Anastasia Kobzeva

Learning the (un)attested

Modeling acquisition of island constraints using symbolic and neural language models

Thesis for the Degree of Philosophiae Doctor

Trondheim, June 2025

Norwegian University of Science and Technology
Faculty of Humanities
Department of Language and Literature

NTNU
Norwegian University of
Science and Technology

NTNU
Norges teknisk-naturvitenskapelige universitet
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor
Faculty of Humanities
Department of Language and Literature

Learning the (un)attested: Modeling acquisition of island constraints using symbolic and neural language models

ISBN TBA (printed ver.)
ISBN TBA (electronic ver.)
ISSN TBA (printed ver.)
ISSN TBA (online ver.)

Doctoral theses at NTNU,

Printed by NTNU Grafisk senter

Abstract

This dissertation explores the learnability of island constraints — structural configurations that restrict filler-gap dependencies — through computational modeling. Island constraints have traditionally been used to motivate complex innate linguistic knowledge, but recently this view has been challenged by modeling studies that suggest that islands can be learned using domain-general learning mechanisms coupled with fewer domain-specific biases than previously assumed, or perhaps none at all. The models in these studies represent distinct approaches that differ fundamentally in their underlying assumptions: top-down symbolic cognitive models and bottom-up neural language models.

The articles in this dissertation explore whether these modeling approaches can represent viable learning strategies for acquiring filler-gap dependencies and islands cross-linguistically, particularly in light of cross-linguistic variation in island effects. To this end, I apply both types of models to Norwegian, a language that presents exceptions to otherwise cross-linguistically attested generalizations — such as in allowing certain extractions from embedded questions.

The results show that both model types can learn certain generalizations from input data, including the unboundedness of filler-gap dependencies and island status of subject phrases in Norwegian. Notably, both models also captured subtle patterns of cross-linguistic variation in the islandhood of embedded questions. However, each model also exhibited distinct limitations: the symbolic model struggled with data sparsity issues and showed limited generalization, while neural language models exhibited brittle representations of filler-gap dependencies, and occasionally over- or under-generalized from their training data. These findings underscore the importance of input and model biases for the question of island learnability, revealing a trade-off between representational and statistical inference biases. More broadly, the results suggest that while domain-general learners can acquire some aspects of island sensitivity, additional constraints or model modifications may be necessary to fully approximate human linguistic behavior.

Acknowledgements

This dissertation would not have been possible without the support, guidance, and encouragement of many people to whom I am deeply grateful.

First and foremost, I would like to thank my main supervisor, Dave Kush, for his unwavering support, generous mentorship, and enthusiasm throughout this journey. Thank you for always making the time to meet and offer thoughtful feedback, and for cheering me up in difficult times. Your supervision has been instrumental in helping me grow as a researcher. I am also sincerely grateful to my co-supervisor, Tal Linzen, for agreeing to take me on and for welcoming me into his lab. I am equally thankful to my collaborator and co-author, Suhas Arehalli, whose help with model training and evaluation has been invaluable. I want to thank both Tal and Suhas for their valuable input on the papers and stimulating discussions. A special thanks goes to Andrew Weir for offering insightful comments on the dissertation draft and helping me navigate the administrative side of things at NTNU.

I would also like to extend my gratitude to my earlier mentors who played an important role in shaping me as a researcher. My BA supervisor, Maria Ivanova, whom I met during the first year of my undergraduate studies, taught me so much about research over the four years that she supervised me. My MSc supervisor, Irina Sekerina, was the one who encouraged me to apply for this PhD position in the first place and guided me through my EMCL+ journey.

At NTNU, I have been fortunate to be part of an inspiring and supportive community. I am thankful to the Øyelab members — Dave Kush, Anna Giskes, Charlotte Sant, Ingrid Bondevik, Myrte Vos, Marta Velnić, Parker Robbins, Sunniva Briså Strætkvern, Anne Dahl, Terje Lohndal, and Marina Sokolova — for fun and insightful lab meetings. Special thanks to Charlotte, Parker, Myrte, Terje, and Dave for co-authoring a paper with me. This journey would not have been nearly as fun without my “PhD ladies” — Anna Giskes, Ingrid Bondevik, Sunniva Briså Strætkvern, and Kristin Klubbo Brodahl. Thank you for welcoming me into the group, organizing fun social

outings, and being great friends and colleagues. Special thanks go to my *skrivepress* peers — Parichehr Afzali, Tom Lorenz, Natalie Devin Hoage, Guro Stensby Sjuls, et al. — for their motivation and solidarity during the long months of dissertation writing. And to all the friends I've made in Norway: thank you for the joy and much-needed distractions along the way.

I want to thank two research assistants who helped me create experimental stimuli in Norwegian and/or collect behavioral data: Sigurd Farstad Iversen and Anna Brandal. I would also like to thank all anonymous reviewers who read my work and offered constructive feedback. While the review process can sometimes be frustrating (yes, Reviewer 2, I'm looking at you), I genuinely learned a great deal from it — and even enjoyed it on occasion.

Lastly, I want to thank my own family, my family-in-law, and all my friends for their love, support, and occasional "Are you done yet?" questions. I am grateful to my parents, Larisa and Sergey, for always believing in me and supporting me through these years (and all the years before that). Most importantly, I want to thank my now-husband, Daniel, for his love, encouragement, and patience. I doubt I could have completed this journey without him. He supported me in countless ways — from helping me troubleshoot occasional programming issues to keeping me sane through it all (a *no thank you* goes to the pandemic, the Russian-Ukrainian war, and other life crises). I wouldn't have made it through in one piece without you.

Contents

List of abbreviations

AJT	Acceptability Judgment Task
APS	Argument from the Poverty of the Stimulus
ATB	Across-The-Board extraction
BERT	Bidirectional Encoder Representations from Transformer
CDS	Child-Directed Speech
COMP	Complement
COP	Copula
CP	Complementizer Phrase
DEF	Definite
ERP	Event-Related Potential
FBCC	Focus-Background Conflict Constraint
FGD	Filler-Gap Dependency
FGE	Filled-Gap Effect
GPT	Generative Pre-trained Transformer
INF	Infinitive
IP	Inflectional Phrase
LFG	Lexical-Functional Grammar
LLM	Large Language Model
LSTM	Long Short-Term Memory
NLM	Neural Language Model
NLP	Natural Language Processing
NP	Noun Phrase
OI	Optional Infinitive
PCFG	Probabilistic Context-Free Grammar
POS	Poverty of the Stimulus
PP	Prepositional Phrase
PS	Phrase Structure
RC	Relative Clause
ReLU	Rectified Linear Unit

RNN	Recurrent Neural Network
RP	Relative Pronoun
RQ	Research Question
SPR	Self-Paced Reading
UG	Universal Grammar
UGE	Unlicensed Gap Effect
VP	Verb Phrase
XCOMP	Open Complement

Part I

Cover Article

Chapter 1

Introduction

Across many cognitive domains, humans make complex generalizations based on limited experience. This logical puzzle, referred to as Plato's problem, raises questions about the origins of human knowledge. As Chomsky put it, 'How comes it that human beings, whose contacts are brief and personal and limited, are nevertheless able to know as much as they do know?' (Chomsky, 1986, p. xxv). This problem is particularly evident in the domain of language acquisition. Language is a complex hierarchical system, yet typically developing children learn it very fast and seemingly effortlessly from exposure to limited and often impoverished input data. The nature of this learning ability is a subject of ongoing debate in the field. The manifestation of Plato's problem in the context of language is often referred to as the poverty of the stimulus (Clark and Lappin (2011), cf. Pearl (2022) for a useful review) or the logical problem of language acquisition (Baker & McCarthy, 1981).

Children acquire their native languages despite facing a threefold poverty of the stimulus (Chomsky, 1986). First, the linguistic input they receive, also known as child-directed speech or primary linguistic data, is noisy and imperfect. It contains incomplete sentences, speech errors, slips of the tongue, and other inaccuracies, yet children reliably detect the linguistic signal despite this noise. Second, it is often claimed that the linguistic input children receive does not fully represent the complexity of all possible structures that children eventually master. These structures are oftentimes missing from the input, making it impossible to learn them based on direct evidence, yet children learning the same language end up making the same generalization leaps that are not fully supported by the data. Lastly, the input lacks evidence for *impossible* structures, yet children learning the same language avoid making certain generalization leaps that are also not warranted by the input data. In other words, the input is ambiguous, making it difficult for the learner to identify the correct hypothesis among many possible ones. Despite this, children learning the same native language end up converging on the same set of complex grammatical rules. This suggests the existence of certain *inductive learning biases* that guide the acquisition process and limit the set of possible grammars that children consider. This logical argument, known as the argument from the poverty of the stimulus (APS), forms the core of nativist accounts, which have played a prominent role in the developing field of language acquisition.

Linguistic nativist (often referred to as just 'nativist') theories assume that inductive biases, which aid learners in navigating the space of possible hypotheses about language structure, are innate and domain-specific (i.e. language-specific; Chomsky, 1971). Such innate, domain-specific biases are often referred to as *Universal Grammar* (UG) and are thought to represent the most fundamental, invariant properties of all natural languages.[1] For the proponents of this approach (Baker & McCarthy, 1981; Crain & Pietroski, 2001; Laurence & Margolis, 2001; Legate & Yang, 2002; Valian, 2009), the gap between constrained generalizations children make within a relatively short time and the ambiguous input is too broad to be bridged without predefined linguistic constraints. One such solution to Plato's problem was offered by the 'Principles and Parameters' framework (Chomsky, 1993). According to this framework, some innate linguistic constraints that define the range of possible grammar representations are universal (e.g., X-bar theory which defines possible structure of phrases), while others represent more flexible parameters that can take on different values (e.g. whether a language allows dropping subjects like Spanish or not, like English). The input is then mapped onto the space of possible analyses defined by the constraints, setting language-specific parameters in place. Such innate biases dramatically reduce the set of hypotheses that children need to consider when acquiring the grammar of their native languages.

However, the domain-specific answer to the logical problem of language acquisition is not the only possible one. *Empiricist* or *functionalist* theories offer a domain-general proposal for solving the puzzle and postulate that innate biases in humans are not language-specific but rather domain-general. According to such accounts, data-driven learning procedures and biases, such as statistical learning mechanisms and pattern-recognition skills (analogy), which are used

[1] This term is ambiguous. Chomsky (1986, p. 28) acknowledges that in his own work, it is used to both describe specific proposals for innate abilities (e.g. Binding Theory or the Empty Category Principle) and to refer to any type of innate linguistic knowledge more broadly. Clark and Lappin (2010) use it to describe any kind of innate abilities, be they language-specific or not.

more broadly across cognitive tasks, are sufficient for solving the problem of language acquisition (e.g. Chater, Clark, Goldsmith, and Perfors 2015; Clark and Lappin 2011; Reali and Christiansen 2005; Valin Jr 1998). Inasmuch as these abilities are innate, these accounts are also nativist, just non-linguistic nativist (Pearl, 2022). In contrast to linguistic nativists, empiricists hold that the input to children is not as impoverished as the former generally assume, so domain-general mechanisms need not be augmented by domain-specific information in order to learn the full range of structures that children come to master (Pullum & Scholz, 2002; Reali & Christiansen, 2005). Related *connectionist* approaches (Elman, 1990; Elman et al., 1996; Rumelhart & McClelland, 1986) use mathematical models known as *neural networks* to study human mental processes and see language system as an emergent property of language use, which is discoverable from regularities in input patterns.[2]

To this day, there is no agreement about the nature of learning biases that guide language acquisition. This debate has been going on for decades, or even centuries if we consider Ancient Greek theorizing as its start (see, for example, Plato's *Cratylus* and Aristotle's *De Interpretatione* for philosophical reflections on language, logic, and reality, and Plato's *Meno* on the origins of all knowledge in humans). The birth date of the modern strand of this research can arguably be attributed to Chomsky's generative revolution in the 60s of the last century. Since then, numerous experimental contributions to this debate have been made, including language learning experiments in adults and children (Ambridge, Rowland, & Pine, 2008; Culbertson, Smolensky, & Legendre, 2012; De Villiers, Roeper, & Vainikka, 1990; Goodluck, Foley, & Sedivy, 1992; Lidz, Waxman, & Freedman, 2003), corpus analyses (Pullum & Scholz, 2002; Reali & Christiansen, 2005), and computational modeling studies (Freudenthal, Pine, & Gobet, 2006; Pearl & Sprouse, 2013a, 2013b; Perfors, Tenenbaum, & Regier, 2011; Wilcox, Futrell, & Levy, 2023; Yang, 2002). Such studies are often centered around particular linguistic phenomena for which there is supposed to be poverty of the stimulus in the input to children.

Island constraints on filler-gap dependencies (Ross, 1967) are one such phenomenon, and also a focus of the present thesis. Filler-gap dependencies are contingencies between words like *what*, known

[2] Although I presented linguistic nativism, empiricism, and connectionism as three separate classes of approaches to language acquisition, in reality, there is a spectrum of different positions (also within one approach, e.g. different versions of UG), and the lines between them are sometimes blurry.

as fillers, and gaps, or positions where fillers are interpreted in a sentence. Islands are environments that block the formation of filler-gap dependencies across their boundaries, such as the *whether*-clause in (1) (throughout the dissertation, gaps are denoted with an underscore and subscripts indicate coreference).

(1) *What$_i$ do you wonder [whether she bought ___$_i$]?

As I describe below in more detail, islands have been used to motivate complex domain-specific learning biases and have long been regarded as support for innateness theories of language acquisition (see Phillips, 2013a, for an extensive review). However, more recently, this view has been challenged by two groups of computational modeling studies done on English. Pearl and Sprouse (Pearl & Sprouse, 2013a, 2013b, 2015) proposed a distributional learning model for island acquisition, which used a domain-general learning mechanism to reduce the complexity of domain-specific biases as compared to previous generative proposals. Importantly, their model did not eliminate domain-specificity completely but rather reduced it. On the other hand, research by Wilcox and colleagues (Wilcox, Futrell, & Levy, 2023; Wilcox, Levy, & Futrell, 2019a, 2019b; Wilcox, Levy, Morita, & Futrell, 2018) examined the generalizations about the distribution of filler-gap dependencies that neural network-based models develop over the course of their training on raw text. They found that such models can learn many properties of *wh*-filler-gap dependencies — including some islands constraints on their distribution — despite not having any in-built language biases. Their results then led Wilcox and colleagues to conclude that in principle, island constraints are learnable from the input using only domain-general mechanisms and representations, and domain-specificity can be altogether eliminated. In this dissertation, I test whether these two learning models can represent viable learning strategies for acquiring filler-gap dependencies and islands *cross-linguistically*.

It is an acknowledged fact that linguistic research is dominated by studies conducted on English. While this facilitates a more straightforward comparison between studies, in many cases English is not representative of the range of phenomena that must be learned across all natural languages (Blasi, Henrich, Adamou, Kemmerer, & Majid, 2022), including island constraints on filler-gap dependencies. As Wilcox and colleagues point out, “the strongest arguments for or against linguistic nativism will hinge on data about the similarities and differences between languages" (Wilcox, Futrell, & Levy, 2023,

p. 37). To this end, I extend the general line of research initiated by Pearl and Sprouse and Wilcox and colleagues to Norwegian, and compare the results to findings for English. As I explain in more detail in Section 1.3, Norwegian (alongside other Mainland Scandinavian languages) presents an exception to otherwise cross-linguistically attested generalizations about islands, which makes it an excellent testing ground for exploring how variable judgment patterns can be induced from the input. In this work, I explore whether the distributional learning model proposed by Pearl and Sprouse (2013a; 2013b; 2015) and neural language models employed by Wilcox and colleagues (2023; 2019a; 2019b; 2018) could successfully learn islands across a wider range of languages and filler-gap dependency types.

