial from Experiment 1. Participants were instructed to “rephrase” the top sentence by filling in the blank, completing the bottom sentence. The trial shown is a DATIVE item; target responses were either “the dog the bone,” a DOUBLE OBJECT response, or “the bone to the dog,” a PREPOSITIONAL DATIVE response.

sunny days we get in Portland”) and unambiguously low acceptability (e.g., “The house that the builder that the contractor fired constructed collapsed”). For each participant, we calculated the difference between the mean rating they gave for high-acceptability fillers and the mean rating they gave for low-acceptability fillers. We excluded 18 participants whose difference score was less than two standard deviations below the mean of all participants’ difference scores. Finally, we excluded 5 participants for having a mean accuracy below 60% on comprehension questions to unambiguous filler sentences. The remaining 295 participants were subject to further analysis-specific exclusions as discussed below.

Procedure

The task was run online using the Ixcel Farm platform (Drummond, 2013) and took most participants under 60 minutes. Participants read informed consent, completed a language background questionnaire, read instructions for the production task, and completed four practice trials with example good and bad responses provided.

Participants then completed the production phase of the experiment, consisting of 64 critical trials and 12 filler trials. Following Morgan and Wagers (2018), production trials were explained to participants as opportunities for them to “rephrase” a context sentence. Each trial consisted of a context sentence above the beginning of a new sentence, as in Figure 3.1.

Participants then read instructions for the acceptability judgment task, completed two practice items, and began the acceptability judgment phase, consisting of 64 critical trials and 12 filler trials. Sentences appeared over a 1 to 9 Likert-like scale, where 1 indicated low acceptability

and 9 high. On approximately 75% of trials, a multiple-choice comprehension question replaced the sentence after participants selected a rating. This was meant to motivate participants to fully read each sentence.

Finally, participants completed the digit span task. This consisted of three blocks of sixteen trials in which participants had to recall series of between two and nine digits. Trials increased in difficulty (number of digits) within each block. Digits were presented one at a time in the center of the browser for 600 ms, with 200 ms of whitespace appearing between the presentation of each digit. In the first block, digits had to be recalled in the order they were presented. In the second block they were recalled in reverse order. In the third block, they were recalled in sequential order (i.e., “3-1-2” would be recalled as “1-2-3”).

Materials

We tested four classes of structures: DATIVE, LOCATIVE, WH-ISLAND, and AGREEMENT. Each of these four classes consists of a pair of structural alternates. In production trials, participants had to (implicitly) choose between these alternates to successfully complete a sentence (Table 3.2). In acceptability trials, we collected ratings of each of the eight structural alternates independently (Table 3.1). Note that the target sentences from the production block were identical to the stimuli in the judgment block. Items were counterbalanced according to a Latin Square design so that the same item might have been seen by one participant in a production trial, another participant in an acceptability trial with one alternate form, and another participant in an acceptability trial but with the other alternate form. The order of stimuli within each block was randomized.

In order to ensure that the acceptability ratings were not confounded by ceiling or floor effects, we included 12 fillers that were meant to be more acceptable than any of the critical items and 12 fillers that were meant to be less acceptable. The “ceiling” fillers were syntactically simple, although varied in structure from item to item (e.g., “There was an enormous painting of a Teddy bear next to the fireplace” and “You lose your earrings every time you go to the beach”).

Table 3.2. Experiment 1 production stimuli and target responses. Participants used the information given by the context sentence to complete the sentence begun by the prompt, as in Figure 3.1.

Structure Class	Stimulus	
DATIVE		
<i>context:</i>	The players learned the plan from the coach.	
<i>prompt:</i>	We all believed that the coach taught...	
<i>options:</i>	...the plan to the players.	PD
	...the players the plan.	DO
LOCATIVE		
<i>context:</i>	The cake had chocolate on it when the chef was done.	
<i>prompt:</i>	Michael groaned watching the chef drizzle...	
<i>options:</i>	...the cake with chocolate.	WITH
	...chocolate on the cake.	ON
WH-ISLAND		
<i>context:</i>	The teacher understood why I scolded the little girl after school.	
<i>prompt:</i>	There goes the little girl that the teacher understood why...	
<i>options:</i>	...I scolded after school.	GAP
	...I scolded her after school.	PRONOUN
AGREEMENT		
<i>context:</i>	The gate to the pastures, gradually falling down due to disrepair, frustrated the farmer.	
<i>prompt:</i>	The farmer was frustrated; the gate to the pastures...	
<i>options:</i>	...was gradually falling down due to disrepair.	SINGULAR
	...were gradually falling down due to disrepair.	PLURAL

Floor fillers were all cases of doubly-center-embedded relative clauses (e.g., “That present that the guitarist that the felon robbed received melted”).

Prior to analysis we excluded two dative items using the verb “award,” for which some responses were ambiguous between dative and locative constructions (e.g., “...awarded an honorary diploma to the speaker” could be the PD form of the DO dative “...awarded the speaker with an honorary diploma” or the ON locative form of the WITH locative “awarded the speaker an honorary diploma”).

3.2.2 Analysis

We analyzed the four structure classes separately, effectively treating them as four sub-experiments. For DATIVE and LOCATIVE items, we used weighted linear regressions to model participants’ production rate of one structure as a function of their preference for that structure, where preference was defined as the difference between a participant’s mean rating of that structure and their mean rating of the alternate. For WH-ISLANDS and AGREEMENT items, we followed the logic of Morgan and Wagers (2018) and used mean GAP/SINGULAR acceptability (not preference) as the lone predictors.

Model weights were a function of the total number of trials that contributed to each participant’s mean production rates and acceptability ratings. Given the nine-point response scale for acceptability rating, each acceptability trial contributed more information than any given production trial, which, after excluding uncodable responses, provided binary data. Specifically, assuming even probabilities over possible responses, acceptability trials contributed $-\log_2(\frac{1}{9}) = 3.17$ bits of information to production trials’ $-\log_2(\frac{1}{2}) = 1$ bit (Shannon, 1948). To calculate weights, we first estimated the total amount of information that contributed to each production and acceptability mean by multiplying the number of bits of information associated with a given trial type (1 for production and 3.17 for acceptability) with the number of trials of that type for a given structure. Thus, for every structure in production and every structural alternate in acceptability, we obtained the mean value and the amount of information that contributed to

that mean. Then, each observation in the model – consisting of a mean production rate and a difference of mean acceptabilities – was weighted by the minimum amount of information associated with any of these three values. That is, a trial was only considered as informative as its least informative component.²

Data coding and trial exclusions

For both Experiments 1 and 2, the production data were manually coded. For each structure class, trials were labeled as one of the two alternates, or they were excluded.

Trials were excluded if the response did not correspond to the structural makeup of one of the two alternates for that particular item type. For datives, this meant either a goal NP followed by a theme NP or a theme NP followed by a goal PP. For locatives, we accepted responses with a theme NP followed by a goal PP or a goal NP followed by a theme PP. For wh-islands, if the response included a resumptive pronoun, the pronoun had to match the intended referent in gender and number. Resumptive pronouns are much more common in subject positions than in object positions (Morgan and Wagers, 2018), suggesting a potential difference in the underlying representation of wh-islands terminating in a subject position and an object position. For this reason, we excluded responses where the gap or resumptive pronoun did not appear in a direct or prepositional object position (e.g., passive constructions). For agreement stimuli, trials were excluded if the response did not include an opportunity for the verb to agree with the subject, as in “the gate to the pastures fell down” (*fall* in the past tense does not show overt agreement marking with the subject).

Because our aim here is to measure structural representations, we included trials where the structure was correct, even if the meaning did not correctly reflect that of the context sentence. Thus, if a participant had responded to the dative prompt in Table 3.2 with “himself something” (instead of “the players the plan”), this was coded as DO because the structure was right, even

²Note that by using surprisal values as weights, we are in keeping with the more standard approach of weighting observations by the variance associated with each observation because the information content as calculated here is directly proportional to variance.

though the meaning was not.

However we did not include trials where meaning could only be expressed using one structure. For instance, “threw a life preserver *at* the swimmer” was not coded as a PD because the same idea cannot be expressed with the DO structure. Production of this structure therefore does not require a choice between structures.

In the acceptability rating data, 3094 trials (13% of total) were excluded because the comprehension question was answered incorrectly. An additional 193 trials (1% of total) were excluded because participants responded in less time than it would have taken to read each word in 150 ms.

While acceptability judgments are often z-scored by participant to reduce the impact of different strategies for mapping from subjective acceptability to the Likert scale, we chose to analyze raw ratings. If individual differences do exist, then centering and scaling participants’ data based on their own means and standard deviations may remove the very same variance we aim to observe.

3.2.3 Results

Data appear in Figure 3.2 and the results of all a priori and exploratory analyses are given in Table 3.3. For dative and locative items, there was a significant positive relationship between preference for one alternate and the likelihood of producing that alternate. For wh-islands and agreement items, the acceptability of the grammatical structure alone did not predict pronoun/agreement attraction error production.

Wondering whether Morgan and Wagers (2018) may have been wrong about the irrelevance of PRONOUN acceptability for determining GAP production, we ran a post-hoc analysis in which we added PRONOUN acceptability. In this model, both GAP and PRONOUN acceptability significantly predicted PRONOUN production: speakers are more likely to produce pronouns the more they like pronouns and the less they like gaps. Model comparison confirmed that adding PRONOUN acceptability significantly improved the model ($F(1, 290) = 11.412, p < .001$), and a