To preview my findings, I find that both model types can learn certain generalizations from input data, including the unboundedness of filler-gap dependencies and island status of subject phrases in Norwegian. Notably, both models also captured subtle patterns of cross-linguistic variation in the islandhood of embedded questions. However, each model also exhibited distinct limitations: the symbolic model struggled with data sparsity issues and showed limited generalization, while neural language models exhibited brittle representations of filler-gap dependencies, and occasionally over- or under-generalized from their training data. These findings underscore the importance of input and model biases for the question of island learnability, revealing a trade-off between representational and statistical inference biases. More broadly, the results suggest that while domain-general learners can acquire some aspects of island sensitivity, additional constraints or model modifications may be necessary to fully approximate human linguistic behavior.

The dissertation is structured as follows: in the remainder of Chapter 1, I introduce filler-gap dependencies, island constraints, and the learnability challenge that they represent (Sections 1.1 and 1.2). The chapter concludes by introducing the variation in island effects observed in Norwegian, and how we can study it to shed light on the origins of islands more broadly (Section 1.3). In Chapter 2, I give an overview of computational modeling as a tool for studying human language acquisition and describe two main types of modeling approaches: symbolic top-down cognitive models and data-driven bottom-up neural language models (Sections 2.2 and 2.3, respectively). In Chapter 3, I give an overview of modeling studies focusing on island learnability in English and how general methodologies used in these studies can be adapted to Norwegian. Chapter 4 gives an overview of five scientific papers that comprise this dissertation. Chapter 5 discusses the main findings of the papers and outlines limitations and directions for future research. Chapter 6 concludes.

1.1 Filler-gap dependencies and island constraints

All natural languages exhibit filler-gap dependencies (FGDs)[3]. This fundamental concept in linguistics refers to contingencies between two positions within a sentence: the *filler*, for example, a *wh*-word *what* in (2-a), and the *gap*, or the position where the filler is interpreted within a sentence. In (2-a), which is a question or a *wh*-filler-gap dependency, there is an interpretive contingency between the *wh*-word and the object of the verb *called*, which is established across a distance. Other types of filler-gap dependencies include topicalizations as in (2-b), relative clauses as in (2-c), and clefts as in (2-d).

(2)
a. What$_i$ are constraints on filler-gap dependencies called ___$_i$?
b. Islands$_i$, Nastia has been working on ___$_i$ a lot.
c. Daniel has learned a lot about the topic$_i$ that Nastia is working on ___$_i$.
d. It is the topic$_i$ that ___$_i$ pops up in all of her papers.

Filler-gap dependencies have played an important role in the development of syntactic theories due to several key properties they exhibit. First, the relationship between the filler and the gap is flexible and structure-dependent: Fillers can license gaps in a variety of syntactic positions, including, for example, subject (2-d), object (2-a), and oblique (2-c) gap positions. Second, filler-gap dependencies are unbounded (Chomsky, 1965): The filler and the gap can appear arbitrarily far apart in a sentence, with several intervening levels of sentential embedding between them, as in (3).

(3) What topic$_i$ did Sergey say (that) Larisa thought (that) Daniel knew (that) Nastia worked on ___$_i$?

Lastly, and most importantly, though filler-gap dependencies are

[3] Often also referred to as long-distance dependencies or unbounded dependencies. The term 'filler-gap dependency' originated from movement-based generative theories that posited that an element like a *wh*-word is fronted leaving a 'trace' in its canonical position (Chomsky, 1973; Ross, 1967). Some other frameworks (e.g. Head-driven Phrase-Structure Grammar, Pollard and Sag (1994); Lexical-Functional Grammar, Kaplan and Bresnan (1995), Construction Grammar Goldberg (1995, 2006)) assume 'trace-less' analyses of such constructions, where the dependency is established between the filler and the head (e.g., the verb *called* in (2-a)). I use the term 'filler-gap dependency' and gap notations simply for convenience and ease of exposition, without commitment to any formal analysis of this phenomenon.

flexible and unbounded, there are *island constraints* on their distribution. The term ‘islands’ was coined by Ross (1967) who first described them as domains that do not allow filler-gap dependencies to be formed across their boundaries (hence, the metaphorical ‘island’ status of such environments). Since then, in a language like English, many different environments have been classified as islands, including coordinate phrases (4-a), embedded questions (4-b), subject phrases (4-c), relative clauses (4-d), and various adjunct clauses (4-e). The examples in (4) are all instances of long-distance filler-gap dependencies crossing island boundaries (denoted with square brackets; the asterisk denotes unacceptability of a sentence).

(4) a. *Which language$_i$ did Sunniva test [___$_i$ and French]?
b. *Which manuscript$_i$ does Dave wonder [when$_k$ Nastia will finish ___$_i$ ___$_k$]?
c. *Which pronouns$_i$ were [all articles on ___$_i$] read by Anna?
d. *Which language$_i$ did Daniel think it was Kristin [who wrote her dissertation in ___$_i$]?
e. *Which variation$_i$ is Ingrid unhappy [if scientists disregard ___$_i$]?

This list of island environments is not exhaustive. With the discovery of more island domains, researchers started to divide them into *weak* and *strong* islands (see Szabolcsi and Lohndal (2017) for a review). Weak islands are domains that allow extraction of certain elements: for example, complex *wh*-phrases (such as *which man*) can arguably be moved out of tenseless *whether*-clauses in English, as shown in (5) (Szabolcsi, 2006, p. 505), making such clauses weak islands in English. Strong islands, on the other hand, never allow filler-gap dependency formation across their boundaries.

(5) Which man are you wondering [whether to invite ___]?

Incorporating weak islands into the list of island domains has resulted in a much longer list, which additionally prompts the question of how children come to acquire these constraints.

1.2 The island learnability challenge

The questions of why islands exist and how children learn about them from limited input data remain unresolved nearly 60 years after their discovery. Islands pose a peculiar learning problem: How do children learn that some unattested structures are possible, whereas other are not? Consider examples in (6) below. If phenomena like noun phrases with prepositional complements and long-distance questions appear in the input in isolation in both subject (6-b) and object positions (6-a), and can even be combined to form long-distance complex object questions (6-c), what prevents learners from combining them to form dependencies like in (6-d) (repeated from (4-c))? The mere absence of such structures in the input does not entail that they are not acceptable. Examples like (6-c) and certainly triply-embedded dependencies like (3) are likely to be absent from the learners' input as well, yet adults judge them as acceptable. If learners generalize beyond their input to learn that examples like (3) and (6-c) are possible, how do they learn that examples like (6-d) are not?

(6) a. Anna read [all articles on cataphoric pronouns].
b. [All articles on cataphoric pronouns] were read by Anna.
c. Which pronouns$_i$ did Anna read [all articles on ___$_i$]?
d. *Which pronouns$_i$ were [all articles on ___$_i$] read by Anna?

The core of the learnability challenge that islands pose is the following: Given the sparse (or even non-existing) relevant data in their input, why do children resist making the leap of generalization that would lead them to island-violating structures, when they clearly make other such leaps (e.g., concluding that filler-gap dependencies are unbounded)? We know for a fact that their caregivers do not tell them that such dependencies are not possible — in other words, there is no explicit negative evidence in the form of corrections (and when adults do provide corrections for other linguistic phenomena, children are shown to disregard them most of the time). Another important part of this learnability challenge is the abstract nature of island constraints. Islands are not tied to specific words or patterns but rather represent abstract rules on the formation of filler-gap dependencies across different clauses.

In terms of developmental trajectories, FGDs and islands seem to be acquired relatively early despite their structural complexity. These constructions rely on hierarchical sentence structure and often involve long-distance co-reference between sentence elements. Previous research has shown that children reliably assign target interpretations to subject and object *wh*-questions as early as 20 months of age (1;8) (Gagliardi, Mease, & and, 2016) and begin producing such structures around 28 months (2;4) on average (Stromswold, 1995). At the same time, children demonstrate sensitivity to island constraints by the age of 3 (De Villiers & Roeper, 1995) or 4 years (De Villiers, Roeper, Bland-Stewart, & Pearson, 2008). This suggests that knowledge of island constraints emerges even before children exhibit adult-like predictive processing of FGDs, which typically develops around age 6 (Atkinson, Wagers, Lidz, Phillips, & Omaki, 2018).

Cross-linguistic consistency in island effects presents an additional piece of the puzzle. Island effects were found to be (mostly) consistent across languages, with domains like coordinate phrases turning out to be islands in one language after the other. All of these pieces of the learnability puzzle led researchers to believe that there must be some learning biases that help children learn about islands. But what do such biases look like, and what is their origin? As foreshadowed earlier, different theoretical frameworks offer contrasting answers to this question. Below I briefly summarize how different linguistic theories answer the question of island learnability.

Linguistic nativists' short answer is that island constraints are not acquired, they are innate. As Phillips put it, islands "are obscure and abstract, and they are a parade case of a linguistic phenomenon that is likely to be difficult to observe in the input that children must learn from. As such, they have been regarded as a good example of the need for Universal Grammar." (Phillips, 2013a, p. 132). In other words, based on the argument from the poverty of the stimulus, nativist approaches assume that there are complex predefined domain-specific constraints on what is a possible dependency that are crucial to the acquisition process.

Many attempts have been made to describe what such constraints might look like. The most prominent *syntax-based* accounts include Chomsky's Subjacency Conditions and its successors (Chomsky's *Barriers* (1986) and the *Phase Impenetrability Condition* (2001)), as well as Huang's Condition on Extraction Domains (1982). Central to these approaches is the assumption that there are constraints on movement operations through which filler-gap dependencies are formed. For example, Subjacency and its successors posit that movement must proceed in a step-wise manner through designated intermediate positions. Island effects arise when certain syntactic configurations block this step-wise movement, effectively acting as "barriers" in the path from the gap to the filler's surface position. In this view, island constraints emerge as a by-product of general, and arguably innate, principles governing syntactic movement.

To illustrate: Chomsky's Subjacency Condition (Chomsky, 1973) stated that a dependency cannot cross more than one bounding node — a certain phrase type intervening between the gap and the filler — in one cyclic application of a movement rule. In English, NP (DP) and IP (S or TP) were considered bounding nodes. Subjacency was proposed to explain several island effects simultaneously, including *wh*-, subject, and complex noun phrase islands. For example, the *wh*-dependency in (7), repeated from (4-b), has two intervening IP nodes between *which manuscript* and its gap site. To move to its surface position, *which manuscript* would have to move in two steps, only crossing one IP node at a time, stopping first in the specifier of the CP. Since that position is occupied by *when*, the movement is prohibited, making dependencies into embedded *when*-questions in English ungrammatical according to Subjacency.

(7) *Which manuscript$_i$ does [$_{\text{IP}}$ Dave wonder [$_{\text{CP}}$ when$_k$ [$_{\text{IP}}$ Nastia will finish ___$_i$ ___$_k$]]]?

Non-syntactic nativist explanations have also been proposed. Several *functionalist* or *constructionist* approaches propose to explain island effects as arising from a clash of information-related properties of the filler and/or its gap site (Abeillé, Hemforth, Winckel, & Gibson, 2020; Cuneo & Goldberg, 2023; Erteschik-Shir, 1973; Goldberg, 1995, 2006).[4] The key observation for such accounts is that filler-gap dependencies like questions and topicalizations are characterized not only by their word order and syntactic structure but also by their discourse function, as determined by their usage. Several (interrelated) discourse or information-structure measures were proposed to account for island effects (e.g., dominance, focus, topic, salience, relevance, backgroundedness; see Liu, Winckel, Abeillé, Hemforth, & Gibson, 2022, for a review). This group of approaches argues for a usage-based model of grammar that is acquired due to domain-general cognitive biases and learning mechanisms, such as a bias towards cooperative communication, statistical learning abilities, and pattern recognition through analogy.

For example, the Focus-Background Conflict constraint (FBCC) of Abeillé et al. proposes that ‘A focused element should not be part of a backgrounded constituent’ (Abeillé et al., 2020, p. 3). Focused elements like *wh*-words in questions represent new or important information, while backgrounded elements like subject phrases are usually introduced earlier in the discourse, and therefore convey already given information. Questioning a complement of a subject phrase then leads to an information clash, where the focused *wh*-word is part of a backgrounded constituent as in (8-a) below. On the

[4] *Semantic* explanations to some island effects have also been proposed, see Abrusán (2014), and Szabolcsi and Lohndal (2017). For a review of non-syntactic explanations to islands, see Newmeyer (2016).

contrary, constructions like relative clauses usually convey already given, backgrounded information. The FBCC would then predict that it should be possible to relativize a complement of a subject noun as in (8-b).[5] The discourse-based/functionalist accounts like FBCC thus predict cross-construction variation in island effects, whereas traditional syntax-based accounts treat *wh*- and RC-dependencies the same (as one homogeneous class — so-called A'-dependencies).

(8) a. *Of which car did the color ___ delight the baseball player?
b. The dealer sold a car of which the color ___ delighted the baseball player.

Another class of approaches shifts the debate on islands from levels of linguistic analyses into the realm of language *processing*. According to such processing (or reductionist, in the terminology of Phillips 2013b) frameworks, island effects arise because of constraints on the processing resources (e.g. working memory capacity) of a human language processor. Such accounts do not view sentences with island violations as ungrammatical or infelicitous but rather too complex to process (Hofmeister & Sag, 2010; Kluender & Kutas, 1993; Liu, Winckel, et al., 2022). According to this view, both long-distance dependencies and island environments are associated with increased processing load in isolation, and taken together they overload the processor, making dependencies into islands to be *perceived* as unacceptable (Hofmeister & Sag, 2010; Kluender & Kutas, 1993; Liu, Winckel, et al., 2022).

To date, there is no agreement about which one of these major types of approaches can account for the majority of empirical data on island effects. Importantly, these approaches are not necessarily mutually exclusive. It is plausible that some island effects are best captured by categorical domain-specific constraints (e.g. coordinate structure island), while others are best explained by frequency-based processing accounts that predict gradient acceptability (e.g., factive and manner-of-speaking islands, cf. Liu, Ryskin, Futrell, & Gibson, 2022). In other words, island effects are not created equal: as the body of experimental work has grown, scholars have uncovered cross-linguistic and cross-construction patterns of variation in island effects. Both types of variation are potentially useful for the refinement of our theoretical frameworks: while cross-construction variation can be espe-

[5] According to Abeillé et al., analogous sentences with preposition stranding like *'The dealer sold a car which the color of ___ delighted the baseball player.'* are less acceptable due to difficulty arising from preposition stranding, independent of the island effect.

cially useful for improving our explanatory categories, cross-linguistic variation gives researchers an opportunity to assess claims about the nature and learnability of island constraints. This dissertation primarily focuses on the latter by modeling filler-gap dependencies and island effects in Norwegian, which are described in the next Section.

1.3 Island constraints in Norwegian

Island effects were long thought to be consistent across languages — as the product of innate constraints or limitations on processing resources, they should allow little to no cross-linguistic variation (Phillips, 2013a; Stepanov, 2007). Some of the first described islands, like the coordinate structure island in (4-a), have been attested in language after language, and still remain good candidates for linguistic universals (Liu, Winckel, et al., 2022; Phillips, 2013a).

However, with the development of theoretical frameworks, scholars have uncovered some patterns of cross-linguistic variation and differences between island types. To illustrate, consider an example (9) from Rizzi (1982, p. 50), which shows that Italian seemingly violates Chomsky's Subjacency condition by allowing extraction from embedded questions:

(9) Tuo fratello, a cui$_i$ mi domando [che storie$_k$
your brother to whom myself wonder.1SG what stories
abbiano raccontato ___$_i$ ___$_k$], era molto preoccupato.
have.SBJN.3PL told was very worried
'Your brother, to whom$_i$ I wonder [which stories$_k$ they told ___$_i$ ___$_k$], was really worried.'

To accommodate such acceptable examples from Italian, Rizzi proposed that the definition of a bounding node in Chomsky's Subjacency condition is language-dependent: while in English IP/TP/S and NP/DP were considered to be bounding nodes, CP/S' and NP/DP were proposed for Italian (Rizzi, 1982). Another pattern of variation came from Hebrew, based on which Reinhart proposed that there might be a second complementizer slot in the syntactic analysis of sentences with FGDs to account for the acceptance of dependencies into *wh*-clauses in the language (Reinhart, 1981). Such examples from Italian and Hebrew led to modifications of prior theoretical proposals by adding language-dependent parametrization in the Principles and Parameters framework (Reinhart, 1981; Rizzi, 1982). According to such parameterized generative proposals, learning involves setting language-specific parameters (e.g., the precise definition of a bounding node, or the number of complementizer slots) in place based on the positive evidence found in the input, while the set of parameters and their possible values is innately specified by the UG.