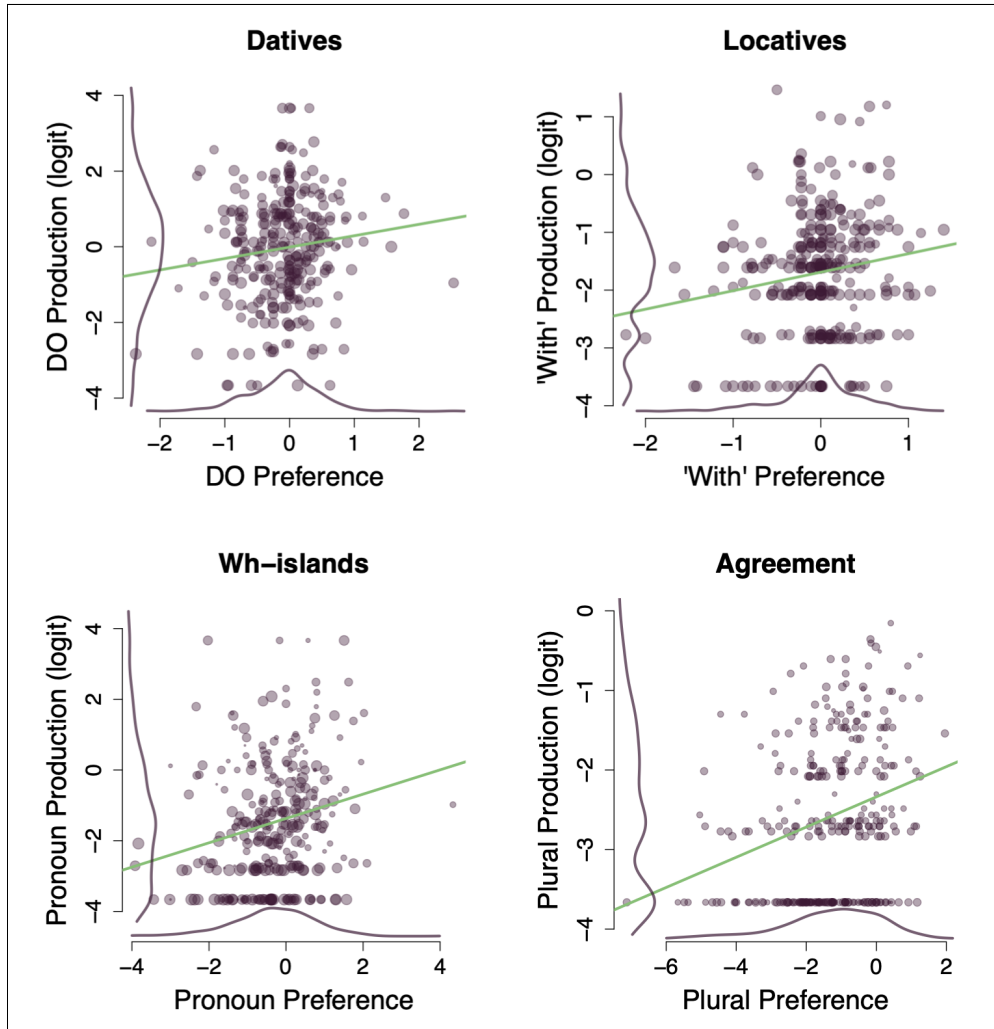


Figure 3.2. Experiment 1 results. Best-fit lines plotted in green; density plots in purple along axes. Point size is proportional to the observation's weight in the model.

Table 3.3. Experiment 1 results: all four structure types.

	β	t	p	
<i>Datives</i>				
Intercept	-0.043	-0.506	.613	n.s.
DO preference	0.297	2.054	.0409	*
<i>Locatives</i>				
Intercept	-1.704	-28.377	< .001	***
WITH preference	0.314	2.669	.008	**
<i>Wh-islands</i>				
Intercept	-1.242	-4.677	< .001	***
GAP acceptability	-0.087	-1.530	.127	n.s.
<i>Post-hoc Analysis</i>				
Intercept	-1.243	-4.582	< .001	***
GAP acceptability	-0.357	-3.843	< .001	***
PRONOUN acceptability	0.326	3.378	< .001	***
<i>Agreement</i>				
Intercept	-2.193	-4.863	< .001	***
SINGULAR acceptability	-0.053	-0.899	.369	n.s.
<i>Post-hoc Analysis</i>				
Intercept	-2.086	-3.817	< .001	***
SINGULAR acceptability	-0.155	-2.503	.013	*
PLURAL acceptability	0.190	4.765	< .001	***
Digit span	-0.083	-1.423	.156	n.s.

model with PRONOUN preference (the difference between PRONOUN and GAP acceptability) was also significant ($\beta_{preference} = 0.321, t = 3.663, p < .001$).

As we predicted, AGREEMENT production was not predicted by the acceptability of SINGULAR structures alone. Agreement attraction errors are attributed to problems with memory retrieval. We wondered if we might be able to account for some of the production data with individual differences in working memory capacity – as indexed by the digit span task or by the comprehension of agreement attraction errors. A post-hoc analysis, also in Table 3.3, revealed that the acceptability ratings of PLURAL stimuli predicted production, but digit-span scores did not; this model was a significant improvement on our a priori model ($F(2, 289) = 13.045, p < .001$). Interestingly, SINGULAR acceptability was also significant in this model.

3.2.4 Discussion

Experiment 1 revealed individual differences for all four structure classes. For dative and locative structures, our predictions were born out: the more a speaker prefers one structural alternate to the other, the more likely they are to produce it.

For wh-islands, however, GAP acceptability alone did not predict production. Instead, a post-hoc analysis revealed that it is the relative acceptability of gaps and pronouns that is predictive – that is, how much more acceptable a sentence like “There goes the little girl that the teacher understood why I scolded her after school” was than the corresponding sentence with a gap, “There goes the little girl that the teacher understood why I scolded __ after school.” In a similar analysis, Morgan and Wagers (2018) found that across different kinds of island structures (not participants), GAP acceptability was a better predictor of production than the relative acceptability of gaps and pronouns. The production system appeared not use representations of resumptive pronouns when choosing a structure. This, they argued, reflects the fact that that there *is* no representation of the resumptive pronoun – or in other words, it is ungrammatical. If it were not for the agreement data, discussed next, our finding that PRONOUN acceptability *does* predict production might contradict Morgan and Wagers’s argument.

Agreement items showed a similar pattern. When production was modeled as a function of the acceptability of the SINGULAR alternate (e.g., “The wooden bridge to the islands was about ten miles off the highway”), the result was not significant. But when PLURAL acceptability was added to the model (“...bridge to the islands *were* about...”), then both SINGULAR and PLURAL acceptability were significant. Unlike with resumptive pronouns, the ungrammaticality of the PLURAL alternate is not typically questioned. Therefore, the agreement data indicate that there is a way for the acceptability of an ungrammatical alternate to predict production. As such, our findings do not clearly bear on the question of whether resumptive pronouns are grammatical.

If PLURAL sentences are indeed ungrammatical, then individual differences cannot reflect differences in a syntactic representation. One possibility, though, is that there may be individual differences in the grammatical status of plural agreement specifically when the singular subject is *notionally* plural – so called *collective* nouns like “family” and “army.” Lending credence to this hypothesis is the fact that such differences exist across dialects of English. In British English, plural verb agreement with collective nouns, as in “The government have decided...,” is not only reportedly acceptable, but according to some proscriptive grammars is the only correct form of agreement (Bock et al., 2006).

Our stimuli contained one collective noun, “information,” and two mass nouns, “enforcement” and “footage,” in the critical subject position. To determine whether differences in the syntactic number feature for these nouns might drive individual differences across items, we excluded these items and re-ran the analysis. The results were the same: the more agreement attraction errors a participants produced, the higher they rated PLURAL items ($\beta = 0.182$, $t = 4.522$, $p < .001$) and the lower they rated SINGULAR items ($\beta = -0.153$, $t = -2.513$, $p = .012$).

Individual differences in AGREEMENT items therefore probably do not reflect differences in syntactic representations. But they may reflect differences in a general cognitive system involved in agreement dependency tracking. While working memory is an obvious candidate, scores on the digit span task did not significantly predict production. Instead, we suggest that the

significant effect of PLURAL acceptability may reflect attentional differences between participants. Participants who were more engaged in the task would be more likely to notice the errors in the acceptability items and assign them low ratings. Higher levels of attention may also have resulted in a better ability to keep track of plural features, resulting in fewer agreement attraction errors in production. Indeed, two post-hoc models showed that participants' mean accuracy on comprehension questions following unambiguous, grammatical stimuli (datives, locatives, and CEILING fillers) significantly predicted their production of agreement errors ($\beta = -0.230$, $t = -2.523$, $p = .012$) and marginally predicted their acceptability ratings of agreement errors ($\beta = -0.308$, $t = -2.205$, $p = .056$; p -values Holm-Bonferroni adjusted for a family of 2 estimates).

We started by asking whether syntactic representations may, like phonemes, have an item-based representation. If so, we argued, we should observe that representations vary from person to person. Experiment 1 showed that there are in fact individual differences in linguistic behaviors. Stronger evidence for item-based representations would be if we could demonstrate a causal relationship between exposure and individual differences. This was one goal of Experiment 2.

Exposure is known to lead to changes in both production and acceptability ratings. For instance, Bock (1986) demonstrated that following exposure to a DO dative, participants were more likely to use the DO structure than they were following exposure to a PD structure. This effect, known as *syntactic priming*, can be seen both within and across modalities (Potter and Lombardi, 1998; Bock et al., 2007; Pickering and Ferreira, 2008), can be seen in patients with anterograde amnesia (Ferreira et al., 2008; Yan et al., 2018), and persists over the course of weeks (Kaschak et al., 2011). These properties have led many researchers to believe that priming effects reflect implicit learning (see Pickering and Ferreira, 2008 for a review and Chang et al., 2006 for a computational model serving as proof of concept).

In acceptability, repeated exposure to structures with low acceptability has been shown to increase their acceptability (Snyder, 2000; Luka and Choi, 2012; Luka and Barsalou, 2005; although this may not be true for all types of unacceptable structures; see Goodall, 2011). This,

too, may reflect a type of implicit learning. The question we aim to address here is whether these two effects reflect the same type of learning, contributing to the strengthening of a syntactic representation.

If so, then the picture that emerges is of a system of item-based representations whose “strengths” are directly proportional the number of times they have been experienced. These strengths are reflected in comprehension as higher acceptability and in production as increased rates of choice relative to other candidate structures.

One perhaps surprising prediction of this view is that common grammatical structures should also show satiation effects. To our knowledge, this has not been shown. It is at odds with a common tacit assumption that patently grammatical structures should have ceiling-level acceptability, modulo extra-grammatical factors such as pragmatic content and processing difficulty.

The hypothesis that structural strength is a direct reflection of the number of exposures an individual has to that structure makes other empirical predictions, too. First, we should expect to see that the more items a person is exposed to, the more their productions and acceptability judgments change. Furthermore, the same amount of exposure should affect different people’s behaviors differently. Being exposed to a structure 18 times should result in a big change for a person who has seen that structure relatively less before, but a much smaller change for someone who has seen that structure relatively more. This predicts that the amount of change in individuals’ production and acceptability ratings should be correlated.

Experiment 2 was designed to test all of these hypotheses by measuring participants’ production and acceptability ratings before and after an exposure regimen. It was also designed to replicate the findings of Experiment 1, while removing one potential confound. Counterbalancing items across conditions and randomization are standard experimental tools used to isolate the effect of a condition, independent of item- and order-specific effects. However, our approach in Experiment 1 was intended to be a psychometric one, not a standard experimental one. That is, our goal was to estimate differences in individuals, not differences between conditions. But as

long as items are counterbalanced and randomized, as they were in Experiment 1, any observed individual differences may reflect artifactual individual differences due to the fact that different participants saw different sentences, and orderings thereof. Experiment 2 rectifies this by fixing the stimulus order and not counterbalancing items across conditions. Any individual differences we might observe are therefore not artifacts of the stimulus design.

3.3 Experiment 2

Experiment two aimed to address two broad questions. The first is whether the finding of individual differences persists when we remove counterbalancing and randomization from the design, as these could potentially drive individual differences. The second is whether exposure leads to changes in production and acceptability that are consistent with what we would expect if syntactic representations were item-based.

The experiment had three phases: pre-test, exposure, and post-test. The pre-test and post-test phases each consisted of a production block followed by an acceptability judgment block, as in Experiment 1. The exposure phase consisted of a single acceptability judgment block in which participants were exposed to (by being asked to rate) either 6 or 18 DOs or PDs, and either 18 or 6 sentences with gaps or resumptive pronouns (counterbalanced such that all participants saw the same total number of stimuli). To ensure that any change we observed in acceptability from pre- to post-test did not reflect differences in the specific items used, we ran a norming task so as to match items across phases for baseline acceptability and production rates.

3.3.1 Norming task

A total of 75 participants participated in the norming task for course credit; 25 were excluded. Production responses were coded as in Experiment 1. The procedure was exactly the same as in Experiment 1.

To reduce the length of Experiment 2, we limited critical items to dative and wh-island structures. As discussed in the introduction, individual differences in datives would represent

the clearest evidence for individual differences in syntactic representations. Wh-islands were chosen for two reasons: the fact that our findings for these items differed from those of Morgan and Wagers (2018) indicate a need for high-powered replication, and given the relatively low frequency of these structures, the effect of exposure may be larger and therefore easier to detect.

Materials from Experiment 1 were updated in various ways. Items with high or low acceptability or production rates relative to other items in their structure class were modified or replaced. Prompts for wh-island production stimuli were also systematically adjusted. Whereas in Experiment 1 the prompt ended at the beginning of the low clause, requiring the participant to minimally write a subject and a verb, now we included the subject in the prompt. This was meant to reduce the number of trial exclusions by preventing participants from producing passive responses.

Stimuli included 84 dative and 84 wh-island items (for each: 28 production, 28 PD/GAP judgment, 28 DO/PRONOUN judgment) and 118 fillers. Fillers consisted of 19 ceiling and 19 floor items (all judgment) and 80 agreement items (26 production, 27 SINGULAR judgment, 27 PLURAL judgment). Each item was arbitrarily assigned to appear in only one of three conditions across participants: production, or acceptability judgment in one of the two alternate forms.

Prior to assigning items to pre- or post-test, we excluded 94 acceptability trials that were answered too quickly. To more accurately estimate items' baseline acceptability, we residualized judgments using mixed effects models (one for each type of structure: GAP, PRONOUN, DO, PD, CEILING, FLOOR, SINGULAR, and PLURAL) with single fixed effects for trial order, random intercepts for participant, and random slopes for trial order. Production data were residualized using models with fixed effects for trial order and the participants' response on the previous trial of the same structure class (meant to account for self-priming effects), random intercepts for participant, and random slopes for trial order and previous responses (except for agreement items, where we removed random slopes in order for the model to converge).

The resulting residual scores were highly reliable. In a series of split-half analyses, we randomly assigned trials to one of two data subsets, calculated item means for each of these

subsets, and then calculated the correlation coefficient of item means. We iterated this process 1000 times and averaged the correlation coefficients for each structure type. In production, these mean correlations were .805 for dative items, .480 for wh-island items, and .871 for agreement items. In acceptability, means were .634 for PD items, .700 for DO, .798 for GAP, .559 for PRONOUN, .771 for SINGULAR, .772 for PLURAL, .663 for CEILING, and .616 for FLOOR.

In assigning items to phases for Experiment 2, we aimed to minimize item-based differences between pre- and post-test. Our goal was not only to match overall acceptability and production rates across phases, but also the order with which items of particular baseline acceptabilities/production rates were presented. Each item was therefore paired with another item of the same structure and with a similar mean. Items which could not be paired were either set aside for use in the exposure phase or excluded if they appeared to behave differently from other items of the same structure. This resulted in the use of 54 dative items (for both pre- and post-test: 11 production, 8 PD, and 8 DO); 56 wh-island items (12 production, 8 GAP, and 8 PRONOUN); and 84 fillers (8 CEILING, 8 FLOOR, 10 agreement production, 8 SINGULAR, 8 PLURAL)).

The result was that pre-test stimuli were very closely matched with post-test stimuli: mean residual acceptability across items in pre-test items was 0.017 (*s.d.* = 0.386) and in post-test items was 0.016 (*s.d.* = 0.380) and mean residual production rate for pre-test was -0.123 (*s.d.* = 0.289) and for post-test -0.122 (*s.d.* = 0.291). The order of item differences was also very closely matched. For the acceptability blocks, when sorted by trial number, the correlation of pre-test item means and post-test item means was $r = .993$; for production blocks it was $r = .992$.

3.3.2 Method

The method and analysis for Experiment 2 were pre-registered on AsPredicted.org after the first six weeks' worth of data (107 participants) had been collected and used to draft the analysis. The pre-registration document can be found at <<<http://aspredicted.org/blind.php?x=sr6xn4>>>.

Participants

A power analysis indicated that 900 participants would be needed to achieve 80% power. We therefore began running and decided on a stopping point of either 900 participants or as many as could be run by the end of October 2019. A total of 843 participants completed the study; 794 participated for course credit and 49 were workers on Amazon's Mechanical Turk who were paid between \$12 and \$15.

Due to the high target number of participants, we opened the study to multilinguals so long as they learned English before the age of 7, self-identified as a native speaker of English, and had no self-reported trace of a non-native accent. We reasoned that if individual differences and the size of priming and satiation effects are in fact dependent on experience, then including multilinguals might make these effects easier to detect given that multilinguals necessarily have less English experience than their monolingual peers.

We excluded 161 participants: 79 did not meet our criteria for being considered a native speaker of English; 46 responded to fewer than 60% of comprehension questions for unambiguous sentences correctly, and 36 were excluded because the difference between their mean ratings of ceiling fillers and floor fillers were lower than two standard deviations below the mean. A total of 682 participants' data were included in the analysis.

Procedure

The task was run online using Ixweb Farm and took most participants under 120 minutes. Participants read informed consent, completed a language background questionnaire, read instructions for production blocks, completed four practice production trials with example good and bad responses, completed a training block of digit span items, and then began the pre-test phase. The experiment was organized into three phases: pre-test, exposure, and post-test. Pre-test and post-test consisted of a production block with 33 trials followed by an acceptability judgment block with 64 trials. Between phases, participants performed a block of 10 digit span task trials to reduce the possibility that any explicit memory for items carried over from phase to phase.

They then completed the exposure block, which consisted of 54 acceptability judgment trials (6 or 18 datives, 18 or 6 wh-islands, 18 plural agreement items, 6 ceiling fillers, and 6 floor fillers). After another digit-span block, they completed the post-test phase, which consisted of a block of 33 production trials followed by a block of 64 trials. Finally, they completed a debriefing questionnaire.

Materials

Materials were designed and normed as discussed for the norming task. All participants saw the same exact sentences in the same exact order, modulo group differences in the exposure phase. The between-subjects manipulations (amount of exposure and which structural alternates were exposed) were implemented such that all participants saw the same number of exposure items, regardless of group: participants who saw 18 wh-island sentences during exposure saw only 6 datives, and participants who saw 6 wh-islands saw 18 datives.

Factors

We manipulated two factors between subjects. The first, STRUCTURE, was a manipulation of which structural alternate participants were exposed to. Roughly half of participants were exposed to GAP wh-islands and half to PRONOUN wh-islands, and of each of those groups roughly half were exposed to DO datives and the other half to PDs. The other manipulation, AMOUNT of exposure, was either 6 items or 18 items. We kept the total number of items in the exposure phase the same for all participants by counterbalancing such that participants who were exposed to 6 wh-island items were exposed to 18 datives and vice versa. This resulted in 8 distinct treatment groups: one where participants were exposed to 6 GAP structures and 18 DO structures; one where participants were exposed to 18 GAP structures and 6 DO structures; one where participants were exposed to 6 PRONOUN structures and 18 DO structures; and so on.

Data coding and trial exclusions

Production data were coded according to the same criteria as in Experiment 1. A total of 14,546 critical trials were excluded from acceptability rating data; 5,242 because the participant responded in less time than it would have taken to read each word in 150 ms, and 9,304 because comprehension question was answered incorrectly. This left 29,050 dative trials and 20,381 wh-island trials to be analyzed.

3.3.3 Pre-registered analyses

Of the six pre-registered analyses, four are detailed below. In the interest of space, we leave out a split-half reliability analysis and a comparison of differences across individuals to changes within individuals (Analyses 4 and 6 in the pre-registration), as the results did not meaningfully contribute to the discussion. These (and all) analyses can be found on OSF under <<<https://osf.io/g3txq>>>.

Individual Differences

The first analysis was meant to determine whether the finding of Individual Differences in Experiment 1 replicates when holding the stimulus design constant across participants. The analysis approach was the same as that of Experiment 1: a weighted linear regression on the data from the pre-test phase. However, rather than only using GAP acceptability to predict production in WH-ISLAND trials (which post-hoc analyses determined was not sufficient for detecting individual differences), we used both GAP and PRONOUN acceptability. For datives, we similarly used the acceptability of both structural alternates.

Priming

The second analysis looked at syntactic priming in production: whether participants produced more of a given structure after having been exposed to that structure, and whether more exposure led to more of an increase in production. This was largely meant as a validity check:

priming is well-attested in experiments with sample sizes under 50.

For each of the two structure classes, a mixed-effects logistic regression (R Core Team, 2018; Bates et al., 2014) was fit to productions from the pre-test and post-test phases. Fixed effects were included for PHASE, which had levels PRE-test and POST-test; for AMOUNT of exposure, with levels 6× or 18×; and for exposed STRUCTURE, with levels GAP and PRONOUN for the wh-island model and DO and PD for the dative model. The models had random intercepts for participants and items; fixed effects that varied within levels of random effects were nested within those random effects such that a random slope was fit for PHASE within participants and for AMOUNT and STRUCTURE within items. When models did not converge, we first removed the random effects correlation structure and then random slopes one by one, in order of least variance accounted for to most (following Barr et al., 2013). All mixed-effects models reported have the maximal random effects structure for which the model converged. For all significant effects of theoretical interest, we also report the results of model comparison to confirm that the effect significantly contributes to model fit.