Norwegian (alongside other Mainland Scandinavian languages) represents a case of variation in island effects and the topic of the present dissertation. While Norwegian exhibits sensitivity to a lot of the same islands that English is sensitive to, like the subject island in (10) below, there are also some important exceptions. In particular, prior traditional work (e.g., Christensen, 1982; Engdahl, 1982; Erteschik-Shir, 1973) and recent experimental research (Kush, Lohndal, & Sprouse, 2018, 2019) show that embedded questions and relative clauses appear to allow for (some) filler-gap dependencies to be formed across their boundaries, as illustrated in examples (11) and (12) below.

(10) *Wh*-dependency into a subject island:

*Hva$_i$	har	brevet	om	___$_i$	skapt	problemer?
What	has	letter.DEF	about		created	problems

‘*What$_i$ did the letter about ___$_i$ create problems?’

(11) RC-dependency into an embedded question:

Compsognathus	er	en	av	de	få	dinosaurene$_i$	vi	vet	hva$_k$
Compsognathus	COP	one	of	the	few	dinosaurs.DEF	we	know	what

___$_i$	spiste	___$_k$.
	ate	

lit. ‘Compsognathus is one of the few dinosaurs$_i$ we know what$_k$ ___$_i$ ate ___$_k$.’[6]

(12) Topicalization-dependency into a presentational relative clause:

Resultatene$_i$	er	det	ikke	bare	vi	som	mener	___$_i$	er
Results.DEF	COP	it	NEG	only	we	REL	believe		COP

enestående …
unique

lit. ‘The results$_i$, it’s not just us who believe ___$_i$ are unique …’[7]

It is unclear how the type of cross-linguistic variation observed in island sensitivity is acquired from the input to Norwegian-learning children. The patterns observed in Norwegian imply that the input to Norwegian learners is structurally different from the input to the learners of English. How exactly the two inputs differ is an open empirical question. Recent corpus findings (Kush, Sant, & Strætkvern,

[6] Source: Compsognathus Wikipedia page

[7] Source: NRK’s article “Gatelagsatsing sparer staten for nesten 100 millioner kroner”

2021) suggest that there is some direct evidence of structures like (11) and (12) in a corpus of child-directed texts, but it is rather infrequent and not representative of the range of constructions that Norwegian speakers consider as acceptable. Is this direct evidence 'enough' for the learners to uncover the patterns of island-insensitivity in Norwegian, and what kind of biases do the learners need to arrive at the correct generalizations? Are the biases proposed for acquiring island constraints in English (Pearl & Sprouse, 2013b; Wilcox et al., 2018) flexible and powerful enough to guide learners toward a different target state when exposed to structurally distinct input, such as in Norwegian? In other words, do the proposed learning models represent viable theories for learning about island facts cross-linguistically? To answer these questions, I use computational modeling to explore whether models with different types of inductive biases can mimic the linguistic behavior of Norwegian speakers based on the input they receive.

Chapter 2

Computational modeling in language acquisition research

For centuries, philosophers, scholars, and rulers alike have been intrigued by the nature of human language development. As reported by Herodotus, Egyptian Pharaoh Psamtik I (663-610 BC) carried out a troubling language deprivation experiment with two infants to see what language the children would acquire from birth. Similar experiments were reportedly conducted by later royal rulers, including Frederick II of Sicily (1192–1250) and James IV of Scotland (1473–1513; see Campbell & Grieve, 1981, for a discussion on the authenticity of these investigations). Fortunately, modern researchers do not resort to such unethical and disturbing methods to study language acquisition. At the same time, these examples represent just the tip of the iceberg of numerous challenges that research in language acquisition faces: it is impossible to alter or influence a child's input due to ethical considerations. Other challenges include difficulty of quantifying child input, as corpus data are not easy to collect and are time-consuming to transcribe, often making it altogether impossible for low-resource languages. Moreover, conducting behavioral experiments with children is challenging because their developing non-linguistic abilities can hinder controlled experimentation. Most importantly, we cannot directly observe cognitive processes that guide children to correct conclusions about their target grammar, their initial state of learning (are they navigating a blank slate?), nor the inference mechanisms that guide the acquisition process. All of the above are components that informative acquisition theories should be able to characterize with precision.

Computational cognitive modeling offers a powerful solution to these issues. Cognitive models are used to hypothesize about the complex cognitive processes that happen in the human mind, including language development. By simulating language acquisition processes, computational modeling allows researchers to make all components of an acquisition theory explicit. It might seem unexpected at first that making the components of an acquisition theory explicit is a benefit of modeling — would not one expect verbal theories[1] to do a good-enough job at specifying them? There is evidence that even when a popular verbal theory seems to describe such components precisely enough, computational models still offer a more detailed account of the learning process and make detailed predictions that better align with empirical data (Jones et al., 2014). The components that go into a model usually include the initial state, the representations to be learned, the learning mechanisms, the learning period, the target state, and the input data. Apart from the input data and the child's linguistic behavior (output), the components that are hypothesized to be relevant to the learning process are latent, which makes it a non-trivial task to sufficiently characterize them in order to be implemented in a model. A benefit of such characterization is that it enables researchers to evaluate whether a learning theory works, but also helps determine precisely what makes it work or fail (Pearl, 2023a, 2023b). To summarize, modeling experiments allow us to 1) identify all the components relevant to an acquisition theory in a precise enough manner to implement them in a model; 2) compare the modeling results against empirical behavioral data; 3) assess what makes the implemented theory work or fail and 4) iteratively refine our theories, providing potential insights into the underlying processes in humans. Moreover, experiments with computational models provide researchers with a high level of control over the learning environment. Unlike human learners, whose linguistic input is highly variable and difficult to track comprehensively, computational models can be trained on datasets that are carefully curated and systematically manipulated. This allows researchers to test specific hypotheses about the role of different kinds of evidence found in the input, and how different language phenomena are acquired based on this evidence.

[1] *Verbal* theories provide theoretical explanations for different empirical data, identifying cognitive mechanisms that are relevant for observed behavior (e.g., phonological working memory plays a primary role in nonword repetition, Gathercole & Baddeley, 1989). Here I juxtapose them with *computational* accounts that provide implementation-level explanations of observed phenomena and make specific predictions for novel data (e.g., a computational implementation of the chunking hypothesis that learns long-term phonological knowledge from naturalistic data and interacts with short-term memory capacity to perform nonword repetition, cf. Jones, Gobet, Freudenthal, Watson, & Pine, 2014).

2.1 Components relevant to language acquisition task

There have been several proposals for characterizing the components relevant to language acquisition task (e.g., Lidz and Gagliardi 2015; Omaki and Lidz 2015; Pearl 2023a, 2023b). It is generally agreed that there are both internal and external components, with the latter including child input and behavior. The *external* components are, in principle, observable: it is possible to observe (and subsequently transcribe) child-directed speech, as well as children's naturalistic production. The CHILDES database is one such valuable source of transcribed (and sometimes annotated) child-adult interactions (MacWhinney, 2000). It is also possible to probe children's developing linguistic representations through clever experimental designs. The diversity of available methods depends on a child's age, and modern experimental techniques allow researchers to conduct experiments from infancy and through all stages of development (Perkins & Lidz, 2023; Syrett, 2023).

There has been less agreement about the *internal* components that enable language acquisition. Several recent proposals characterizing such components come from Omaki and Lidz (2015), later more fully described by Lidz and Gagliardi (2015), and more recently adapted by Pearl (2023a, 2023b). The proposed model of language acquisition is presented in Figure 2.1. The internal components are organized in clusters. The input data that can be perceived by the child at this stage (e.g., segmented and parsed given the cognitive abilities in the perceptual encoding cluster) results in *perceptual intake*. In production, the child also relies on the representations that can be perceptually encoded at this stage of development. Both perceptual encoding and production clusters include extralinguistic systems, which limit the amount of information that can be perceived, encoded, and generated by the child. Examples of such systems include attention and memory limitations, theory of mind considerations, etc. The main component relevant to language acquisition — namely, the learning component — includes applying filters and biases to the encoded information to produce *acquisitional intake*. Acquisitional intake is part of the perceptual intake that is useful for comparing possible features of the developing grammar against the set of possible grammars defined by a set of hypothesized biases (be they domain-specific like UG or domain-general like bias for simplicity, Perfors et al. 2011). The developing grammar is iteratively updated by performing inference on the acquisitional intake in the *inference engine* cluster. The inference engine typically involves non-linguistic abilities like probabilistic inference or some sort of statistical learning. The developing grammar in turn feeds the perceptual processes through which new input is analyzed, and the production systems that generate observable behavior.

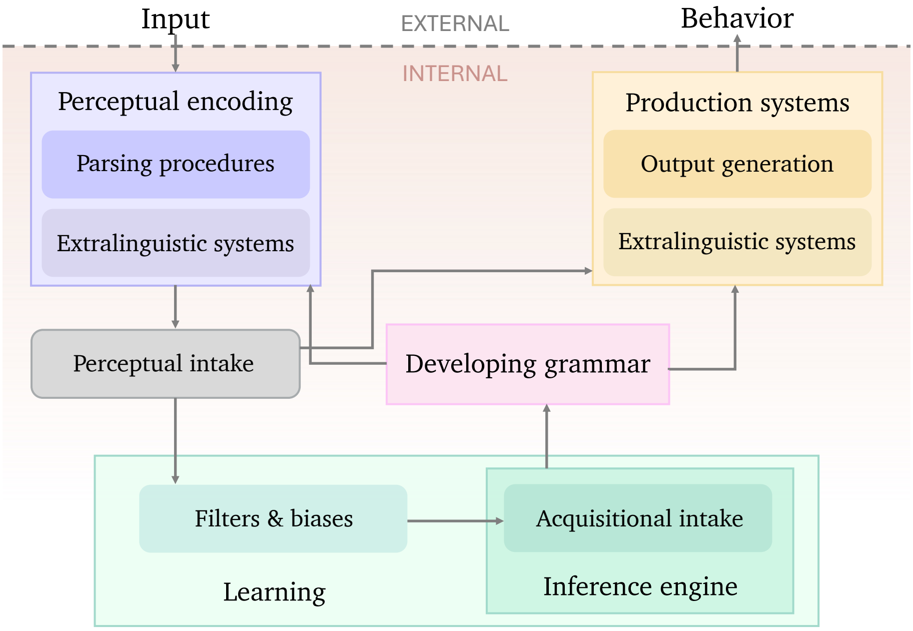


Figure 2.1: Components relevant to the acquisition process, adapted from Pearl (2023a, 2023b). Observable external components include input and behavior. Internal unobservable components include perceptual encoding of the input signal (resulting in perceptual intake), learning from the encoded information by applying filters and biases to produce acquisitional intake, and performing inference on this intake. The inference engine updates the developing grammar, which in turn feeds the perceptual processes through which new input is analyzed and the production systems that generate observable behavior.

To date, there is no consensus about the details of the internal components. What are the specific filters and biases that apply to perceptual intake? What kind of representations constitute the developing grammar? What is the particular inference mechanism used? Depending on a linguistic phenomenon at hand, models for its acquisition might give different answers to these questions. The proposed acquisition process offers a *general framework* which is useful to keep in mind when evaluating acquisition models. In this dissertation, I focus on two particular types of models — traditional symbolic cognitive models and neural language models — that have very different assumptions about the types of the components that go into the model, the types of representations that the model learns, and the way inference operates over those representations.

2.2 Symbolic cognitive models

Traditional cognitive models approximate the acquisition processes that are hypothesized to take place in the human mind. Such *symbolic cognitive models* usually commit to some symbolic representations (e.g., words that are organized into phrases, and phrases that are organized into hierarchical structures) and inference mechanisms that are specified by a certain acquisition theory (e.g., Perfors et al. 2011). The acquisition process specified in Figure 2.1 is quite complex, and implementing all relevant components in a single model is a challenging task. Therefore, in the field of language acquisition modeling, it is not uncommon to focus on a specific part of this process, whether it relates to a specific level of linguistic representation (e.g., phonetics/phonology) or a specific feature of the developing grammar (e.g., whether a language has a *wh*-movement), abstracting away from other components that are deemed less relevant to the process being modeled. In other words, computational models are inevitably simpler than the target system they aim to describe. This discrepancy between the model and the target system has led to many scientific controversies, as all of the components are thought to be highly interrelated. Our methodology-driven choice to abstract away from something can always be questioned by others who believe that this 'something' is crucial and cannot be omitted (Frank, 2023b).

Which components usually go into a symbolic computational model, and how do we implement them in practice? This choice depends on the process being modeled, and here I will mostly focus on models of *syntactic* acquisition. Based on a recent review by Pearl (2023b), I characterize below the modeling components that go into models of syntactic acquisition.

- **Initial state:** These are the abilities and representations that the modeled learner is supposed to have prior to the acquisition of some syntactic property. For instance, it is reasonable to assume that speech stream segmentation abilities as well as some parsing abilities are a prerequisite for acquiring higher order syntactic rules such as *wh*-movement. It is not uncommon that the initial state also (sometimes implicitly) includes learner's extra-linguistic abilities such as working memory capacity or theory of mind considerations. In other words, the initial state for a model describes all of the clusters identified in Figure 2.1 at their initial stage.

- **Input:** This modeling component seems to be natural to describe as the following: 'The input to the model is an approximation of child-directed input'. However, this is not always the case: many existing models of syntactic acquisition equate model's input to perceptual intake in Figure 2.1, given some idealized perceptual encoding capacities (i.e., when all of the model's input is relevant for updating the developing grammar, and no information is 'lost' on the way). In other work, input representations can already be shaped by certain biases (e.g., bias for hierarchical structure of language), which then equates such input to the hypothesized acquisitional intake from Figure 2.1 (Pearl & Sprouse, 2013a, 2013b). However, with computational advances and the development of theories in cognitive domains other than language, modern models start to incorporate specific hypotheses about perceptual encoding abilities and extralinguistic constraints, where the model's input becomes more closely related to human input (for instance, see various models implementing memory constraints and biases, such as working memory capacity and recency effects: De Santo 2020; Dickson, Futrell, and Pearl 2024; Freudenthal, Pine, Aguado-Orea, and Gobet 2007; Freudenthal et al. 2006, as well as Pearl 2023a for a recent review).

- **Filters and biases:** As foreshadowed in the introduction, the nature of inductive biases in human language learning is a polarizing topic. Researchers often struggle to describe them precisely enough in verbal models, creating even more controversy around an already contentious topic. In computational cognitive models, however, such filters and biases need to be explicitly spelled out to be implemented into a model. The filters and biases that have been proposed in previous research can be broadly grouped into three main types. The first type is a bias on *data representation*: do learners consider all possible representations for the input data, or do they have some pre-specified biases for how the data should be represented? An example of such a bias can be a preference to represent language structures hierarchically, as described by Universal Grammar, rather than linearly. Another type of biases concerns *subsetting input data*: do learners learn from all of the data available in the input, or do they filter/subset it in a specific way to make the learning process more feasible? Examples of such biases include a preference to only learn from unambiguous data (Pearl, 2008), chunking of data into smaller pieces (Pearl & Sprouse, 2013b), or subsetting the input to include only the most relevant data for learning (e.g., learn mainly from main clauses while initially disregarding embedded clauses, Lightfoot 1991). Additionally, there are biases on the *inference mechanism*: what kind of procedures do learners use when updating the developing grammar (see examples below)? Importantly, there is likely to be a *trade-off* between the complexity of biases that go into the model: a modeled learner can equally succeed with weak representational biases and strong inference mechanisms, or with simpler inference procedures operating over more sophisticated input representations/more restricted subset of input data.

- **Inference mechanism:** This is a type of mechanism that updates the developing grammar given new data points of acquisitional intake and the hypothesized filters and biases. Nowadays, it is widely agreed that the inference mechanisms in humans include some form of *statistical learning*[2] (Krogh, Vlach, & Johnson, 2013). The most common examples of such mechanisms include *sensitivity to frequency patterns* in the input data (i.e., counting different types of elements in the input), *reinforcement learning* where the probabilities of competing representations are updated in a specific way depending on whether they can account for a new data point or not, and *Bayesian inference* that involves prior assumptions about the probability of different options (for a more exhaustive list of options and their detailed descriptions, see Pearl, 2023b). In modern computational models, it is common to assume that such statistical learning mechanisms are *domain-general* in nature; for instance, Bayesian models are successfully employed in many different domains of cognition such as visual perception, motor control,

[2] Previously, though, *triggering models* of language acquisition assumed that the features of the target grammar were learnt based on a few relevant examples, without the need for statistical inference, see Yang (2002).

pattern recognition, decision making, etc. (Chater, Oaksford, Hahn, & Heit, 2010; Griffiths, Kemp, & Tenenbaum, 2008).

- **Learning period:** This component refers to the phase during which the learner acquires and refines the target knowledge about the phenomenon in question. Different stages of the learning period can be modeled by manipulating the amount and the type of input data (corresponding to either perceptual or acquisitional intake) on which inference is performed in order to update the developing grammar.

- **Target state:** This typically refers to the grammatical knowledge that adult speakers possess regarding the modeled phenomenon. Alternatively, it can denote a proxy of children's knowledge at a specific developmental stage if the model focuses on earlier stages. Target state can describe both specific grammatical constraints that a language obeys (e.g., in English, verbs must agree with their subjects in number) or some behavioral data collected from a representative group of native speakers (e.g., acceptability judgment data assessing passivizability of certain verbs, Leong & Linzen, 2024). Given that complex linguistic phenomena (such as islands) likely arise from a combination of different factors (e.g., syntactic and extra-syntactic), researchers need a comprehensive way to compare model generalizations to the target generalizations. Since grammatical knowledge is latent, such characterization sometimes can only be achieved through controlled experiments.[3]

To build a symbolic computational model, all of the above-mentioned components should be specified. This allows researchers to 'look inside' the model when assessing its successes and failures. Importantly, if the model succeeds, this lends support to a particular acquisition theory and describes how acquisition *could* proceed, but it does not mean that it necessarily proceeds this way. Since the the majority of the components relevant to the learning task are unobservable, and there is likely to be a trade-off between their complexity, different model

[3] When conducting controlled experiments, researchers hope that the design and the stimuli effectively isolate the relevant grammatical knowledge, i.e., *competence*, (Chomsky, 1965), from *performance* issues and random noise, but this is not always guaranteed. Unless performance pressures are explicitly specified in a model (e.g., De Santo 2020; Dickson et al. 2024), the modeling target state is an approximation of competence, which is in turn inferred from measurable performance. This distinction is important to keep in mind when evaluating modeling results.

implementations can appear to be equally successful at representing a particular target state. To give an example, below I review two different types of computational learning models that implement different theories explaining why in some languages children go through an optional infinitive stage in their linguistic development (Haegeman, 1995; Wexler, 1994).

The phenomenon of *optional infinitives* (OI) or *root infinitives* refers to the observation that in some languages, during early stages of development, children often use non-finite verb forms (infinitives) where finite forms (tensed verbs) are required. Some examples are provided in (1). Interestingly, there appears to be some cross-linguistic variation in OI frequency and duration of the OI stage. For instance, children learning languages with rich verb morphology (like Spanish or Russian) tend to move past the OI stage more quickly than those learning languages with poorer morphological systems (like English).

(1) a. Papa have it (English)
b. Zähne putzen (German)
teeth brush-INF
c. Mama spat' (Russian)
mother sleep-INF

Legate and Yang (2007) presented a *variational learning model* which embodies a parametric account for learning OI cross-linguistically. This model posits that during learning, children set language-specific parameters based on the positive evidence found in their input. Therefore, this model presupposes that there is a strong domain-specific bias in the form of UG that narrows down the hypothesis space that the learner navigates during the acquisition process. Instead of navigating a blank slate, the child only needs to learn about the specific parameter settings of their target grammar, for which possible values are innately specified by the UG. The parameter they proposed as relevant to the OI stage is [±Tense]. [+Tense] languages, like Spanish, express tense morphosyntactically, whereas [−Tense] languages, like Chinese, do not mark tense on verbs but use other means such as particles expressing verbal aspect.

According to Legate and Yang, [+Tense] and [−Tense] grammars compete with each other during learning. This competition is probabilistic, with each parameter option having an initial probability. The modeled learner uses reinforcement learning mechanism to update these probabilities: each example with overt verbal morphology in the acquisitional intake rewards the [+Tense] option and punishes

[−Tense] option, which corresponds to increasing and decreasing their probabilities, respectively. Consequently, the learners gradually increase their preference for the correct grammatical form based on the statistical properties of the input they receive. Over time, the correct grammar becomes more dominant as it is reinforced by the input. Legate and Yang's variational learning model therefore presupposes that children's acquisitional intake includes (at least partially) parsed sentences that are relevant for parameter-setting in the developing grammar.

Languages with richer verbal morphology systems provide more cues to the learners, influencing the rate at which they overcome the OI stage. Legate and Yang tested this prediction on three languages — Spanish, French, English — and found an alignment between model predictions and empirical data. Their variational learning model thus explained the brief OI stage in Spanish, the prolonged OI stage in English, as well the intermediate status of the OI stage in French.

Another type of model that was proposed for explaining the OI stage is the MOSAIC (Model of Syntax Acquisition in Children) model (Freudenthal et al., 2006). MOSAIC employs an incremental probabilistic learning mechanism that interacts with the distributional properties of the input. Compared to the variational learning model described above, this model exhibits fewer (if any) domain-specific biases. It is characterized by a bias toward attending to utterance-final elements, combined with domain-general sensitivity to frequency, and mechanisms for chunking and generating novel utterances. The model is resource-limited, meaning that it processes input incrementally and is biased toward learning from the ends of utterances, akin to recency-based memory effects. This bias mirrors the observed tendency in children to focus on utterance-final elements, which often include the infinitive if the utterance contains a compound finite verb with an auxiliary or a modal at the beginning (e.g, *Can papa* ***have it?***).

The model simulates how children gradually acquire the correct use of finite verb forms by building a network of interconnected nodes (i.e., sentence elements such as words or phrases) that are learned from child-directed speech. As children hear more examples of correct verb usage, they learn to produce progressively longer utterance-final phrases, leading to a decrease in OI errors as a function of the amount of input to which they are exposed. In the model, this process is implemented as the network growth, where new words are added to the network and links are established between sentence elements based on their distributional properties. Sentence elements that occur in similar contexts become connected by generative links, which allow the model to generate novel utterances based on the utterances already encountered in the input. Compared to the variational learner described above, this model includes very different assumptions about the biases that shape the acquisitional intake and the representations in the developing grammar.

Freudenthal et al. tested their model production on a range of languages including English, Dutch, German, French, and Spanish. Their findings showed that MOSAIC could account for the cross-linguistic variation in the rate of OI errors. Moreover, they found that the model captured other cross-linguistic differences between the tested languages (such as differential use of modal constructions in German and Dutch; Freudenthal et al., 2007, p.332), providing a larger coverage of empirical data than the variational learner model of Legate and Yang. The authors concluded that together, the utterance final bias and the distributional statistics of the input are sufficient to explain the variation in the occurrence of OI errors across languages.

2.3 Neural language models

A distinct modeling approach that has been shown to provide a broad coverage of empirical data uses *neural language models* to study language learnability, rooted in the connectionist tradition in linguistics (Elman, 1990; Elman et al., 1996; Rumelhart & McClelland, 1986). Neural language models are instances of artificial neural networks — computational systems inspired by the interconnected neuronal system in the human brain. In essence, an artificial neural network is a network of small computing units, often referred to as neurons or cells. Each cell takes a vector of input values, processes them using a set of *weights* and a function (so called ‘activation function’), and produces a single output value.[4] The representational power of modern neural networks arises from the combination of these cells into larger networks, which are organized into multiple layers. These layers usually include an input layer, one or (typically) more hidden

[4] Apart from weights, *biases* also influence output calculation. In this overview, I omit this term to avoid confusion with *learning biases* discussed earlier. In essence, weights are coefficients that input values are adjusted by, while biases are constants that are added to the adjusted value to produce the output. Simplifying somewhat, output calculation in a neural network can be seen as an extension of a simple linear function where weights are analogous to slopes and biases are similar to intercepts. The added complexity then stems from two facts: (i) weights and biases are typically matrices and vectors, respectively, whereas slopes and intercepts are real numbers; (ii) a non-linearity in output calculation is introduced by passing the output through an activation function (e.g. sigmoid, ReLu, or tanh).

layers, and an output layer. The computation of outputs proceeds iteratively from one layer to the next. This layered structure allows neural networks to model and capture various non-linear relationships between inputs and outputs, making them powerful tools for various tasks including natural language processing (NLP). The success of neural language models in the field of NLP led to an increased interest in using them as tools or subjects in linguistics, sparking a round of debates about their potential contribution to (the evaluation of) linguistic theories; see Pater (2019) and responses it generated (Dunbar 2019; Linzen 2019; Pearl 2019; Rawski and Heinz 2019, a.o.)

A neural language model computes a probability distribution over words in its vocabulary to predict the next word in a given sequence of words. One type of models used in this dissertation is of Long Short-Term Memory architecture (LSTM; Hochreiter and Schmidhuber (1997)) — a type of a Recurrent Neural Network (RNN) that was developed to capture long-term sequences in training data, making it popular for tasks like language modeling and time-series prediction. LSTM language models have been shown to learn linguistically adequate representations of various types of long-distance dependencies, including subject-verb agreement (Gulordava, Bojanowski, Grave, Linzen, & Baroni, 2018), and, important to our purposes, filler-gap dependencies (Wilcox, Futrell, & Levy, 2023; Wilcox et al., 2019a, 2019b, 2018). Conceptually, the LSTMs architecture is built upon the same fundamental principles as the RNN architecture, which I shortly describe below (cf. Arehalli & Linzen, 2024; Jurafsky & Martin, 2024, for in-depth explanations).

RNNs process sequences of inputs — in the case of language models, vectorized representations of sentences from the training corpus — to produce an output vector by maintaining hidden states across time steps. Hidden states, often denoted by h, can be thought of as RNN's memory that gets updated with every time step i. The hidden state at the time step i is a function of the input word vector at this time step, x_i, and the network's hidden state at the previous time step, h_{i-1}. For a Simple Recurrent Network (Elman, 1990), this equation is shown in 2.1, where σ represents a non-linear activation function (such as tangent or sigmoid), and Ws are weight matrices (W_{hx} is input-to-hidden weight matrix and W_{hh} is hidden-to-hidden weight matrix). Weight matrices are key components of RNNs that transform inputs and hidden states — they determine how much influence each input or hidden state has on the output $\hat{y}$, which is mathematically expressed in 2.2 (with W_{hy} being hidden-to-output weights).

$$h_i = \sigma(W_{hx}x_i + W_{hh}h_{i-1}) \tag{2.1}$$

$$\hat{y}_i = W_{hy}h_i \tag{2.2}$$

Weight matrices are initially set to random values and are adjusted during the training process. During the first training stage (called the forward pass), the input vector and the previous hidden state are multiplied by their respective weight matrices. The results are summed and passed through an activation function (e.g., tanh) to produce the new hidden state as in 2.1. The hidden state is then multiplied by the hidden-to-output weight matrix to produce the output as in 2.2. This stage is followed by loss calculation, where the produced output is compared to the actual target using a loss function (e.g., a predicted word vector is compared to the vector of the word found in that position in the training corpus). This comparison produces a loss value (i.e., an error), which measures how far the predicted output is from the actual target. In the backward pass, the gradients of the loss are calculated with respect to the network's parameters in order to update the weights. The loss is minimized by repeating the forward and backward passes multiple times to define optimal weights — in a process called backpropagation through time (Rumelhart, Hinton, & Williams, 1986; Werbos, 1990). This is schematically shown in Figure 2.2.

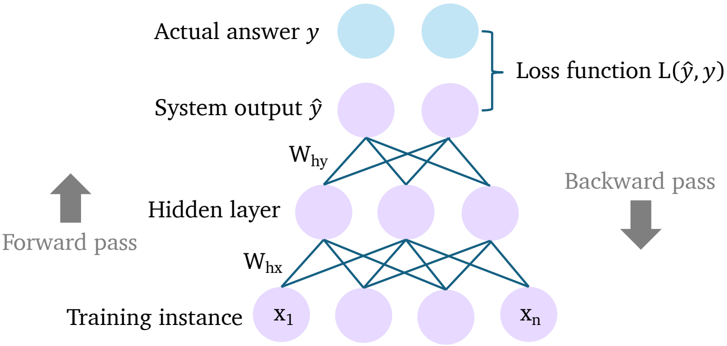


Figure 2.2: Schematic representation of backpropagation through time; circles represent RNN cells. During the forward pass, the estimate $\hat{y}$ is computed, while in the backward pass, the weights W_{hx} and W_{hy} are updated.

In the final stage, the model makes a prediction by applying well-defined weights to some unseen data, which constitutes the output generation. In the case of language models, there is a softmax layer — a final transformation that converts the raw output values (i.e., logits) into a probability distribution over the possible output classes (i.e., the model's vocabulary).

A similar training procedure applies to LSTM RNNs, which involves forward passes, loss calculation, and backpropagation through time. However, during the training of simple RNNs, the repeated multiplication of gradients through many time steps can lead to the so-called vanishing gradient problem, where gradients become exceedingly small and hinder the network's ability to learn long-term dependencies. LSTMs address this issue by incorporating gating mechanisms; these include input, forget, and output gates that regulate the flow of information. These gates allow LSTMs to maintain a more constant error gradient, effectively preserving information over longer sequences. By controlling what information is added to or removed from the cell state, LSTMs mitigate the vanishing gradient problem, making them more suitable for tasks that require processing of longer dependencies.

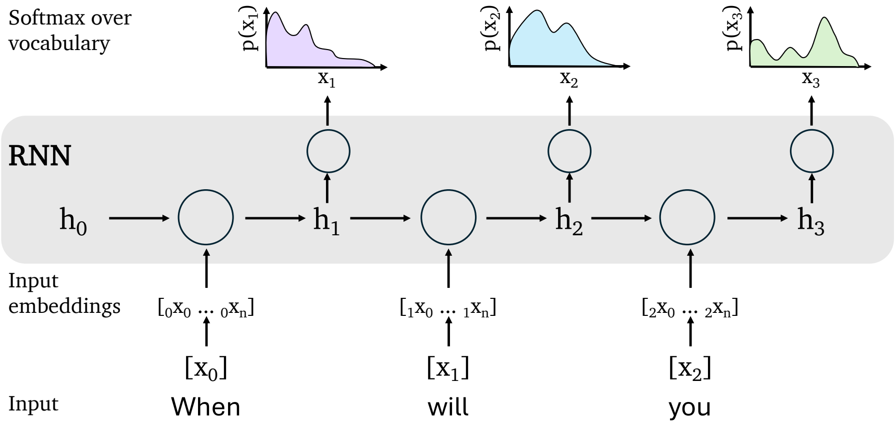


Figure 2.3: Adapted from (Arehalli & Linzen, 2024). In a language modeling task, each word is mapped to an embedding (i.e., a vector representation). The current word vector x_i is merged with the representation of all previous words in the hidden state h_{i-1}. To predict the i-th word, a softmax function is applied to the system output to produce a probability distribution over the i-th word, $p(x_i)$. During training, the model's weights are adjusted to maximize the likelihood of the actual word that appears in the sentence in position i.