Satiation

The third analysis looked at syntactic satiation. Again, two linear mixed-effects were fit on acceptability judgments from the pre- and post-test phases, one model for each of the structure classes. The dependent variable was acceptability ratings of the exposed structure – for example, GAP ratings for participants exposed to sentences with gaps and PRONOUN ratings for participants exposed to pronouns. The fixed- and random-effects structures and process of removing random slopes (when necessary) was identical to those in Analysis 1. For wh-islands, this analysis was intended as a validity check, as satiation is attested in such structures (Snyder, 2000). Showing satiation for straightforwardly grammatical structures like datives, however, would be a novel result.

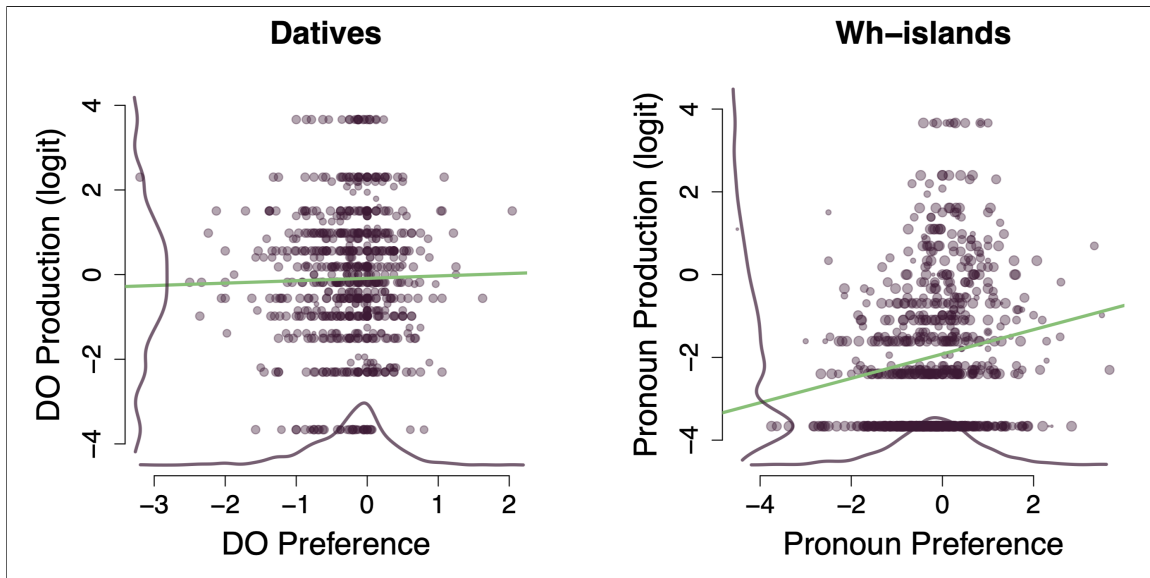


Figure 3.3. Experiment 2: Production as a function of preference in data from the pre-test phase. Point size is proportional to the observation’s weight in the model. Best-fit lines plotted in green; density plots in purple along axes.

Correlation between size of priming effect and satiation effect

The fourth analysis aimed to determine whether participants with relatively bigger changes in the production rate of a given structure from pre-test to post-test in acceptability were also participants with relatively bigger changes in preference for that structure. A linear model was run for each of the two structure classes predicting the amount of change in production as a function of the amount of change in preference.

3.3.4 Results

Results of each analysis are discussed below in turn.

Individual Differences

Plots of participants’ productions as a function of preference for one alternate vs. the other are given in Figure 3.3. The finding of individual differences in Experiment 1 replicated for wh-islands, but not for datives. Again, both GAP and PRONOUN acceptability significantly predicted the likelihood of producing a pronoun.

Table 3.4. Experiment 1 results: Individual differences in the pre-test data.

	β	t	p	
<i>Datives</i>				
Intercept	0.723	1.359	.174	n.s.
DO acceptability	0.023	0.217	.828	n.s.
PD acceptability	-0.124	-1.113	.266	n.s.
<i>Wh-islands</i>				
Intercept	-1.922	-9.746	< .001	***
PRONOUN acceptability	0.341	4.982	< .001	***
GAP acceptability	-0.346	-4.945	< .001	***

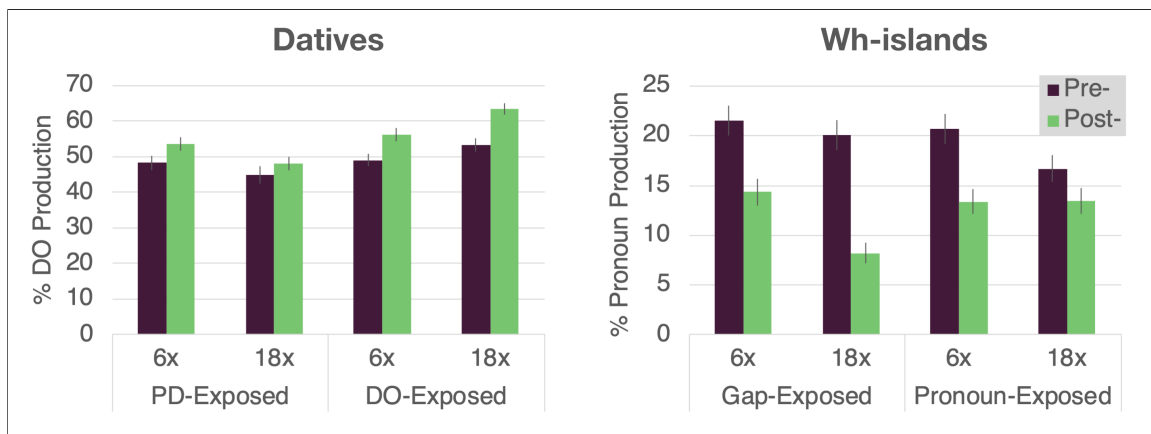


Figure 3.4. Experiment 2: Production data for dative (left) and wh-island (right) stimuli.

Priming

Production data are shown in Figure 3.4 and the results of the priming models appear in Table 3.5. For datives, the exposure regimen seems not to have worked. The model revealed no significant effects. (A marginal effect of the interaction of STRUCTURE and AMOUNT reflects a small between-group difference *prior* to exposure, and therefore does not bear on priming.)

For wh-islands, there was an across-the-board decrease in pronoun production in the post-test phase ($\chi^2(1) = 39.601, p < .001$). Among gap-exposed participants, this decrease was bigger (the significant interaction between PHASE and AMOUNT; $\chi^2(1) = 7.931, p = .004$). There was also a significant three-way interaction reflecting the fact that this augmented decrease was not present in the pronoun-exposed group ($\chi^2(1) = 10.406, p = .001$).

Table 3.5. Experiment 2 results: Priming.

	β	t	p	
<i>Datives</i>				
Intercept	-0.134	-0.438	.661	n.s.
PHASE: POST	0.331	0.803	.422	n.s.
AMOUNT: 18×	-0.232	-1.316	.188	n.s.
STRUCTURE: DO	0.037	0.205	.838	n.s.
Interaction: PHASE × AMOUNT	-0.123	-0.764	.445	n.s.
Interaction: PHASE × STRUCTURE	0.155	0.937	.349	n.s.
Interaction: × AMOUNT × STRUCTURE	0.497	1.885	.059	.
3-way Interaction	0.349	1.442	.149	n.s.
<i>Wh-islands</i>				
Intercept	-2.342	-10.616	< .001	***
PHASE: POST	-1.876	-6.337	< .001	***
AMOUNT: 18×	-0.019	-0.068	.946	n.s.
STRUCTURE: PRONOUN	0.117	0.403	.687	n.s.
Interaction: PHASE × AMOUNT	-0.979	-2.830	.005	**
Interaction: PHASE × STRUCTURE	-0.081	-0.232	.816	n.s.
Interaction: AMOUNT × STRUCTURE	-0.608	-1.541	.123	n.s.
3-way Interaction	1.566	3.249	.001	**

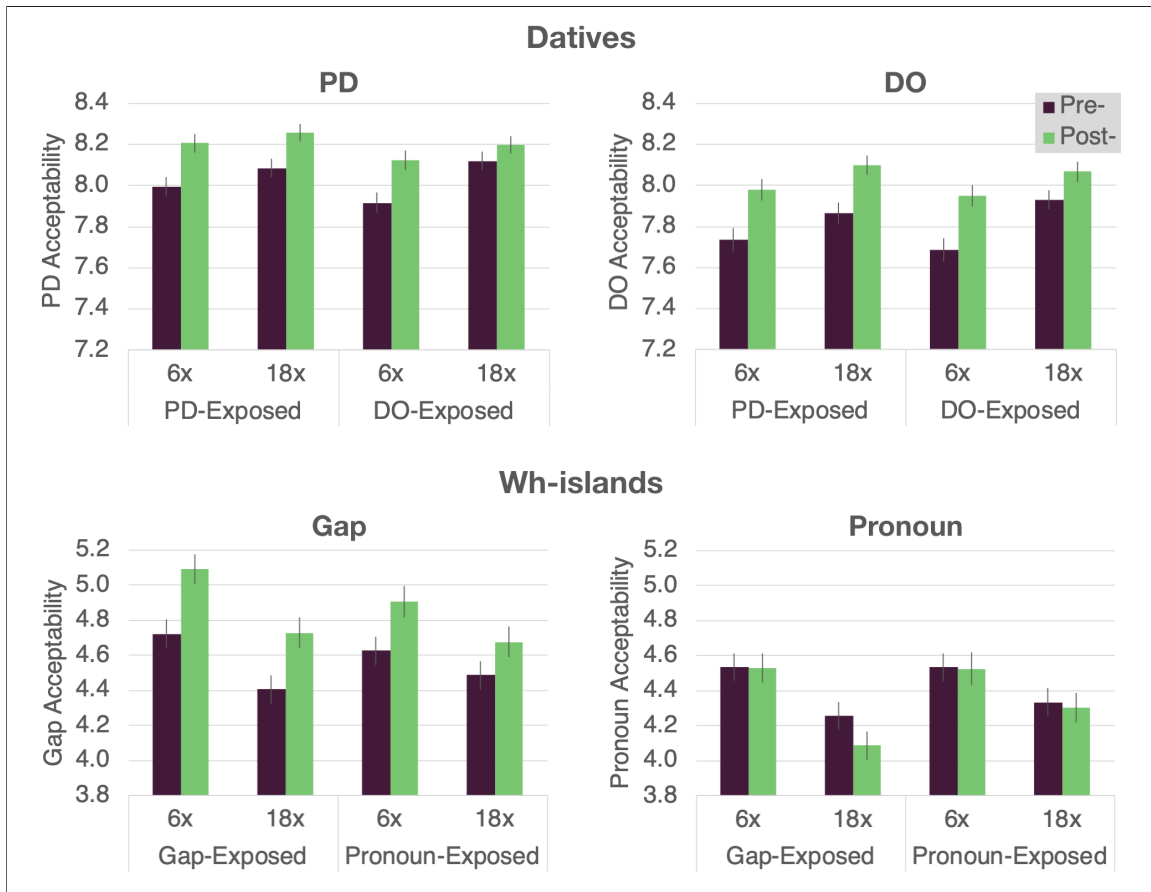


Figure 3.5. Experiment 2: Acceptability ratings for all four structure types by condition.