Although the LSTM RNN models have achieved impressive results on various NLP and linguistic tasks, the state-of-the-art in the field, and another type of NLMs used in this dissertation, is the *Transformer* (Vaswani et al., 2017). Since their invention a few years ago, Transformers have driven what many call 'a revolution in artificial intelligence' (Emanuilov, 2024) and have become the foundation for *Large Language Models* (LLMs) like BERT (Bidirectional Encoder Representations from Transformers; Devlin, Chang, Lee, & Toutanova, 2019) and GPT (Generative Pre-trained Transformer; Radford, Narasimhan, Salimans, & Sutskever, 2018). Transformers were initially designed for applications like machine translation, and were first introduced as so-called sequence-to-sequence models, or encoder-decoder models. Nowadays, there is a great variety of Transformer model architectures (see Lin, Wang, Liu, & Qiu, 2021, for a review). In what follows, I outline the basic principles behind Transformers focusing on left-to-right (also called autoregressive or decoder-only) models that are trained to predict the next token in a sequence (for an in-depth introduction to Transformer architecture, see Jurafsky & Martin, 2024, Chapters 9-11).

At the heart of Transformer technology is the *self-attention* or *multi-head attention* mechanism. This mechanism can be thought of as a way to build contextual representation of a token by leveraging information from all other tokens in the input sequence (i.e., attending to them), leading to more nuanced and context-aware processing. In essence, self-attention is a weighted sum of context vectors (calculated in a very sophisticated way) that determines the relationship between each element in a sequence and all other elements. This mechanism enables Transformers to capture long-distance dependencies and context more effectively than previous architectures. Unlike RNNs, which process data sequentially, Transformers can handle sequences in parallel — that is, attention computation happens in parallel for each token in a sequence. This parallel processing capability dramatically improves computational efficiency, allowing for faster training and inference times.

Attention is one of the three major components of the architecture, with the other two being feedforward network and layer normalization. Together, they constitute one Transformer block (also sometimes referred to as layer), as illustrated in Figure 2.4. Such blocks are then stacked to create the decoder; block stacking allows for deeper and more powerful networks. As with RNNs, the the architecture begins with input embeddings, where each token is converted into

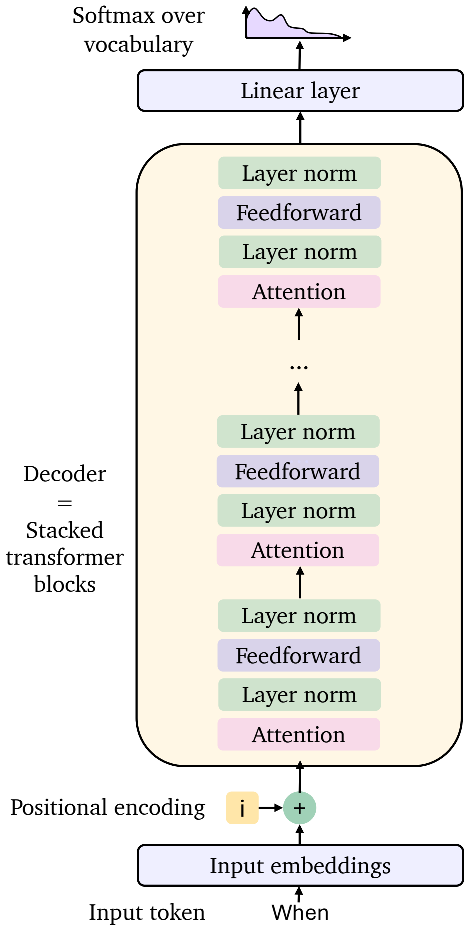


Figure 2.4: Adapted from (Jurafsky & Martin, 2024, Chapter 9). Transformer language model (decoder-only architecture).

a dense vector representation. To retain the order of tokens, positional encodings are added to these embeddings, as the self-attention mechanism itself does not inherently capture positional information. Following the self-attention mechanism, the output is passed through a feedforward neural network. This component consists of two linear transformations with a ReLU activation function in between, introducing non-linearity and enabling the model to learn complex patterns. Each component (self-attention and feedforward network) is followed by layer normalization, which helps stabilize training and facilitate gradient flow through the network.

The decoder consists of multiple identical blocks, each containing a self-attention mechanism and a feedforward neural network. The output of one layer serves as the input to the next layer. At the pre-final step, a linear layer is applied to transform the decoder's output into logits that match the size of the model's vocabulary. At the final step, a probability distribution over the vocabulary for the next token is produced, typically using a softmax function (just like with RNNs). This distribution is used to predict the next token based on the input sequence.

2.3.1 NLMs as models of human language acquisition

With traditional cognitive models, researchers explicitly implement mechanisms that are hypothesized to mimic cognitive processes in the human mind, with those processes operating over symbolic representations. Unlike traditional models, neural network models develop *distributed representations* over the course of their training. This poses significant interpretability challenges when using such models as tools for studying human language acquisition and processing. It is an open question how such distributed representations compare to symbolic representations that many scientists believe are utilized by humans. This emergent behavior also makes it challenging to precisely describe the mechanisms that a neural network employs (however, see studies analyzing behavior of individual neurons with and without causal interventions, as well as ablation studies: Boguraev, Potts, & Mahowald, 2025; Elazar, Ravfogel, Jacovi, & Goldberg, 2021; Finlayson et al., 2021; Lakretz et al., 2019; O'Connor & Andreas, 2021, a.o.). Lastly, and perhaps most importantly, the nature of inductive biases that such models possess is not yet properly understood.

NLMs are usually viewed as weakly-biased learners that have a general preference for maximizing data likelihood. These systems are often described as domain-general (i.e., linguistically-neutral) learners that can learn from any arbitrary type of vectorized input, not limited to language data, and that there is no obvious linguistic bias encoded within the models (Futrell & Mahowald, 2025; Wilcox, Futrell, & Levy, 2023). However, there are other (more general) biases that the models are likely to exhibit. These biases might or might not align well with biases postulated for human learners — today, little is known about this relationship. This is an area of active research, so it is possible that future research will shed more light on this issue.

One of the most notable inductive biases observed in NLMs so far is their preference for information locality (Futrell & Mahowald, 2025). This bias reflects a tendency to favor sequences where elements that predict each other are close together. For example, in human language, syntactic dependencies often involve adjacent or nearby words, such as modifiers being placed close to the nouns that they describe. This property of locality is pervasive in human language, and NLMs appear to mirror it in their processing. The autoregressive nature of NLMs, where predictions are made incrementally based on preceding tokens, naturally aligns with this bias. This similarity suggests that the bias toward locality may be a shared property between NLMs and human language processing, even though the underlying mechanisms differ.

An example of a bias in where NLMs arguably differ from human learners is the relative insensitivity of the former to word order (Futrell & Mahowald, 2025). While human language relies heavily on word order to convey grammaticality and meaning, studies have shown that NLMs, and in particular Transformers, are less sensitive to word order (Gupta, Kvernadze, & Srikumar, 2021; O'Connor & Andreas, 2021).[5] Instead, they tend to focus on lexical content and contextual information, which can often compensate for variations in word order. Interestingly, this mirrors some findings in human language use, where word order is sometimes redundant with other cues, such as case markers or semantic context (Mahowald, Diachek, Gibson, Fedorenko, & Futrell, 2023). This redundancy explains why bag-of-words methods, which ignore word order entirely, were historically successful in many natural language processing tasks (Futrell & Mahowald, 2025).

No matter the (dis)similarity of the biases, their origins in NLMs still differ fundamentally from those in human learners. Human inductive biases are often theorized to stem from innate constraints (be they domain-general or domain-specific) that restrict the forms of possible languages and guide learning. In contrast, NLM biases arise from their architecture, optimization processes, and, perhaps most importantly, training objective (McCoy, Yao, Friedman, Hardy, & Griffiths, 2024). These biases are not hardwired but emerge as a result of the models' exposure to large amounts of linguistic data and their optimization for specific tasks.

The possible points of divergence between human learners and NLMs include not only inductive biases, but also learners' input and learning environment more broadly. The input to human learners and

[5] This is a point of divergence between the two NLM architectures, as RNNs, unlike Transformers, have a built-in notion of linear order.

NLMs learners differs significantly in quantity, modality, and richness (Warstadt & Bowman, 2022). Human learners acquire language through grounded, multimodal environments that include sensorimotor stimuli, prosody, and interaction, while NLMs learners typically rely on text-only input. For example, children are exposed to tens of millions to around 100 million words by puberty, often in spoken or signed form, enriched with prosodic cues and contextual grounding (Gilkerson et al., 2017; Hart & Risley, 1992). In contrast, popular NLM models like GPT-3 are trained on hundreds of *billions* of words, far exceeding human exposure (Frank, 2023a). Additionally, human learners benefit from interactive learning environments that incentivize communicative success, whereas NLMs learners optimize for statistical reproduction of text distributions (McCoy et al., 2024). Efforts to bridge this gap, such as multimodal models trained on text-image pairs or transcribed speech datasets like CHILDES, have yet to replicate the complexity of human input (MacWhinney, 2000).

Due to these differences in learning conditions between humans and NLMs, there is ongoing debate about the utility of NLMs for studying the developmental trajectories of human language acquisition. One extreme position holds that NLMs themselves can serve as theories of language acquisition (Piantadosi, 2023) — a claim that has received substantial critique (Cuskley, Woods, & Flaherty, 2024; Katzir, 2023; Kodner, Payne, & Heinz, 2023). At the other end of the spectrum, some scholars argue that NLMs offer no meaningful insight into human linguistic cognition (Bolhuis, Crain, Fong, & Moro, 2024; Chomsky, Roberts, & Watumull, 2023; Moro, Greco, & Cappa, 2023).

Between these poles, there is a range of middle-ground approaches. These acknowledge the differences between NLMs and human learners but suggest that NLMs can still be informative for studying linguistic learnability. However, the conclusions drawn from such work — and the degree of optimism — vary considerably. Some researchers argue that the ability of NLMs to acquire complex linguistic patterns challenges traditional assumptions about the necessity of strong domain-specific inductive biases. According to this view, NLMs provide a proof of concept for gradient, usage-based theories of language, which posit that linguistic structure emerges from statistical regularities in language use (Futrell & Mahowald, 2025; Wilcox, Futrell, & Levy, 2023).

A growing body of work supports the use of NLMs in acquisition research, provided that their inherent advantages over human learners — such as vast training data — are carefully controlled (Frank, 2023b; Warstadt & Bowman, 2022; Warstadt et al., 2023). This line of research explores how NLMs can be used to test specific acquisition hypotheses or to simulate learning under different conditions, by, for example, manipulating their training data (Leong & Linzen, 2024; Patil et al., 2024). Other scholars adopt a more cautious stance, recognizing the potential of NLMs to inform questions of learnability, but emphasizing that their current utility remains limited (Lan, Chemla, & Katzir, 2024; Vázquez Martínez, Heuser, Yang, & Kodner, 2024; Ziv, Lan, Chemla, & Katzir, 2025). These studies often investigate whether NLMs truly acquire linguistic generalizations or rely on superficial heuristics and shortcuts (McCoy, Pavlick, & Linzen, 2019), and stress the importance of scrutinizing both NLM training data and evaluation methods (Vázquez Martínez et al., 2024). In this dissertation, I adopt a cautious middle-ground perspective. I use NLMs to explore what kinds of generalizations are supported by the input given weak domain-general biases, without making any commitments about how human-like those learned generalizations are.

2.3.2 Evaluation of NLMs

As discussed above, assessing the internal components of a neural network model is a challenging task. To assess NLM's linguistic abilities, it is often more insightful to probe the representations that the model learns within the pressures that it is given (e.g., the model's inductive biases, its training data and task; Arehalli & Linzen, 2024). When it comes to linguistic representations, it is common to probe whether a given model can learn formal properties of natural languages like hierarchical structure, grammatical constraints, and recursive rules that allow embedding phrases within phrases. Phenomena like question formation (Ahuja et al., 2024; McCoy, Frank, & Linzen, 2020; Yedetore, Linzen, Frank, & McCoy, 2023), subject-verb agreement (Arehalli & Linzen, 2024; Gulordava et al., 2018; Linzen, Dupoux, & Goldberg, 2016), passivization (Leong & Linzen, 2024; Mueller & Linzen, 2023), center embedding (Hu, Gauthier, Qian, Wilcox, & Levy, 2020; Wilcox et al., 2019a), and filler-gap dependencies (Chowdhury & Zamparelli, 2018; Lan et al., 2024; Wilcox, Futrell, & Levy, 2023; Wilcox et al., 2018) serve as testing cases.

The engineering community tests these models on standard, straightforward tests that use sentences that are likely to be present in the models' training corpus (such process is often referred to as 'benchmarking'). This is a potential shortcoming as we cannot say whether the models' performance on such tests reflects their mastery of abstract rules that govern natural languages or just rote memorization of the training examples (McCoy, Min, & Linzen, 2020). Therefore, researchers often employ methods that resemble psycholinguistic experiments with human participants that test a specific linguistic phenomenon, also refereed to as 'targeted evaluation' (Futrell et al., 2019). Such targeted evaluation methods include testing the models on minimal pairs of sentences of varying grammaticality, cloze tests evaluating grammatical preferences of the model by assessing model-generated continuations of a sentence given a prompt, various probing tasks, and evaluating how the models fare as incremental processors (cf. Warstadt et al., 2019). The probability that the model assigns to each sentence in a minimal pair is taken to indicate whether the model has learned the appropriate linguistic representations. For instance, if the model assigns a higher overall probability to a sentence (2-a) as compared to (2-b), this is taken as evidence that the model has learned that verbs in English must agree with their subjects in number, and that *the key to the cabinets* is a subject phrase with a singular head. There are benchmarks of such minimal pairs for English called BliMP (Warstadt et al., 2020) as well as other test suites assessing syntactic knowledge of NLMs (Hu et al., 2020).

(2) a. The key to the cabinets is on the table.
b. *The key to the cabinets are on the table.

While whole sentence probability can be sufficient for comparing minimal pairs of sentences for some linguistic phenomena, other phenomena might require more nuanced comparisons and measures. In such cases, it is common to explore how NLMs fare as incremental processors by looking at *surprisal* — an information-theoretic measure of predictability of a word in context. Surprisal is calculated according to the formula in 2.3, where x_i is the current word, h_{i-1} is the network's hidden state before consuming x_i, and p is the conditional probability calculated from the network's softmax layer.

$$S(x_i) = -log_2 p(x_i|h_{i-1}) \tag{2.3}$$

In psycholinguistic research, surprisal is more broadly defined as the negative logarithm of the probability of a word given its preceding context. This measure is central to *surprisal theory*, which is one of the most popular modern approaches to language processing. This theory, originally proposed by Hale (2001) and further developed by Levy (2008), posits that predictability of a word in a sentence, quantified as surprisal, directly influences the cognitive effort required to process it. Less predictable words, which have higher surprisal values, are associated with greater processing difficulty (Hale, 2001; Levy, 2008). This relationship has been supported by various studies showing that surprisal — including surprisal values derived from language models — is a strong predictor of both behavioral measures, such as reading times or gaze duration, and neural responses, such as the amplitude of event-related brain potentials (Michaelov, Bardolph, Van Petten, Bergen, & Coulson, 2024; Shain, Meister, Pimentel, Cotterell, & Levy, 2024; Smith & Levy, 2013; Wilcox, Meister, Cotterell, & Pimentel, 2023).

Chapter 3

Modeling FGDs and islands: Approaches and methods

The two modeling approaches described in the previous chapter have been applied to a variety of learning tasks, including the acquisition of filler-gap dependencies and island constraints on them in English. In this chapter, I focus on two lines of research that are particularly relevant to the experiments presented in the articles constituting this thesis: the symbolic modeling work of Pearl and Sprouse (Pearl & Sprouse, 2013a, 2013b), and the neural language modeling studies on FGDs learnability, primarily those conducted by Wilcox and colleagues (Wilcox, Futrell, & Levy, 2023; Wilcox et al., 2019a, 2019b, 2018). These sets of studies are described in Sections 3.1 and 3.2 respectively.

In this dissertation, I explore whether the two proposed learning models represent viable theories for learning about island facts cross-linguistically by applying them to Norwegian data. Before describing these modeling studies in detail, I outline a set of main research questions that this thesis aims to address. As discussed in the previous chapter, the language acquisition process involves both observable and internal components. The observable components include the input and the resulting linguistic behavior, with the latter serving as the modeling target. Regardless of model type, the goal is to reproduce the target linguistic behavior based on the input that the model receives. The research questions then naturally target these observable components and the models' inductive biases that allow them to generalize from — and beyond — the input:

1. Can the models induce the (non-)island status of different domains in Norwegian?

2. What kind of evidence does the input provide, and how does it influence model generalizations?

3. Can the models generalize beyond the input?

In this chapter, I explain how these research questions are adapted to two modeling approaches described below. I also discuss the methodological challenges and decisions involved in applying these models to Norwegian data. Finally, I summarize all research questions — including the model-specific ones — that this dissertation aims to answer (Section 3.3).

3.1 The modeled learner of Pearl and Sprouse

The modeled learner of Pearl and Sprouse (2022; 2013a; 2013b) is an example of a symbolic top-down cognitive model proposing how children might acquire the distribution of acceptable filler-gap dependencies and island constraints on them from transcribed and syntactically annotated child-directed speech. The researchers proposed a learning mechanism and tested it on a set of *wh*-filler-gap dependencies in English. According to this model, a learner must pay attention to the phrasal categories that contain FGDs called *container nodes.* Container nodes are major phrases like NP, VP, IP, CP, where CP nodes are annotated with lexical information about the complementizer head in order to distinguish between different types of embedded clauses (e.g., declarative vs. interrogative). An example of a container node sequence for a sentence *What did he think she saw?* is in (1-b). To accommodate shorter sequences, and to track filler-gap dependency boundaries, START and END nodes are appended to each container node sequence.

The learning mechanism involves breaking down sequences of container nodes into smaller units called container node *trigrams*, as shown in (1-c). During the learning period, the modeled learner tracks the frequency of such trigrams to induce a frequency distribution over previously encountered trigrams. The learner then uses this distribution to infer probabilities of novel sequences, calculated as the product of probabilities of the container node trigrams, which is shown in (1-d).

(1) a. [$_{\text{CP}}$ What did [$_{\text{IP}}$ he [$_{\text{VP}}$ think [$_{\text{CP}_{null}}$ [$_{\text{IP}}$ she [$_{\text{VP}}$ saw ___]]]]]]?

b. Sequence: START-IP-VP-VP$_{null}$-IP-VP-END

c. Trigrams:

START-IP-VP

IP-VP-CP$_{null}$

VP-CP$_{null}$-IP

CP$_{null}$-IP-VP

IP-VP-END

d. P(START-IP-VP-VP$_{null}$-IP-VP-END) = P(START-IP-VP) * P(IP-VP-CP$_{null}$) * ... * P(IP-VP-END)

By treating the probability of a given dependency as the product of trigram probabilities, the modeled learner can generalize beyond the specific set of FGDs encountered during training. This approach allows the model to assign a non-zero probability to any FGD whose container node sequence consists entirely of previously observed trigrams. Therefore, grammatical dependencies composed of known trigrams will be identified as possible FGDs. Conversely, any sequence containing one or more unknown trigrams should have a probability of zero, as at least one of the trigrams being multiplied has a zero probability. In practice, Pearl & Sprouse use a smoothing technique that assigns a count of 0.5 to unknown trigrams, resulting in very low, but non-zero, probabilities for unattested dependencies. If island violating structures are unattested in the input, then the overall probability of an island structure will be close to 0. For instance, in example (2) below — which is a dependency into a *whether*-island — the container node CP$_{whether}$ should be unattested if the learner has never seen examples of extraction out of *whether*-clauses.

(2) [$_{CP}$ *What did [$_{IP}$ he [$_{VP}$ wonder [$_{CP_{whether}}$ whether [$_{IP}$ she [$_{VP}$ saw ___]]]]]]?

Therefore, the learner employs probabilistic reasoning: a dependency cannot cross a very low probability region of the syntactic structure. The idea behind this reasoning is somewhat close to Chomsky's Subjacency condition except that the dependency is given a gradient probability rating rather than a binary grammaticality status.

3.1.1 Model input and target state

As the input to the model, Pearl and Sprouse used transcribed child-adult interactions from the CHILDES database (MacWhinney, 2000). Their sample consisted of of 31,247 *wh*-dependencies extracted from several corpora that contained child-directed speech to 25 children aged between 1.5 and 5 years old. The utterances were automatically parsed using Penn Treebank style annotation and subsequently verified by human annotators, who also inserted appropriate gaps into the tree structures. When compared against the language acquisition model presented in Figure 2.1 in the previous chapter, such model input is best described as the acquisitional intake.

To define the modeling target state, Pearl and Sprouse used experimental acceptability judgment task (AJT) data from native English speakers. Ideally, such behavioral data should come from children of the same age as the target audience of the model input data. However, not many behavioral studies with children have been conducted to study the developmental trajectories for island acquisition (though see De Villiers & Roeper, 1995; De Villiers et al., 2008). Consequently, as an approximation of the final target state, Pearl and Sprouse used AJT data from adult speakers from Sprouse, Wagers, and Phillips (2012). This way, the probabilities induced by the learner were compared to acceptability ratings, based on the assumption that grammaticality (as mediated through syntactic probability) is a major factor contributing to acceptability (albeit not the only one).

The AJTs in Sprouse et al. followed an established 2×2 design for quantifying island effects (Sprouse, 2007). This design manipulates the DISTANCE between the filler and the gap (*Short vs. Long*) and the presence of an island STRUCTURE (*No Island vs. Island*). An example experimental item set for testing for island effects in English is in (3), where DISTANCE modulates whether the *wh*-filler *who* is associated with a gap in the matrix clause (*Short*) or the embedded clause (*Long*). STRUCTURE manipulates whether the embedded clause is a declarative complement clause (*No Island*) or an embedded *whether*-question (*Island*). The FGD in the *Long, Island* condition corresponds to an 'island violation' in English, as the filler is associated with a gap located inside an embedded question.

(3) a. *Short, No Island*
Who$_i$ ___$_i$ thinks that Olav stole the necklace?
b. *Short, Island*
Who$_i$ ___$_i$ wonders whether Olav stole the necklace?
c. *Long, No Island*
What$_i$ does the detective think that Olav stole ___$_i$?
d. *Long, Island*
What$_i$ does the detective wonder whether Olav stole ___$_i$?

According to this design, the island effect is defined as super-additive interaction between the *Long* level of DISTANCE and the *Island* level of STRUCTURE. The island effect manifests when (i) the *Long, Island* condition receives significantly lower ratings compared to all other conditions, and (ii) the decreased acceptability of this condition cannot be solely explained by the simple effects of dependency length and structural complexity. The residual drop in unacceptability is defined as the island effect. The presence or absence of this effect can be visually examined using interaction plots, as shown in Figure

3.1. If dependency length and structural complexity are sufficient to account for the relative unacceptability of the *Long, Island* condition, there will be no island effect, and the plotted ratings for the four conditions will form two parallel lines. Conversely, if an island effect is present, the plot will display two non-parallel lines, with the *Long, Island* condition having the lowest ratings.

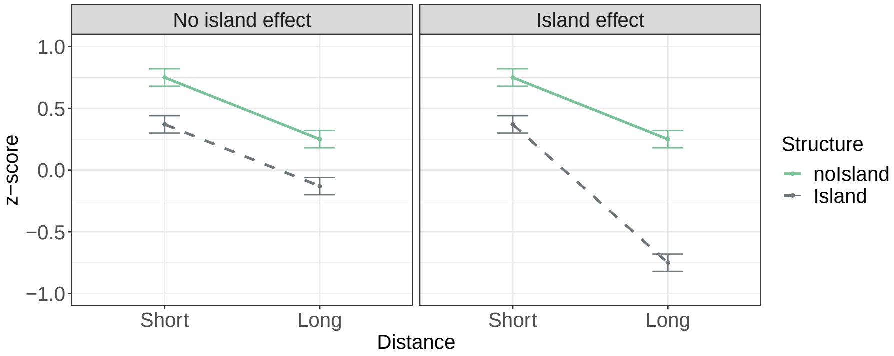


Figure 3.1: Visual definition of island effects: The left panel represents the absence of an effect, and the right panel represents a pattern when the island effect is present.

Pearl and Sprouse assessed their modeled learner by comparing its induced probabilities against adult behavioral data on four island types. They plotted the induced probabilities and z-scores representing acceptability ratings side-by-side, and found good qualitative alignment between them.[1] Pearl and Sprouse concluded that their modeled learner showed an overall feasibility of a distributional learning strategy applied to a naturalistic and developmentally plausible corpus.

3.1.2 Model biases

Although the learning model described above substantially reduces the amount of prior domain-specific knowledge compared to traditional generative proposals such as Subjacency, it still incorporates several complex biases, many of which are specific to language. First, the learner proposed by Pearl and Sprouse (2013b) relies on access to hierarchical syntactic structure. Second, the model includes a bias for

[1] Despite their training corpus being two times smaller than what they estimate to be an amount of data needed for learning about islands (Pearl & Sprouse, 2013a, 2013b; Phillips, 2013a).

attending to a specific subset of the input: namely, *wh*-dependencies and their associated container node trigrams. Another domain-specific bias involves encoding embedded clauses with lexical information about the complementizer head. These features reflect assumptions about the structure and granularity of the input that the modeled learner is designed to track.

The learner also incorporates biases in its inference mechanism. These include the ability to chunk longer sequences into smaller units (in this case, trigrams), to track the frequency of different elements in the input, and to use those frequencies to compute probabilities. These are likely to be general-purpose learning biases that are not specific to language.

The proposed learning strategy raises difficult questions about how such biases might arise in the learner. The reliance on hierarchical structure and attention to specific syntactic configurations reflects domain-specific knowledge, which, as the authors rightfully note, is required for all other syntax-related linguistic phenomena (e.g., agreement, question formation, passivization, etc.). To date, there is no agreement as to whether this domain-specific knowledge is innate or derived from experience (for a detailed discussion, see Pearl & Sprouse, 2013b, p. 47). The authors argue that the two additional domain-specific biases — attending to a particular subpart of the input and categorizing embedded clauses — are likely to be derived from knowledge of syntactic structure, and therefore acquired through exposure to linguistic input. In contrast, the biases related to chunking, frequency tracking, and probabilistic inference are more plausibly grounded in innate domain-general cognitive abilities.

3.1.3 Applying the learner to Norwegian data

I test the general feasibility of the proposed learning strategy by applying it to Norwegian data. In doing so, I aim to address the three main empirical research questions outlined above through the following steps. First, I apply the symbolic model to Norwegian data to assess whether it constitutes a viable strategy for learning about island constraints cross-linguistically. Specifically, I investigate whether the learner can acquire a different set of island generalizations in Norwegian compared to English. Second, I examine the evidence available in the input to Norwegian learners to evaluate how it influences the model's generalizations. Third, I extend the range of tested filler-gap dependencies to include RC-dependencies, in order to assess whether the learner can acquire target generalizations for a broader set of constructions — also in light of potential cross-construction variation. This also allows me to evaluate the model's capacity to generalize beyond the input.

Two additional research questions have to do with how much the success of this learning strategy depends on the idiosyncratic assumptions of the Pearl and Sprouse's modeled learner. First, is the learner's success formalism-dependent, or are there other representational formats that would also support distributional learning of island facts? By answering this question, I address one previous point of criticism related to this acquisition strategy that "it shuttles the debate from island constraints to one particular parse schema, and whether or not it can be learned using only domain-general assumptions." (Wilcox, Futrell, & Levy, 2023, p. 7). To this end, I explore whether the modeled learner can succeed with an input annotated within Lexical-Functional Grammar (LFG) formalism. LFG provides hierarchical syntactic representations that encode functional relations (e.g., subject, object) and clause types more compactly than the phrase structure representations used by Pearl and Sprouse, resulting in shorter dependency paths.

The difference between dependency lengths in the two formalism allows me to answer another additional methodological question: Can the learner succeed with a smaller n-gram window size? A smaller n-gram size (e.g., bigrams) focuses on highly local properties of structural paths, making it easier to generalize beyond the input corpus and reducing the risk of data sparseness. Moreover, it reduces the total number of distinctive n-grams that the learners needs to track, as this number is calculated as the number of categories of container nodes to the power of n-gram window size (Phillips, 2013a). On the other hand, for the learner to succeed with a smaller n-gram window size, all island violations should be attributable to local illicit bigrams of structure. As such, this methodological choice might result in the learner missing some non-local dependencies.

There are other methodological modifications that I made in order to apply the idea behind the learner of Pearl and Sprouse to Norwegian data. First, as Norwegian unfortunately lacks large-enough corpora of child-directed speech, I used annotated children's fiction corpus from NorGramBank (Dyvik et al., 2016) as the approximation of model input. This difference in input corpus modality has a number consequences. On the one hand, written input is likely to contain greater syntactic complexity compared to child-directed speech, but also fewer questions and embedded questions that are prominent in child-adult interactions. Therefore, this corpus may overrepresent RC-dependencies and underrepresent *wh*-dependencies compared to child-directed speech. On the other hand, the corpus used in this study is approximately five times larger than the original corpus of Pearl and Sprouse, which might potentially yield better frequency estimates for structures in question. Additionally, because there is no variance in the probability values induced by the learner for any condition, there is no formal statistical way to (i) check for statistically significant interaction in the induced probabilities (the same way as an interaction can identified in behavioral data); (ii) formally compare the modeling results and human judgments. Consequently, alignment between modeling and behavioral results is assessed qualitatively by inspecting the interaction plots. To address this issue somewhat, I used bootstrapping techniques to construct confidence intervals for the probabilities assigned to dependency paths, providing a measure of uncertainty around the modeling results.[2]

3.2 Neural language models and island constraints

More recently, researchers have started using neural language models to explore learnability questions pertaining to filler-gap dependencies and islands. This line of work aims to offer a new perspective on the necessity of domain-specific biases for island acquisition. Unlike the symbolic model of Pearl and Sprouse, which relies on structured representations and targeted attention to specific syntactic configurations, NLMs are trained on raw text and operate with relatively weak, general inductive biases. If NLMs are able to approximate human-like generalizations about islands — despite lacking explicit syntactic knowledge — this could suggest that some aspects of island sensitivity may emerge from statistical learning over large-scale input, rather than requiring innate, language-specific constraints.

Chowdhury and Zamparelli (2018; 2019) used two types of RNNs to investigate how well the probabilities assigned by the models align with human acceptability judgment ratings. They calculated whole-sentence probabilities for sentences with subject and relative clause

[2] Pearl and Bates also used sampling techniques to create representative learning samples that matched estimated input quantity for higher- and lower-socioeconomic status children, and to subsequently analyze whether there is a difference between the modeled judgments from these two groups.

island violations, as well as affirmative and yes/no-interrogative control sentences, as shown in (4) (alongside other comparisons).

(4) a. *Who$_i$ did John see the person that dated ___$_i$? [WH-INTER]
b. Did John see the person that dated Mary? [Y/N-INTER]
c. John saw the person that dated Mary. [AFFIRMATIVE]

Their results showed that the models could exhibit some sensitivity to islands as they assigned less whole-sentence probability to sentences that violated island constraints compared to affirmative control sentences, (4-a) vs. (4-c). However, they also found that the models assigned a lower probability to grammatical yes/no-questions when compared to ungrammatical sentences with island violations, (4-b) vs. (4-a). They concluded that model probabilities were mostly affected by the cumulative effect of increasing syntactic complexity, and that the models are "unable to induce a more abstract notion of grammaticality" (Chowdhury & Zamparelli, 2018, p.133).