Satiation

Acceptability data appear in Figure 3.5 and the results of the satiation models are given in Table 3.6. For datives, a significant main effect of AMOUNT and a marginal main effect of STRUCTURE reflect group differences *prior* to exposure. Contrary to what we predicted, a significant interaction of PHASE and AMOUNT reflects the fact that the increase in ratings from pre- to post-test was reduced for participants exposed to 18 sentences rather than 6 ($\chi^2(1) = 4.198$, $p = .040$).

For wh-islands, the exposure regimen again seems not to have worked, as the previously attested satiation effect was not present in the data. There was a significant effect of AMOUNT, reflecting a baseline group difference, but no other effects were detected.

Table 3.6. Experiment 2 results: Satiation.

	β	t	p	
<i>Datives</i>				
Intercept	7.662	58.370	< .001	***
PHASE: POST	0.283	1.691	.101	n.s.
AMOUNT: 18×	0.263	2.543	.011	*
STRUCTURE: DO	0.319	1.718	.092	.
Interaction: PHASE × AMOUNT	−0.156	−2.037	.045	*
Interaction: PHASE × STRUCTURE	−0.083	−0.350	.729	n.s.
Interaction: × AMOUNT × STRUCTURE	−0.174	−1.244	.214	n.s.
3-way Interaction	0.127	1.212	.229	n.s.
<i>Wh-islands</i>				
Intercept	4.686	22.684	< .001	***
PHASE: POST	0.395	1.529	.135	n.s.
AMOUNT: 18×	−0.326	−1.968	.049	*
STRUCTURE: PRONOUN	−0.130	−0.429	.669	n.s.
Interaction: PHASE × AMOUNT	−0.058	−0.442	.659	n.s.
Interaction: PHASE × STRUCTURE	−0.368	−0.996	.326	n.s.
Interaction: AMOUNT × STRUCTURE	0.116	0.470	.639	n.s.
3-way Interaction	< 0.001	0.001	1.000	n.s.

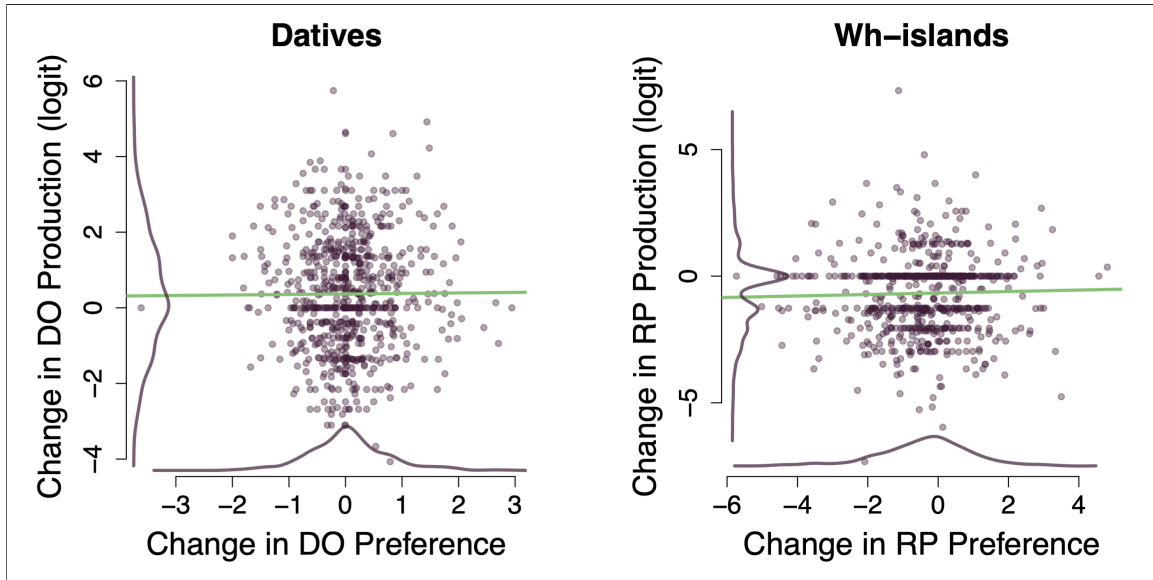


Figure 3.6. Experiment 2: The amount of change in production rate of DOs (left) or gaps (right) from pre-test to post-test as a function of the corresponding amount of change in preference for that structure from pre-test to post-test, where preference for a given structure is the difference between a participant’s mean rating of that structure and their mean rating of the alternate. Model fit plotted in green; density plots in purple along axes.

Correlation between size of priming effect and satiation effect

Figure 3.6 shows the size of the change in production rate of gaps or DOs from pre-test to post-test as a function of the size of their change in preference for that structure. The results of the corresponding analyses are given in Table 3.7. Perhaps not surprisingly given the results of the priming and satiation analyses, no significant relationship was detected for either wh-islands or datives.

Table 3.7. Experiment 2 results: Relationship between the size of the priming effect and the size of the satiation effect.

	β	t	p	
<i>Datives</i>				
Intercept	0.366	6.496	< .001	***
$\Delta_{post-pre}$ DO preference	0.014	0.172	.863	n.s.
<i>Wh-islands</i>				
Intercept	-0.664	-11.570	< .001	***
$\Delta_{post-pre}$ PRONOUN preference	0.029	0.725	.468	n.s.

3.3.5 Discussion

Experiment 2 had two goals. First, we aimed to replicate the finding of individual differences from Experiment 1 while controlling for the possible confound introduced by counterbalancing. The effect was successfully replicated in the wh-island stimuli, but not for datives.

It is possible that the dative effect in Experiment 1 was a Type I error. However, in light of the significant effects in other structures, this seems unlikely. More likely, we think, is that even with 682 participants, Experiment 2 may have been underpowered. The individual differences effect for datives in Experiment 1 was very small, accounting for only 1.4% of the variance (adjusted $R^2 = .011$). And while 682 participants is unusually high for experimental psychology, it is worth noting that each participant only contributed one data point to the model.

The second goal of Experiment 2 was to determine whether individual differences in the strength of syntactic representations might be attributed to individual differences in experience with those representations. To investigate this, we exposed participants to either DOs or PDs, and to either gaps or resumptive pronouns, and aimed to measure the resulting changes in production rates and acceptability ratings of these structures.

For datives, the only significant change from pre-test to post-test is an interaction reflecting the fact that increased exposure to PDs was associated with a reduction in the amount by which PD acceptability increased. There is no immediately obvious reason why this would be the case, and we think it likely that this effect is spurious. In the satiation data, effects were also not consistently present.

The wh-islands data were similarly unexpected. Participants produced significantly fewer pronouns in post-test than pre-test, regardless of which structure they were exposed to. This may reflect self-priming, as gaps are more common in these structures than pronouns. The fact that this effect was bigger for gap-exposed participants than for pronoun-exposed participants may be a sign that exposure had some effect, although this was not enough to counteract the across-the-board increase in gap production. Again, there was no evidence for satiation in the

wh-islands.

A number of post-hoc tests were performed in an attempt to identify possible explanations for the lack of priming and satiation, but no clear answer emerged. The inevitable conclusion seems to be that the exposure regimen did not work. There could be a number of reasons for this. One possibility is that the fact that the study was run online means that participants may have been distracted. However, Experiment 1 was run online and found a number of predicted effects, one of which was replicated in Experiment 2. Furthermore, we took reasonable measures to minimize potential artifacts of attentional differences by including comprehension questions after exposure trials and removing critical trials where comprehension questions were answered incorrectly.

A more likely explanation may have to do with the priming mechanism (and presumably the satiation mechanism as well, although less is known about this). In a computational model of production, Chang et al. (2006) model syntactic priming effects as the result of learning from prediction error. If a similar mechanism is at play in human cognition, then when the system expects a PD but gets a DO, it increases the “strength” of the DO representation to reduce the risk of future errors. Critically, however, for learning to result from prediction errors, there must be prediction.

However, prediction may not always be automatic, and there are many reasonable possibilities for why participants in Experiment 2 may not have been engaged in predictive processing. The task was long. Compared to priming studies that take place in the lab, where social pressures from experimenters may encourage engagement, there was little to encourage participants to stay engaged, and the many pragmatically odd sentences may have been exceedingly difficult to predict. If priming and satiation rely on prediction error, then this could explain why we do not see these commonly-observed effects in our data.

3.4 General Discussion

Syntax is traditionally thought of as a purely abstract set of representations. However, in the recent history of cognitive psychology, many other representations which were once thought to be purely abstract have been re-conceived in light of evidence that they are tightly linked to experience. Lexical representations, for instance, exhibit some categorical properties, as one would expect of an abstract representation. But they also show gradient sensitivity to features of their usage, like frequency and neighborhood density (Vitevitch et al., 1997; Shulman et al., 1978; Slowiaczek and Pisoni, 1986).

To determine whether syntax similarly shows gradient properties, we looked for individual differences in syntactic representations. Such differences, we hypothesized, should be apparent in behavioral measures that rely on syntactic representations, including sentence production and comprehension. In production, the “stronger” the representation, the more it should be chosen relative to other candidate structures. In comprehension, the “stronger” the representation, the higher its perceived acceptability should be.

Experiment 1 detected individual differences in four pairs of structural alternates: datives, locatives, wh-islands, and agreement attraction stimuli. A post-hoc analysis showed that participants who answered comprehension questions less accurately produced more agreement errors and rated such errors higher (although this latter finding was only marginally significant after multiple comparisons correction). This suggests that individual differences for this structure class may have been driven by differences in attention. For the rest, individual differences seem to reflect what we have called the “strength” of syntactic representations.

Experiment 2 aimed to determine whether these differences could be attributed to differences in experience by directly manipulating experience with an exposure regimen. In particular, if a structure’s strength corresponds to the number of times it has been experienced, then more exposure should lead to bigger changes in behavior. We tested this prediction by manipulating the *amount* of exposure participants received to one dative alternate and one wh-island alternate.

The experience-based account of individual differences also predicts that participants whose representational strength changes relatively more should show relatively bigger changes in both production rates and acceptability ratings. We therefore also looked for evidence of a correlation between the priming and satiation effect sizes within participants.