In another set of studies, which is more relevant for the current dissertation, Wilcox and colleagues (2023; 2019a; 2019b; 2018), investigated whether different types of neural language models can learn the distribution of acceptable *wh*-FGDs and island constraints on them in English from unannotated text input. Wilcox and colleagues used the models' surprisal as their dependent variable. Compared to investigations by Chowdhury and Zamparelli, Wilcox and colleagues analyzed *word-by-word* surprisal values that the models assigned to test sentences, rather than whole sentence probabilities. This allowed them to simulate how the models would fare as incremental language processors and pinpoint potential sources of ungrammaticality or unpredictability in a given sentence, much like psycholinguistic investigations with humans that use self-paced reading or eye-tracking methods. They argued that equating whole-sentence probabilities to the models' notion of grammaticality could have masked some important generalizations about FGDs that the models have learned (for example, expecting gaps in certain syntactic positions more than in others).

3.2.1 Diagnosing NLMs' sensitivity to FGDs

Wilcox and colleagues' experimental stimuli were created according to a 2×2 design manipulating the presence of a filler and the presence of a gap, as shown in (5). This design relies on the fact that the relationship between fillers and gaps is *bidirectional*: fillers require gaps to be interpreted, and gaps require fillers to be properly licensed. Wilcox and colleagues tested the models' sensitivity to FGDs by comparing two minimal sentence pairs, by measuring the difference between +FILLER, −FILLER conditions in each.

(5) a. −FILLER, −GAP:
I know that the lion devoured the gazelle at sunrise.
b. +FILLER, −GAP:
*I know what the lion devoured the gazelle at sunrise.
c. +FILLER, +GAP:
I know what the lion devoured ___ at sunrise.
d. −FILLER, +GAP:
*I know that the lion devoured ___ at sunrise.

In −GAP conditions, they tested whether the models exhibit an expectation for a gap after encountering a filler by measuring the surprisal values that the models assign to a filled noun phrase at a potential gap site (at *the gazelle* in (5)). If the models are sensitive to encountering a filled NP where a gap is expected, the difference in surprisal between (5-b) and (5-a) should be positive. Such effects are known as *filled-gap effects* in psycholinguistic literature (Stowe, 1986). In +GAP conditions, they measured how presence of a filler influences the processing of a later gap by comparing the surprisal values that the models assign to the region following the gap (at *at sunrise* in (5)). In our experiments, we called this comparison *unlicensed gap effect* (article C1, Kobzeva, Arehalli, Linzen, & Kush, 2022). If the models are sensitive to the fact that gaps should be licensed, encountering a gap in (5-c) should be less surprising than in (5-d), and the surprisal difference between them should be negative.

Wilcox et al. compared minimal pairs of sentences with and without island violations. They found that the models' sensitivity to the bidirectional relationship between fillers and gaps is attenuated within islands, suggesting that the models learned the relevant constraints in English. They argued that their results present empirical evidence against the argument from the poverty of the stimulus for the learning of islands insofar the models that they employed do not possess any domain-specific biases.

3.2.2 Applying NLMs to Norwegian data

Although Wilcox et al.'s results are intriguing, suggesting that NLMs can approximate human-like sensitivity to FGDs and island constraints, more recent studies complicate this picture. Empirical evidence increasingly shows that NLMs struggle when tested on more complex syntactic environments and do not, in fact, capture the underlying linguistic generalizations that govern FGDs (Bhattacharya & van Schijndel, 2020; Chaves, 2020; Da Costa & Chaves, 2020; Howitt, Nair, Dods, & Hopkins, 2024; Lan et al., 2024). These studies question the robustness of earlier findings and challenge the view that FGD acquisition can be supported by domain-general biases only.

First, NLM performance has been shown to vary depending on the type of FGD being tested (Howitt et al., 2024; Ozaki, Yurovsky, & Levin, 2022). For example, NLMs were shown to perform best on *wh*-dependencies, while constructions with tough-movement (e.g., *'He is tough to play against ___'*) are learned to a lesser extent (Howitt et al., 2024; Ozaki et al., 2022). Howitt et al. conducted an augmented corpus training analysis and found that that the frequency of particular constructions modulates the model's recognition of the filler-gap dependency types, in line with Ozaki et al. (2022). They concluded that NLMs do not learn a unified, abstract representation behind filler-gap dependencies but rather a series of piecemeal generalizations and shallow heuristics.

Second, the limited cross-linguistic work available indicates that NLM success may not generalize across languages. For instance, Suijkerbuijk, de Swart, and Frank (2023) report that an LSTM model trained on Dutch does not induce the island status of embedded questions in the language, contrary to the results of a behavioral AJT experiment. This result raises questions about whether NLMs can successfully learn island facts across languages. This is particularly relevant for the current thesis, which focuses on Norwegian — a language with its own unique distribution of acceptable and unacceptable FGDs.

Together, these findings invite follow-up questions about the ability of NLMs to induce complex generalizations that underlie FGDs and islands across constructions and languages. To this end, we conduct modeling experiments with NLMs to answer the previously specified main empirical questions. First, we explore whether NLMs can learn patterns of cross-linguistic variation in island facts. Second, we conduct a restricted corpus analysis of NLM training data to explore whether the input to such models contains evidence for typologically marked structures. Third, as with the learner of Pearl and Sprouse, we extend the set of tested FGDs by including RC-dependencies to see if NLMs can learn a broader set of FGDs.

Before testing NLMs on complex environments like islands, we explore whether they can learn basic properties of FGDs such as flexibility of filler-gap licensing and unboundedness. This is an important prerequisite for conducting island experiments. In order to interpret a model's behavior in island contexts, we must first establish that it has learned the baseline generalizations about FGDs. For example, if a model fails to show sensitivity to doubly-embedded FGDs, then any apparent island effect may simply reflect a broader failure to acquire doubly-embedded FGDs, rather than a specific sensitivity to island constraints. Conversely, if the model demonstrates robust generalization across embedded clauses and across different syntactic positions, we can more confidently attribute any drop in performance in island contexts to the presence of learned constraints. Thus, testing for basic FGD properties serves as a necessary diagnostic for evaluating whether the model has acquired the relevant grammatical representations that make island effects meaningful.

Additionally, we test whether different model architectures (LSTM RNN/Transformer) achieve similar results, and conduct a controlled cross-linguistic comparison of island learnability in English and Norwegian. The motivation for comparing architectures is two-fold. First, many previous studies on filler-gap learnability were conducted specifically with LSTM RNNs (Chaves, 2020; Howitt et al., 2024; Ozaki et al., 2022; Suijkerbuijk et al., 2023; Wilcox et al., 2018), but the state-of-the-art architecture for all natural language processing tasks is the Transformer. In order to facilitate the comparison with previous studies, and to keep up with the current standard, both model types are included into the analysis. Second, this comparison is also somewhat theoretically motivated: if both architectures, despite their structural differences, converge on similar generalizations about filler-gap dependencies and island constraints, this strengthens the argument that such patterns are learnable from statistical input alone. Conversely, if only one architecture succeeds while the other fails, this has a potential to shed more light on the types of architectural biases (i.e. recurrence vs. attention) that are necessary for learning these phenomena.

Similarly, the cross-linguistic comparison between English and Norwegian allows us to assess whether the same learning mechanisms can accommodate variation in island sensitivity across languages in an experimentally controlled setting, when model architectures and test items are closely matched. If a model can learn different island patterns depending on the input language, and there are no other factors that can contribute to this results (e.g., variation in experimental items) this would provide a stronger proof of concept that domain-general statistical learners can, in principle, acquire language-specific constraints — an essential requirement for any theory of cross-linguistic acquisition.

Finally, it remains unclear how well NLMs predict human linguistic behavior in the processing of filler-gap dependencies. In particular, the experimental design described above relies on the assumption that FGD-formation is an active and incremental process, and that active gap-filling is suspended within island domains (Phillips, 2006; Stowe, 1986; Traxler & Pickering, 1996). However, there is limited behavioral research investigating active gap-filling in light of potential cross-linguistic variation in island effects. To address this gap, we examine whether Norwegian speakers actively predict gaps within embedded questions — a non-island domain in the language — and, if so, how well NLMs can model this linguistic behavior.

3.3 Overview of research questions

To summarize, the studies conducted by Pearl and Sprouse and Wilcox and colleagues found that two types of differently biased statistical models can learn the distribution of acceptable *wh*-filler-gap dependencies in English and associated island constraints on them (at least, to some extent). However, it is unclear yet whether such models can serve as models for learning FGDs and islands given the possibility of cross-linguistic variation. Moreover, given the general unacceptability of island-violating structures in English, the models are unlikely to encounter such examples during training. The patterns of variation found in Norwegian make it a great testing ground for exploring how the evidence found in the input influences model generalizations. Additionally, it is unclear whether the results obtained will hold for a wider range of constructions with filler-gap dependencies, such as relative clauses. In this dissertation, I address these three empirical research questions, alongside some model-specific questions discussed above, which are summarized in Figure 3.2. In the next section, I give an overview of five studies that were conducted to answer these questions.

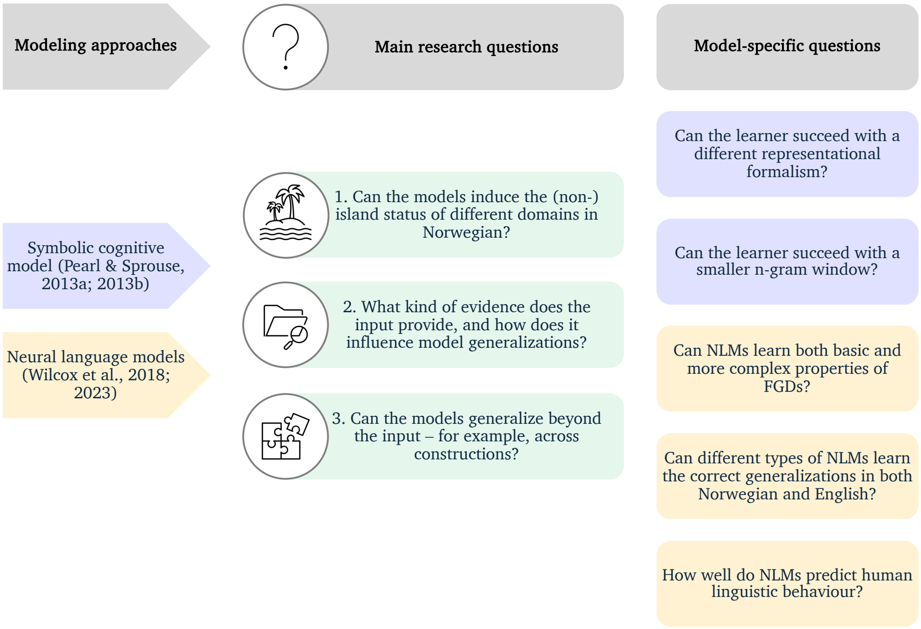

Modeling approaches
Main research questions
Model-specific questions
Can the learner succeed with a different representational formalism?
1. Can the models induce the (non-) island status of different domains in Norwegian?
Can the learner succeed with a smaller n-gram window?
Symbolic cognitive model (Pearl & Sprouse, 2013a; 2013b)
2. What kind of evidence does the input provide, and how does it influence model generalizations?
Neural language models (Wilcox et al., 2018; 2023)
Can NLMs learn both basic and more complex properties of FGDs?
3. Can the models generalize beyond the input – for example, across constructions?
Can different types of NLMs learn the correct generalizations in both Norwegian and English?
How well do NLMs predict human linguistic behaviour?

Figure 3.2: Modeling approaches used in this thesis, together with main research questions that are common for both approaches (in green), and model-specific questions (in purple or yellow).

Chapter 4

Research dissemination

This thesis summarizes most of the research conducted during the PhD period. It comprises five articles that were published in two types of scientific outlets: as conference proceedings articles (C1-C2) and as journal articles (J1-J3). These five articles are listed below.

4.1 List of articles

C1 **Kobzeva, A.,** Arehalli, S., Linzen, T. & Kush, D. (2022). LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian. In J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 2974–2980). URL: https://escholarship.org/uc/item/012683gb

C2 **Kobzeva, A.,** Arehalli, S., Linzen, T., & Kush, D. (2023). Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian. In T. Hunter & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics* (Vol. 6, pp. 175–185). DOI: 10.7275/qb8z-qc91

J1 **Kobzeva, A.** & Kush, D. (2024). Grammar and Expectation in Active Dependency Resolution: Experimental and Modeling Evidence from Norwegian. *Cognitive Science, 48*(10), e1350. DOI: 10.1111/cogs.13501

J2 **Kobzeva, A.** & Kush, D. (2025). Acquiring Constraints on Filler-Gap Dependencies from Structural Collocations: Assessing a Computational Learning Model of Island-insensitivity in Norwegian. *Language Acquisition.* DOI: 10.1080/10489223.2024.2440340

J3 **Kobzeva, A.,** Arehalli, S., Linzen, T. & Kush, D. (2025). Learning Filler-Gap Dependencies with Neural Language Models:

Testing Island Sensitivity in Norwegian and English. *Accepted to the Journal of Memory and Language.*

Apart from the main articles listed above, one additional paper was published during the PhD period together with the members of the ØyeLab (EyeLands Lab). Due to its collaborative nature, this paper is not part of the current thesis[1]:

J4 **Kobzeva, A.,** Sant, C., Robbins, P. T., Vos, M., Lohndal, T., & Kush, D. (2022). Comparing Island Effects for Different Dependency Types in Norwegian. *Languages, 7*(3), 195–220. DOI: 10.3390/languages7030197.

4.2 Notes on author contributions

4.2.1 Note on articles J1 and J2

Articles J1 and J2 are joint works with Assistant Professor Dave Kush, who supervised this thesis. For article J1, I was responsible for creating the initial version of the experimental items, administering the technical experimental setup, recruiting participants, collecting data, performing statistical analysis, and visualizing the results. I also gathered, analyzed, and visualized data for the modeling experiment. Both authors contributed to conceptualizing the experimental idea and design, finalizing the experimental items, and writing the manuscript (including drafting, reviewing, and editing).

For article J2, I created the script for extracting filler-gap dependency paths from annotated data, implemented the modeled learner in Python, analyzed the modeling results, and wrote the first draft of the manuscript. Both authors contributed to conceptualizing the modeling experiment and manually checked parts of the automatically extracted data. Both authors also contributed to the writing process, including editing and revision.

4.2.2 Note on articles C1, C2, and J3

Articles C1, C2, and J3 were co-authored with Assistant Professor Suhas Arehalli, Associate Professor Tal Linzen (co-supervisor), and Assistant Professor Dave Kush (main supervisor). I was primarily responsible for creating the experimental stimuli, gathering modeling

[1] However, the behavioral data collected in this study served as the modeling target state for the symbolic model presented in article J2.

data, conducting statistical analyses, and visualizing the results. DK supervised this process. SA trained the Transformer model reported in J3 for English, while I trained all Norwegian models under his guidance. In terms of manuscript writing, DK and I drafted the initial drafts, with TL and SA providing feedback and comments. Both DK and I were involved in editing and revision of the papers, while TL and SA provided further comments and feedback on the revised versions.

4.3 Article summaries

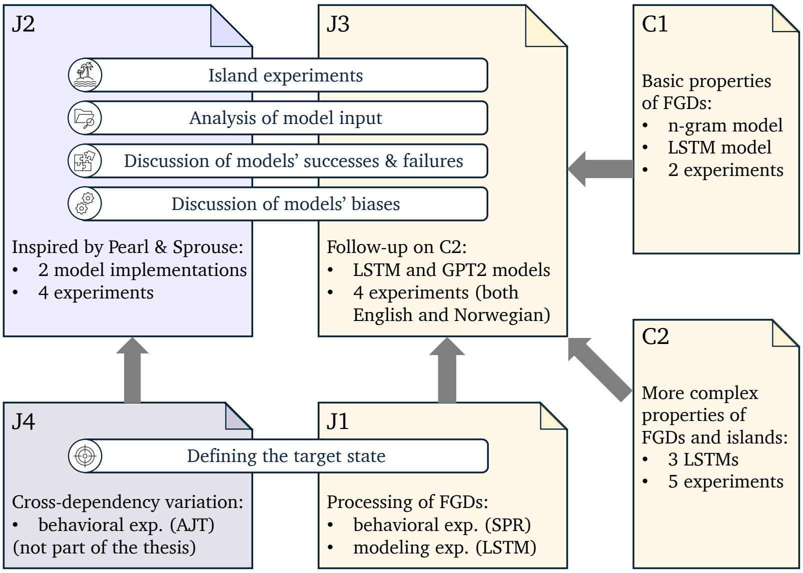

J2
J3
C1
Island experiments
Analysis of model input
Discussion of models' successes & failures
Discussion of models' biases
Basic properties of FGDs:
• n-gram model
• LSTM model
• 2 experiments
Inspired by Pearl & Sprouse:
• 2 model implementations
• 4 experiments
Follow-up on C2:
• LSTM and GPT2 models
• 4 experiments (both English and Norwegian)
C2
More complex properties of FGDs and islands:
• 3 LSTMs
• 5 experiments
J4
J1
Defining the target state
Cross-dependency variation:
• behavioral exp. (AJT)
(not part of the thesis)
Processing of FGDs:
• behavioral exp. (SPR)
• modeling exp. (LSTM)

Figure 4.1: Graphical representation of the relationship between articles. Articles in purple relate to the cognitive model inspired by Pearl and Sprouse, while yellow articles present experiments with neural language models, which were inspired by work of Wilcox et al. (2018).