Experiment 2 replicated the finding of individual differences for wh-islands, but not for datives. Given that individual differences were detected for all four structures in Experiment 1, we believe the lack of evidence for individual differences in datives in Experiment 2 likely reflects a Type II error, although there is clearly a need for further high-powered replication.

Experiment 2 also failed to find evidence of priming and satiation. This was particularly surprising with 682 participants given that these effects are frequently reported with sample sizes under 50. It appears that the exposure regimen did not work. Probably as a result, we found no evidence to support the two predictions about properties of the priming and satiation effects: that more exposure would lead to bigger effects, and that the size of effects would correlate within individuals.

These questions stood not only to link individual differences in syntactic representations to experience, but also to address open questions about the nature of learning. Specifically, we have proposed that syntactic representations have varying “strengths,” and that these may be direct reflections of the number of exposures a speaker has had to a structure over the course of their lifetime. But another possibility is that strength reflects learning not from mere exposure, but from how unexpected a particular exposure is – that is, error-based learning (as in Chang et al., 2006).

One way to tease these two possibilities apart would be to determine whether 6 vs. 18 exposures to a particular structure makes a difference in its strength. If mere exposure leads to learning, then 18 exposures should lead to bigger changes than 6. However, when considering prediction error, 6 consecutive exposures to the same structure may lead participants to expect that structure even more, and therefore to be less likely to predict the alternate structure. With continued exposure to the same structure, then, there should be diminishing impact on the

strength of the structure. The two hypotheses thus make different predictions about the impact of different amounts of priming. Unfortunately, given the failure of the exposure regimen in Experiment 2, we are forced to leave these questions for future research.

3.4.1 What is syntactic strength?

We have referred to whatever property of syntactic representations that varies from speaker to speaker as its “strength.” There are a few possibilities for what strength might map onto at a cognitive level. One possibility is that it reflects something about the representation itself: the amount of representational architecture that is dedicated to it.

In lexical access, there are many theories about the underlying cause of frequency effects. For example, early work posited a serial search mechanism for lexical access, where words were searched for in order of their frequency (Forster et al., 1976). With the advent of connectionist models, frequency effects were recast in terms of spreading activation, for example, as the result of differences in words’ baseline activations (McClelland and Rumelhart, 1981). Something similar might also explain differences among syntactic representations.

Another possibility is implemented in Chang et al.’s (2006) computational model of sentence production. Here, priming is achieved by updating the weights to hidden units in a connectionist model following prediction errors. Thus, it could also be that strength reflects not something about the representation itself, such as its baseline activation, but in its connections to other representations.

3.5 Conclusion

Here we have argued that, similar to phonetic and lexical representations, syntactic representations should be thought of as existing at least in part in a continuous representation space. We argue this on the basis of the finding of individual differences in syntactic representations for multiple structures, although a failure to replicate this finding for dative constructions indicates the need for further replication.

These differences, if real, reflect relatively minor variations in speakers' likelihood of producing a structure and how acceptable they find that structure. We attribute them to an underlying property of the representation, which we refer to as its "strength," and suggest that strength may be a function of experience, consistent with current thinking about gradient differences in other types of representations.

A high-powered experiment that aimed to link these differences to differences in experience failed to produce consistent changes in production and comprehension behavior after an exposure regimen. This is not taken as evidence for or against any particular theory, as the expected changes are well-attested in the literature. Instead, it is assumed to reflect a problem in the experimental design. It is therefore left for future work to determine what factors contribute to a structure's strength, and in particular if this property is simply a reflection of a lifetime of experience.

Chapter 4

Conclusion

This dissertation reports three projects investigating the mental representation of syntax, or sentence-level linguistic structure. Taken together, these projects addressed extant problems in the field and demonstrate novel evidence suggesting a need to modify the way that syntactic representations are generally characterized.

Chapter 1 addressed a prominent hypothesis regarding the nature of resumptive pronouns in English. These pronouns pose a problem for traditionally held views about the relationship between grammaticality, acceptability, and production in that they are reliably produced, but have low acceptability. In order to account for this state of affairs, previous researchers have argued that they are ungrammatical, but that speakers nonetheless use them to help the comprehension on the part of their interlocutor. A series of high-powered experiments using different paradigms showed that, rather than helping the listener, resumptive pronouns reduce the accuracy of comprehension. This is interpreted as support for competing theories, according to which resumption is the result of production processes gone awry.

Having verified that production and comprehension are reliable metrics for probing syntactic representation in Chapter 1, the following two chapters used these to address questions about the nature of particular representations. For example, Chapter 2 asked a question about whether a class of representations known as *long-distance dependencies* are represented as several individual structures, or with one general representation. While a general representation

has long been modeled in various theories of syntax, the particular acquisition mechanism proposed by some theories of language acquisition make it difficult to imagine how such a representation could ever be derived.

It was predicted that if these structures had a general representation, then knowledge of one type would amount to knowledge of all types. In four artificial language learning tasks, participants were taught some types of relative clauses, and tested on their knowledge of other types. In each experiment, participants who learned that types of relative clauses they were exposed to also learned the types they had not been exposed to, indicating that there must be a representation of relative clauses that is general enough to account for both the trained and the untrained structures.

Chapter 3, too, asked a question about the nature of syntactic representations: whether they are purely abstract, or if they show gradient properties, consistent with what is known about other linguistic representations. In two experiments, evidence was found that native English speakers' syntactic representations vary, with some participants both finding the same structure more acceptable than others and producing that structure more. These individual differences were interpreted as reflecting variation in what was referred to as the "strength" of the representations. It is suggested that this property may simply reflect frequency: how much experience a person has with a structure. This would be parallel to item-based knowledge that is evident in lexical and phonetic representations.

Chapters 2 and 3 provide evidence that bear on a higher-level issue, too. Specifically, both of these projects portray syntactic representations as objects, just like words or phonemes. In Chapter 2, we demonstrate that these objects are abstracted over to form an even more general representation, and in Chapter 3 we demonstrate that they show gradient properties, just like other symbolic knowledge. But, while the literature does sometimes portray syntactic representations as symbols, they are in fact more commonly portrayed as an architecture: a system of abstract frames or procedures that coordinate morphemes to generate or decode strings. This characterization is necessary to account for the fact that syntactic representations are at

some level responsible for controlling how the system manipulates lexical representations.

How these two characterizations may be unified is not clear. It is possible that syntactic representations do exist as distinct objects, and that the architectural component of their behavior is derived from some intermediate sequencing system that builds structures out of words (as in Chang et al., 2006). Indeed, some generative theories of syntax involve a process like this known as *spellout*, which linearizes morphemes according to the specifications of nonlinear syntactic representations. At a very high level, this is consistent with the picture we have painted here: syntactic representations being represented as objects, and used by some other system to sequence morphemes. While these models are not explicitly intended to map onto cognition, their utility in capturing particular syntactic phenomena may be a sign that there are in fact cognitive processes like these theoretical ones.

But it is also possible that syntax is ultimately more like procedural knowledge (as in Ullman, 2001). On this view, if the finding of individual differences in Chapter 3 is on the right track, then what differs across individuals would not be a representation, but instead something about the sequencing mechanisms. The data presented in this dissertation are consistent with both of these possibilities. We suspect that a better understanding of the neural circuitry underlying sentence-level linguistic processes stands to shed light on these questions.

As a final note, this dissertation may also provide a new way to think about the problem of learnability first posed by Chomsky (1959). One of the core problems pointed out in that work is that human language is of a type that is not learnable solely on the basis of positive evidence. One type of structure which poses such a problem is the long-distance syntactic dependency. However, if syntactic representations are represented as objects which can be abstracted over, as we suggest in the general discussion of Chapter 2, then it may be possible for long-distance dependencies to be represented as abstractions over syntactic representations, just as syntactic representations are abstractions over words. If this is the case, then it reframes the problem in such a way that may not require innate knowledge of the non-context-free components of syntax, as Chomsky argues. Instead, all that is needed is the ability to form abstract representations of

extant representations. Further work is needed to determine whether such a proposal can indeed account for the types of syntactic representations seen in human languages, and whether this proposal is indeed functionally distinct from Chomsky's.

Taken together, the three projects presented here make a number of contributions. First, they provide evidence in support of the continued practice of using production and acceptability judgments to probe grammaticality. They use these methods to probe differences in grammatical representations, and in so doing demonstrate that these representations vary across individuals and should be thought of like other types of linguistic objects. Finally, they confirm that theoretical models of syntax are correct in characterizing representations of certain families of structures as unitary, indicating a need for cognitive models of representation and learning to account for a highly complex form of representation.

Bibliography

- Ackerman, L., Frazier, M., and Yoshida, M. (2018). Resumptive pronouns can ameliorate illicit island extractions. *Linguistic Inquiry*, 49(4):847–859.
- Alexopoulou, T. and Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, pages 110–160.
- Altmann, G. T. (2004). Language-mediated eye movements in the absence of a visual world: The ‘blank screen paradigm’. *Cognition*, 93(2):B79–B87.
- Altmann, G. T. and Kamide, Y. (2004). Now you see it, now you don’t: Mediating the mapping between language and the visual world. *The interface of language, vision, and action: Eye movements and the visual world*, pages 347–386.
- Altmann, G. T. and Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, 111(1):55–71.
- Altmann, G. T. and Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive science*, 33(4):583–609.
- Ariel, M. (1999). Cognitive universals and linguistic conventions: The case of resumptive pronouns. *Studies in Language. International Journal sponsored by the Foundation “Foundations of Language”*, 23(2):217–269.
- Asudeh, A. (2004). *Resumption as resource management*. PhD thesis, Stanford University.
- Asudeh, A. (2011). Local grammaticality in syntactic production. *Language from a Cognitive Perspective*, pages 51–79.
- Baayen, R. H. and Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, 3(2):12–28.
- Barr, D. J. (2008). Analyzing ‘visual world’ eyetracking data using multilevel logistic regression. *Journal of memory and language*, 59(4):457–474.

- Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.
- Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- 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.
- Beltrama, A. and Xiang, M. (2016). Unacceptable but comprehensible: the facilitation effect of resumptive pronouns. *Glossa*, 1(1):1.
- Bennett, R. (2009). English resumptive pronouns and the highest-subject restriction: A corpus study. *Trilateral (TREND) Linguistics Weekend, UC Santa Cruz*.
- Berko, J. (1958). The child's learning of english morphology. *Word*, 14(2-3):150–177.
- Bever, T. G. and Poeppel, D. (2010). Analysis by synthesis: a (re-) emerging program of research for language and vision. *Biolinguistics*, 4(2-3):174–200.
- Bjork, R. A. (1994). Memory and metamemory considerations in the. *Metacognition: Knowing about knowing*, 185.
- Bley-Vroman, R. (1989). What is the logical problem of foreign language learning. *Linguistic perspectives on second language acquisition*, 4:1–68.
- Bley-Vroman, R., Felix, S. W., and Loup, G. L. (1988). The accessibility of universal grammar in adult language learning. *Interlanguage studies bulletin (Utrecht)*, 4(1):1–32.
- Bock, K. (1986). Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.
- Bock, K., Cutler, A., Eberhard, K. M., Buttefield, S., Cutting, J. C., and Humphreys, K. R. (2006). Number agreement in british and american english: Disagreeing to agree collectively. *Language*, pages 64–113.
- Bock, K., Dell, G. S., Chang, F., and Onishi, K. H. (2007). Persistent structural priming from language comprehension to language production. *Cognition*, 104(3):437–458.
- Bock, K. and Miller, C. A. (1991). Broken agreement. *Cognitive psychology*, 23(1):45–93.
- Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28.

- Cann, R., Kaplan, T., and Kempson, R. (2005). Data at the grammar-pragmatics interface: the case of resumptive pronouns in English. *Lingua*, 115(11):1551–1577.
- Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).
- Chang, F., Dell, G. S., and Bock, K. (2006). Becoming syntactic. *Psychological review*, 113(2):234.
- Chipere, N. (2001). Variations in native speaker competence: Implications for first-language teaching. *Language Awareness*, 10(2-3):107–124.
- Chomsky, N. (1959). Chomsky, n. 1959. a review of bf skinner’s verbal behavior. *language*, 35 (1), 26–58.
- Chomsky, N. (1981). *Lectures on government and binding*. Number 9. Walter de Gruyter.
- Chomsky, N. (1991). Some notes on economy of derivation and representation. *Anuario del Seminario de Filología Vasca “Julio de Urquijo”*, pages 53–82.
- Chomsky, N. (1992). *The minimalist program*. MIT press.
- Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter.
- Chomsky, N. and Lightfoot, D. W. (2002). *Syntactic structures*. Walter de Gruyter.
- Christiansen, M. H. (2001). *Using artificial language learning to study language evolution: Exploring the emergence of word order universals*. na.
- Christiansen, M. H. and Chater, N. (2015). The language faculty that wasn’t: A usage-based account of natural language recursion. *Frontiers in Psychology*, 6:1182.
- Clemens, L. E., Coon, J., Pedro, P. M., Morgan, A. M., Polinsky, M., Tandet, G., and Wagers, M. (2015). Ergativity and the complexity of extraction: A view from Mayan. *Natural Language & Linguistic Theory*, 33(2):417–467.
- Clemens, L. E., Morgan, A., Polinsky, M., and Xiang, M. (2012). Listening to resumptives: An auditory experiment. In *Poster presented at the 25th Annual CUNY Conference on Human Sentence Processing, New York*.
- Comrie, B. (2008). Prenominal relative clauses in verb-object languages. *Language and Linguistics*, 9(4):723–733.

- Cook, V. J. (1975). Strategies in the comprehension of relative clauses. *Language and Speech*, 18(3):204–212.
- Culbertson, J. (2012). Typological universals as reflections of biased learning: Evidence from artificial language learning. *Language and Linguistics Compass*, 6(5):310–329.
- Culbertson, J. and Adger, D. (2014). Language learners privilege structured meaning over surface frequency. *Proceedings of the National Academy of Sciences*, 111(16):5842–5847.
- Culbertson, J. and Newport, E. L. (2017). Innovation of word order harmony across development. *Open Mind*, 1(2):91–100.
- Culbertson, J., Schouwstra, M., and Kirby, S. (2016). Word order universals reflect cognitive biases: Evidence from silent gesture. In *The Evolution of Language: proceedings of the 11th International Conference*. doi, volume 10.
- Culbertson, J. and Smolensky, P. (2012). A bayesian model of biases in artificial language learning: The case of a word-order universal. *Cognitive science*, 36(8):1468–1498.
- Culbertson, J., Smolensky, P., and Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, 122(3):306–329.
- Culicover, P. W. and Jackendoff, R. (2005). *Simpler syntax*. Oxford University Press on Demand.
- Dabrowska, E. (2008). The later development of an early-emerging system: The curious case of the polish genitive.
- Dabrowska, E. (2012). Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism*, 2(3):219–253.
- Dabrowska, E. and Street, J. (2006). Individual differences in language attainment: Comprehension of passive sentences by native and non-native english speakers. *Language Sciences*, 28(6):604–615.
- Davidson Sorkin, A. (2012). Obama wins battleship – with bayonets.
- Davis, H. (2010). A unified analysis of relative clauses in st’át’imcets. *Northwest Journal of Linguistics*, 4(1):1–43.
- Dickey, M. W. (1996). Constraints on the sentence processor and the distribution of resumptive pronouns. *Linguistics in the Laboratory*, 19:157–192.
- Diessel, H. and Tomasello, M. (2005). A new look at the acquisition of relative clauses. *Language*, pages 882–906.

- Donnelly, S. and Verkuilen, J. (2017). Empirical logit analysis is not logistic regression. *Journal of Memory and Language*, 94:28–42.
- Drummond, A. (2013). IbeX farm. *Online server: <http://spellout.net/ibexfarm>*.
- Dryer, M. S. (2013). Order of relative clause and noun. In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Dunn, M., Greenhill, S. J., Levinson, S. C., and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, 473(7345):79.
- Eilish, B. (2019). Bad guy.
- Enochson, K. and Culbertson, J. (2015). Collecting psycholinguistic response time data using amazon mechanical turk. *PloS one*, 10(3):e0116946.
- Erteschik-Shir, N. (1992). Resumptive pronouns in islands. In *Island constraints*, pages 89–108. Springer.
- Evans, N. and Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and brain sciences*, 32(5):429–448.
- Fadlon, J., Morgan, A. M., Meltzer-Asscher, A., and Ferreira, V. S. (2019). It depends: Optionality in the production of filler-gap dependencies. *Journal of Memory and Language*, 106:40–76.
- Fedzechkina, M., Jaeger, T. F., and Newport, E. L. (2012). Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902.
- Fedzechkina, M., Newport, E. L., and Jaeger, T. F. (2016). Miniature artificial language learning as a complement to typological data. *The usage-based study of language learning and multilingualism*, pages 211–232.
- Ferreira, F., Bailey, K. G., and Ferraro, V. (2002). Good-enough representations in language comprehension. *Current directions in psychological science*, 11(1):11–15.
- Ferreira, F. and Swets, B. (2005). The production and comprehension of resumptive pronouns in relative clause “island” contexts. *Twenty-first century psycholinguistics: Four cornerstones*, pages 263–278.
- Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual review of psychology*, 70:29–51.

- Ferreira, V. S., Bock, K., Wilson, M. P., and Cohen, N. J. (2008). Memory for syntax despite amnesia. *Psychological Science*, 19(9):940–946.
- Ferreira, V. S. and Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, 40(4):296–340.
- Fitz, H., Chang, F., and Christiansen, M. H. (2011). A connectionist account of the acquisition and processing of relative clauses. *The acquisition of relative clauses*, 8:39–60.
- Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard university press.
- Forster, K., Wales, R., and Walker, E. (1976). New approaches to language mechanisms. *Accessing the Mental Lexicon*, pages 257–276.
- Foschi, M. (2000). Double standards for competence: Theory and research. *Annual review of Sociology*, 26(1):21–42.
- Fox, B. A. (1987). The noun phrase accessibility hierarchy reinterpreted: Subject primacy or the absolutive hypothesis? *Language*, pages 856–870.
- Frazier, L. (1987). Syntactic processing: evidence from dutch. *Natural Language & Linguistic Theory*, 5(4):519–559.
- Gart, J. J., Pettigrew, H. M., and Thomas, D. G. (1985). The effect of bias, variance estimation, skewness and kurtosis of the empirical logit on weighted least squares analyses. *Biometrika*, 72(1):179–190.
- Gass, S. (1979). Language transfer and universal grammatical relations. *Language learning*, 29(2):327–344.
- Gass, S. and Ard, J. (1980). L2 data: Their relevance for language universals. *TESOL Quarterly*, pages 443–452.
- Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472.
- Gibson, E. and Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248.
- Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.
- Goodall, G. (2011). Syntactic satiation and the inversion effect in english and spanish wh-

- questions. *Syntax*, 14(1):29–47.
- Greenberg, J. H. (1963). Some universals of grammar with particular reference to the order of meaningful elements. *Universals of language*, 2:73–113.
- Hall, M. L., Mayberry, R. I., and Ferreira, V. S. (2013). Cognitive constraints on constituent order: Evidence from elicited pantomime. *Cognition*, 129(1):1–17.
- Halle, M. and Stevens, K. (1959). Analysis by synthesis. In *Proceeding of the Seminar on Speech Compression and Processing*, volume 2, page D7.
- Halle, M. and Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE transactions on information theory*, 8(2):155–159.
- Han, C.-h., Elouazizi, N., Galeano, C., Görgülü, E., Hedberg, N., Hinnell, J., Jeffrey, M., Kim, K.-m., and Kirby, S. (2012). Processing strategies and resumptive pronouns in english. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, pages 153–161. Cascadilla Proceedings Project Somerville, MA.
- Hatch, E. (1971). The young child’s comprehension of relative clauses.
- Heestand, D., Xiang, M., and Polinsky, M. (2011). Resumption still does not rescue islands. *Linguistic Inquiry*, 42(1):138–152.
- Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, 44(4):311–338.
- Hoffman, M. D. and Gelman, A. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.
- Hofmeister, P. and Norcliffe, E. (2013). Does resumption facilitate sentence comprehension? In *The core and the periphery: Data-driven perspectives on syntax inspired by Ivan A. Sag*, pages 225–246. CSLI Publications.
- Huettig, F. and Altmann, G. T. (2005). Word meaning and the control of eye fixation: Semantic competitor effects and the visual world paradigm. *Cognition*, 96(1):B23–B32.
- Huettig, F., Rommers, J., and Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta psychologica*, 137(2):151–171.
- Jaeger, T. F. and Levy, R. P. (2007). Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856.
- Järvikivi, J., van Gompel, R. P., and Hyönä, J. (2017). The interplay of implicit causality, structural heuristics, and anaphor type in ambiguous pronoun resolution. *Journal of psycholinguistic*

research, 46(3):525–550.

- Kaiser, E., Runner, J. T., Sussman, R. S., and Tanenhaus, M. K. (2009). Structural and semantic constraints on the resolution of pronouns and reflexives. *Cognition*, 112(1):55–80.
- Kako, E. (1999). Elements of syntax in the systems of three language-trained animals. *Animal Learning & Behavior*, 27(1):1–14.
- Kam, C. L. H. and Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive psychology*, 59(1):30–66.
- Kang, S. H., Gollan, T. H., and Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic bulletin & review*, 20(6):1259–1265.
- Kaschak, M. P. (2006). What this construction needs is generalized. *Memory & cognition*, 34(2):368–379.
- Kaschak, M. P., Kutta, T. J., and Schatschneider, C. (2011). Long-term cumulative structural priming persists for (at least) one week. *Memory & cognition*, 39(3):381–388.
- Keenan, E. L. and Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic inquiry*, 8(1):63–99.
- Keenan, E. L. and Hawkins, S. (1987). The psychological validity of the accessibility hierarchy. *Universal grammar*, 15:60–85.
- Keffala, B. and Goodall, G. (2011). Do resumptive pronouns ever rescue illicit gaps in english. In *Poster presented at CUNY 2011 Conference on Human Sentence Processing*.
- Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.
- Kliegl, R., Masson, M. E., and Richter, E. M. (2010). A linear mixed model analysis of masked repetition priming. *Visual Cognition*, 18(5):655–681.
- Koopman, H. (1983). Control from comp and comparative syntax. *The linguistic review*, 2(4):365–391.
- Koornneef, A. W. and Sanders, T. J. (2013). Establishing coherence relations in discourse: The influence of implicit causality and connectives on pronoun resolution. *Language and cognitive processes*, 28(8):1169–1206.
- Kroch, A. S. (1981). On the role of resumptive pronouns in amnestying island constraint

- violations. In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill.*, number 17, pages 125–135.
- Kwon, N., Lee, Y., Gordon, P. C., Kluender, R., and Polinsky, M. (2010). Cognitive and linguistic factors affecting subject/object asymmetry: An eye-tracking study of prenominal relative clauses in Korean. *Language*, pages 546–582.
- Lau, E. (2016). The role of resumptive pronouns in Cantonese relative clause acquisition. *First Language*, 36(4):355–382.
- Levelt, W. J. (1993). *Speaking: From intention to articulation*, volume 1. MIT Press.
- Levy, R. (2008). A noisy-channel model of rational human sentence comprehension under uncertain input. In *Proceedings of the conference on empirical methods in natural language processing*, pages 234–243. Association for Computational Linguistics.
- Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of psycholinguistic research*, 25(1):93–115.
- Luck, S. J. and Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, 390(6657):279.
- 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.
- Luka, B. J. and Choi, H. (2012). Dynamic grammar in adults: Incidental learning of natural syntactic structures extends over 48 h. *Journal of Memory and Language*, 66(2):345–360.
- MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in psychology*, 4:226.
- Marácz, L. K. (1984). Postposition stranding in Hungarian. *GAGL: Groninger Arbeiten zur germanistischen Linguistik*, (24):127–161.
- Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., and Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, 94:305–315.
- McCauley, S. M. and Christiansen, M. H. (2011). Learning simple statistics for language comprehension and production: The cappuccino model. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.
- McClelland, J. L. and Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. an account of basic findings. *Psychological review*, 88(5):375.

- McCloskey, J. (1990/2011). Resumptive pronouns, \bar{A} -binding and levels of representation in Irish. In Hendrick, R., editor, *Syntax of the Modern Celtic Languages*, volume 23 of *Syntax and Semantics*, pages 199–248. Academic Press, New York and San Diego. Republished in Rouveret (2011), pp 65–119.
- McCloskey, J. (2002). Resumption, successive cyclicity, and the locality of operations. *Derivation and explanation in the minimalist program*, 5:184–226.
- McDaniel, D. and Cowart, W. (1999). Experimental evidence for a minimalist account of english resumptive pronouns. *Cognition*, 70(2):B15–B24.
- Miller, G. A. (1962). Some psychological studies of grammar. *American psychologist*, 17(11):748.
- Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., and Fedorenko, E. (2018). High local mutual information drives the response in the human language network. *BioRxiv*, page 436204.
- Montag, J. L. and MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in 8- and 12-year-old children and adults. *Journal of Experimental Psychology: General*, 144(2):447.
- Morgan, A. M. and Wagers, M. W. (2018). English resumptive pronouns are more common where gaps are less acceptable. *Linguistic Inquiry*, 49(4):861–876.
- Nicenboim, B., Logačev, P., Gattei, C., and Vasishth, S. (2016). When high-capacity readers slow down and low-capacity readers speed up: Working memory and locality effects. *Frontiers in psychology*, 7:280.
- Nicenboim, B. and Vasishth, S. (2016). Statistical methods for linguistic research: Foundational ideas—part ii. *Language and Linguistics Compass*, 10(11):591–613.
- Noble, O. (2017). 24 things Trump does better than anybody (according to Trump).
- Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cognitive psychology*, 47(2):204–238.
- Ochs, E. and Schieffelin, B. B. (1994). Language acquisition and socialization: Three developmental stories and their implications. In Blount, B. G., editor, *Language, Culture, and Society*, pages 470–512. Waveland Press, Inc., Prospect Heights, IL, USA.
- Paape, D., von der Malsburg, T., and Vasishth, S. (2019). Quadruplex negatio invertit? The on-line processing of depth charge sentences. Under review at *Journal of Semantics*. Preprint available at <https://psyarxiv.com/uw64a>.

- Park, Y. A. and Levy, R. (2011). Automated whole sentence grammar correction using a noisy channel model. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 934–944. Association for Computational Linguistics.
- Pepperberg, I. M. (1981). Functional vocalizations by an african grey parrot (*psittacus erithacus*). *Zeitschrift für Tierpsychologie*, 55(2):139–160.
- Phillips, C. (2006). The real-time status of island phenomena. *Language*, pages 795–823.
- Pickering, M. J. and Ferreira, V. S. (2008). Structural priming: A critical review. *Psychological bulletin*, 134(3):427.
- Pickering, M. J. and Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, 27(2):169–190.
- Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological studies in language*, 45:137–158.
- Pilley, J. W. and Hinzmann, H. (2013). *Chaser: Unlocking the genius of the dog who knows a thousand words*. Houghton Mifflin Harcourt.
- Polinsky, M., Clemens, L. E., Morgan, A. M., Xiang, M., and Heestand, D. (2013). Resumption in english. *Experimental syntax and island effects*, page 341.
- Pollard, C. and Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.
- Potter, M. C. and Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *Journal of Memory and Language*, 38(3):265–282.
- Prince, E. F. (1990). Syntax and discourse: A look at resumptive pronouns. In *Annual meeting of the berkeley linguistics society*, volume 16, pages 482–497.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Roediger III, H. L. and Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on psychological science*, 1(3):181–210.
- Romoli, J., Santorio, P., and Wittenberg, E. (in preparation). Fixing De Morgan’s laws in counterfactual antecedents.
- Ross, J. R. (1967). Constraints on variables in syntax.

- Saldana, C., Oseki, Y., and Culbertson, J. (2019). Do cross-linguistic patterns of morpheme order reflect a cognitive bias? In Goel, A., Seifert, C., and Freksa, C., editors, *Proceedings of the 41st Annual Meeting for the Cognitive Science Society*, pages 994–1000. Cognitive Science Society.
- Saldana, C., Smith, K., Kirby, S., and Culbertson, J. (2018). Is regularisation uniform across linguistic levels? Comparing learning and production of unconditioned probabilistic variation in morphology and word order.
- Senghas, A., Kita, S., and Özyürek, A. (2004). Children creating core properties of language: Evidence from an emerging sign language in nicaragua. *Science*, 305(5691):1779–1782.
- Shakespeare, W. J. (1996). *Hamlet*. T. J. Spencer (Ed.), The new Penguin Shakespeare. London, England: Penguin Books.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.
- Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic inquiry*, 23(3):443–468.
- Shulman, H. G., Hornak, R., and Sanders, E. (1978). The effects of graphemic, phonetic, and semantic relationships on access to lexical structures. *Memory & Cognition*, 6(2):115–123.
- Slowiaczek, L. M. and Pisoni, D. B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory & cognition*, 14(3):230–237.
- Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, 31(3):575–582.
- Tabor, W., Galantucci, B., and Richardson, D. (2004). Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*, 50(4):355–370.
- Thompson, S. P. and Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language learning and development*, 3(1):1–42.
- Tily, H., Frank, M., and Jaeger, F. (2011). The learnability of constructed languages reflects typological patterns. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.
- Tomasello, M. (2000a). First steps toward a usage-based theory of language acquisition. *Cognitive linguistics*, 11(1/2):61–82.
- Tomasello, M. (2000b). The item-based nature of children’s early syntactic development. *Trends in cognitive sciences*, 4(4):156–163.

- Tomasello, M. (2009). The usage-based theory of language acquisition. In *The Cambridge handbook of child language*, pages 69–87. Cambridge Univ. Press.
- Trueswell, J. C. and Kim, A. E. (1998). How to prune a garden path by nipping it in the bud: Fast priming of verb argument structure. *Journal of memory and language*, 39(1):102–123.
- Ullman, M. T. (2001). The declarative/procedural model of lexicon and grammar. *Journal of psycholinguistic research*, 30(1):37–69.
- Vitevitch, M. S., Luce, P. A., Charles-Luce, J., and Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and speech*, 40(1):47–62.
- von der Malsburg, T., Poppels, T., and Levy, R. P. (2018). Implicit gender bias in linguistic descriptions for expected events: The cases of the 2016 us and 2017 uk election.
- Wagers, M. W., Borja, M. F., and Chung, S. (2018). Grammatical licensing and relative clause parsing in a flexible word-order language. *Cognition*, 178:207–221.
- Williams, G. P., Kukona, A., and Kamide, Y. (2018). Spatial narrative context modulates semantic (but not visual) competition during discourse processing.
- Wu, F., Kaiser, E., and Vasishth, S. (2018). Effects of early cues on the processing of chinese relative clauses: Evidence for experience-based theories. *Cognitive science*, 42:1101–1133.
- Wu, T. (2011). The syntax of prenominal relative clauses: A typological study. *Linguistic Typology*, 15(3):569–623.
- Yan, H., Martin, R. C., and Slevc, L. R. (2018). Lexical overlap increases syntactic priming in aphasia independently of short-term memory abilities: Evidence against the explicit memory account of the lexical boost. *Journal of Neurolinguistics*, 48:76–89.
- Yip, V. and Matthews, S. (2007). Relative clauses in cantonese-english bilingual children: Typological challenges and processing motivations. *Studies in Second Language Acquisition*, 29(2):277–300.
- Zaenen, A., Engdahl, E., and Maling, J. M. (1981). Resumptive pronouns can be syntactically bound. *Linguistic Inquiry*, pages 679–682.