The research work summarized in the next subchapters involved training and evaluating two different kinds of models across a multitude of experiments, which is published in a variety of scientific outlets. To facilitate article review and comparison, the relationship between the five articles that comprise the current thesis is visualized in Figure 4.1

(as advised by Nygaard and Solli 2020). Articles in purple relate to the traditional cognitive model inspired by Pearl and Sprouse, while articles in yellow present experiments employing psycholinguistic assessment of neural language models introduced by Wilcox, Futrell, and Levy (2023); Wilcox et al. (2018). Below I give article summaries in logical (rather than chronological) order.

4.3.1 Summary of J2

The study aimed to investigate whether a simple computational learning model could acquire knowledge of filler-gap dependencies and island constraints on them in Norwegian. The computational modeling approach used in the study was inspired by the work of Pearl and Sprouse on English *wh*-dependencies (Pearl & Sprouse, 2013a, 2013b). We focused on assessing whether the proposed distributional learning model, which tracks n-grams over structurally annotated child input, could successfully capture island a set of constraints in Norwegian, given the possibility of cross-linguistic variation.

As input, the modeled learner received syntactically parsed sentences with *wh*- and relative clause (RC) filler-gap dependencies. The learner tracked syntactic structures along the path from filler to gap, called *container nodes*. The learner then used the probabilities of attested container nodes to estimate the probability of novel FGDs. This probability was calculated as the product of probabilities of its container node trigrams. The modeling results were compared to human behavioral data from an acceptability judgment experiment.

Compared to the original model of Pearl and Sprouse, we made several modifications. First, as model input we used a corpus of child-directed Norwegian text annotated using the Lexical-Functional Grammar (LFG) formalism (instead of phrase structure (PS) formalism). LFG assigns distinct constituent structure (c-structure) and functional structure (f-structure) representations to sentences, with f-structure encoding predicate-argument structure and hierarchical functional relations. This allowed us to track more fine-grained information about functional relations in a sentence through a richer inventory of labels. This, in turn, allowed us to test different n-gram window sizes, and we ran separate versions of the modeled learner that tracked the frequencies of container node bigrams and trigrams. Additionally, we used bootstrapping techniques to construct confidence intervals around the probability estimates for tested FGDs, providing a measure of uncertainty around the predicted probabilities.

The main results showed that the proposed learning strategy could capture some patterns of island-insensitivity in Norwegian when direct evidence was present in the input. Specifically, the learner successfully induced the non-island status of RC-dependencies into embedded questions in Norwegian. However, the modeled learner failed to learn other important patterns due to a lack of relevant data in the training corpus. For example, the modeled learner struggled with embedded subject questions because the input lacked direct evidence of this undoubtedly grammatical structure. We also found that f-structure bigrams might be sufficient for capturing the relevant co-occurrence statistics, given that f-structure paths tend to be shorter and encode structural features more compactly than their PS counterparts. This result emphasized the trade-off between formalism complexity and n-gram window size for this learning strategy.

We concluded that while the distributional learning model can approximate human-like knowledge of filler-gap dependencies and island constraints in some cases, it is not sufficient to fully recover this knowledge cross-linguistically given limited input data. The study further highlighted the importance of direct evidence in the input for successful acquisition and suggested that additional biases or modifications to the proposed model may be necessary to overcome data sparsity issues.

4.3.2 Summary of C1

The article C1 titled *"LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian"* presented a preliminary study on learnability of Norwegian filler-gap dependencies by neural language models. The study aimed to determine if the previous success of neural language models in learning FGDs in English (Wilcox et al., 2018) is due to specific properties of the English language or if it reflects a more general capability of these models to learn abstract syntactic generalizations across languages and constructions.

We trained a Long Short-Term Memory Recurrent Neural Network model (LSTM RNN) and a baseline n-gram model to explore their ability to learn basic facts about *wh-* and RC-filler-gap dependencies in Norwegian. We tested how the two models would fare as incremental language processors by looking at *surprisal*, which measures how (un)predictable a word is in context. Our experiments used a 2×2 factorial design introduced by Wilcox et al. (2018). The design manipulated the presence of a filler and the presence of a gap in a sentence, as illustrated by an experimental item in (1).

(1) a. −FILLER, −GAP:
She knows that the priest revealed the secret
b. +FILLER, −GAP:
*She knows what the priest revealed the secret
c. −FILLER, +GAP:
*She knows that the priest revealed ___
d. +FILLER, +GAP
She knows what the priest revealed ___
...in front of the guests at the party.

We investigated the models' ability to learn FGDs by looking at how the presence of a filler affects surprisal in two different pairwise comparisons. *Filled-gap effects* were measured by comparing surprisal associated with an NP in −GAP conditions. If a model shows sensitivity to FGDs, it should exhibit positive filled-gap effects at the region of the filled NP (e.g., surprisal at *the secret* in (1-b) should be higher than in (1-a)). Filled-gap effects indicate whether the presence of a filler triggers an active expectation for a gap in an upcoming NP position. *Unlicensed gap effects* were measured by comparing surprisal associated with a gap in the +GAP conditions. We expect unlicensed gap effects to be negative (e.g., surprisal at *in front of* should be larger in (1-c) compared to (1-d)). The unlicensed gap effects measure how 'surprised' the model is to find a gap without a licensing filler, arguably tapping into the model's notion of grammaticality.

We ran two experiments. Experiment 1 tested the models' ability to learn that fillers can license gaps in different syntactic positions, adding the factor POSITION to the 2×2 design specified above (POSITION had three levels: subject, direct object, and oblique). Experiment 2 examined whether the models' representation of FGDs is robust to intervening material. To do so, we manipulated the linear distance between the filler and the gap by varying the length of a phrase modifying a subject that came between the filler and the gap (the phrase MODIFIER had four levels: no, short, medium, and long).

Our results showed that the LSTM model could learn that fillers can be associated with gaps in different syntactic positions by showing the expected filled-gap effects and unlicensed gap effects, while the baseline n-gram model failed to learn all but subject filler-gap dependencies. The LSTM model also demonstrated robustness to intervening material, exhibiting non-zero filled-gap effects and unlicensed gap effects even with increased linear distance between the filler and the gap, while the baseline n-gram model showed zero effects across all conditions. The results were consistent across both *wh*-and RC-filler-gap dependencies, indicating that the LSTM's ability to learn FGDs is not limited to a single construction type.

We interpreted these results as evidence that LSTM RNNs can learn basic syntactic generalizations about FGDs in languages other than English, suggesting that the models' success is not solely due to the distributional properties of English. The findings support the conclusion that LSTMs have a robust ability to learn abstract generalizations about FGDs across different languages and dependency types, highlighting the potential of general-purpose neural language models to capture abstract syntactic structures.

4.3.3 Summary of C2

The article C2 titled "Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian" followed up on the article C1 by testing the models on island environments. The article aimed to determine if LSTMs can capture not only the basic properties of filler-gap dependencies, but also island constraints on their distribution, given the possibility of cross-linguistic variation in island facts. Moreover, to test the robustness of the previous results, we trained two additional LSTMs and presented the averaged results across three models.

We focused on four potential island environments: subject phrases and temporal islands, which are islands in both English and Norwegian, and two types of embedded questions, which are islands in English but not in Norwegian. We ran five experiments. Experiment 1 tested the models' ability to handle FGDs across multiple layers of sentential embedding, thereby testing whether the models can associate fillers and gaps through increased *structural* distance, which is a prerequisite for island experiments. Experiments 2 and 3 examined whether the models could learn that subject phrases and temporal adjunct clauses are islands in Norwegian, similar to what Wilcox et al. (2018) found for English. Experiment 4 tested if the models could learn that embedded adjunct questions are not islands in Norwegian. Experiment 5 tested embedded polar questions (*whether*-islands) and presented a direct comparison between English and Norwegian LSTMs. We used the same factorial design measuring filled-gap effects and unlicensed gap effects as in C1.

The results of Experiment 1 showed that filled-gap effects and unlicensed gap effects decreased as the number of sentential embeddings increased. Despite the reduction, the effects remained non-zero even at the largest structural distance (five levels of sentential embedding), suggesting that the models learned that FGD formation is unbounded. Experiments 2 and 3 showed that the models could learn that subject phrases and temporal adjuncts are islands in Norwegian, similar to English. Experiment 4 that tested embedded questions found an asymmetry between the effects. Filled gap effects were small or close to zero across both control and test conditions, while unlicensed gap effects were large. Importantly, both effects in the island condition were comparable to the effects in the declarative control, suggesting that the models treat EQs and embedded declarative clauses similarly with respect to FGD formation in Norwegian. Finally, Experiment 5 that presented a comparison between English and Norwegian on *whether*-islands showed that both effects were smaller in English than in Norwegian. In Norwegian, robust filled gap effects were observed in both declarative control and *whether*-clause environments, while in English, no filled gap effect was observed inside a *whether*-island. We concluded that the models could learn that *whether*-clauses are not islands in Norwegian, unlike English.

We interpreted these results as evidence that LSTMs can induce complex syntactic knowledge, including cross-linguistic differences in island effects, from their input data. This finding suggested that the input to the models must provide sufficient evidence for learning the distribution of FGDs and island constraints in different languages, paving the way for potential reassessment of the importance of domain-specific learning biases in acquiring island constraints from the input.

4.3.4 Summary of J3

The article J3 titled “Learning Filler-Gap Dependencies with Neural Language Models: Testing Island Sensitivity in Norwegian and English" followed up on articles C1 and C2 and investigated whether two types of NLMs can learn complex linguistic generalizations about FGDs and island constraints in English and Norwegian. In particular, we conducted a controlled cross-linguistic comparison aimed to determine whether NLMs can learn that FGDs are structurally unbounded in both English and Norwegian, can induce island constraints common to both languages, and learn patterns of cross-linguistic variation in island facts. Moreover, we asked whether the Norwegian models received direct evidence of island-crossing dependencies in their training data and conducted a restricted corpus analysis to answer that question.

We used two LSTM RNNs (as previously) and trained two Transformer-based models (specifically GPT-2-small variant) on Norwegian and English Wikipedia texts. We conducted four experiments to test the models' ability to learn FGDs and island constraints. In the experiments, we used the same factorial design measuring filled-gap effects and unlicensed gap effects as in C1 and C2. Experiment 1 tested whether the models could learn that FGDs are unbounded by varying the number of embedded clauses between the filler and the gap. Experiment 2 examined sensitivity to subject islands, where gaps within subject phrases should be ungrammatical in both English and Norwegian. Experiment 3 explored whether the models could learn that embedded polar questions are islands in English but not in Norwegian. Experiment 4 tested the models' ability to recognize that embedded adjunct questions are islands in English but not in Norwegian. Finally, the Norwegian Wikipedia corpus analysis explored whether the models received direct evidence of FGDs crossing into embedded questions and how such 'island-violating' examples aligned with our experimental items.

The results showed mixed evidence regarding the models' ability to learn FGDs and island constraints. The models successfully learned some generalizations, such as the unbounded nature of FGDs without overt complementizers, the island status of subject phrases in both languages and the differential island status of embedded polar questions in Norwegian and English. However, the models struggled with more complex environments: In the case of unboundedness, the models showed significantly reduced filler effects when the complementizer introducing the embedded clause was present, suggesting potential undergeneralization compared to human judgments. The English models also displayed non-zero unlicensed gap effects in subject positions inside embedded adjunct questions, thus establishing ungrammatical FGDs into *wh*-islands in English, suggesting overgeneralization. The corpus analysis revealed that the Norwegian models received scarce direct evidence for dependencies crossing into embedded questions, suggesting a potential cross-dependency generalization, but more research should be done to shed light on the exact nature of models' generalizations.

We concluded that while NLMs can acquire some sophisticated generalizations about FGDs in English and Norwegian, their overall predictions still diverge from human judgments. The models' successes in learning cross-linguistic variations suggest that domain-general learning procedures can recover some patterns from the input.

However, the failures to approximate target generalizations, such as complementizer-dependent boundedness and overgeneralization of wh-islands, indicate that domain-specific biases may still be necessary for accurate language acquisition. Thus, the current evidence does not support the claim that FGDs and constraints on them can be acquired without domain-specific biases.

4.3.5 Summary of J1

This study investigated the mechanisms behind active filler-gap dependency resolution during sentence comprehension in Norwegian. The primary goals were to determine why active gap-filling is suspended in island domains and to assess whether this behavior is best explained by grammar-based or simple processing accounts. The study also aimed to explore if active filler-gap processing could be understood within an expectation-based framework by comparing surprisal values derived from a neural language model to reading times from a human behavioral experiment. Two experiments were conducted to address these goals.

Active gap-filling refers to the process where comprehenders eagerly posit gaps in upcoming positions before confirming their true location, a behavior supported by evidence from various languages and methodological frameworks (Lee, 2004; Omaki et al., 2015; Stowe, 1986; Traxler & Pickering, 1996). Active gap-filling has been shown to be suspended inside island domains such as embedded questions (Phillips, 2006; Stowe, 1986; Traxler & Pickering, 1996). Simple processing accounts suggest that this suspension is due to inherent processing complexity of islands given limited resources of the human language processor (Hofmeister & Sag, 2010; Kluender & Kutas, 1993). Such accounts posit that structural complexity of filler-gap dependencies crossing into island domains makes it too taxing for the language processor to maintain unintegrated fillers in memory. In contrast, grammar-based accounts argue that the suspension is guided by language-specific grammatical constraints, predicting that active gap-filling is only blocked in languages where embedded questions are grammatical islands (Phillips, 2006; Wagers & Phillips, 2009). The two accounts make divergent predictions for the processing of filler-gap dependencies that cross into embedded questions in Norwegian — a non-island domain in the language.

In Experiment 1, a self-paced reading study, we tested whether native Norwegian speakers actively posit gaps in embedded declarative clauses and embedded questions. Participants read sentences with filler-gap dependencies with an oblique gap inside an embedded clause, as well as matched sentences without a dependency. Their reading times were measured to identify filled-gap effects as manifested by increased reading times on filled subject and object NPs in sentences with a filler-gap dependencies as compared to ones without. The results from Experiment 1 showed that Norwegian participants exhibited filled-gap effects in both embedded declarative clauses and embedded questions, indicating active gap-filling in both domains. We concluded that our findings supported grammar-based accounts of active gap-filling, suggesting that the suspension of gap-filling in certain domains is guided by language-specific grammatical rules rather than language-agnostic limitations on human sentence processor.

When it comes to the second goal of the study, we asked if active filler-gap processing can be understood as a special case of probabilistic ambiguity resolution within an expectation-based framework (Hale, 2001; Levy, 2008). To do so, we tested whether word-by-word surprisal values from a neural language model could predict the location and magnitude of filled-gap effects in our behavioral data. Experiment 2 found that while surprisal values could predict the location of filled-gap effects, they significantly underestimated the magnitude of these effects compared to the empirical data. These findings suggest that either additional mechanisms beyond simple predictability are needed to explain the full cost of filled-gap effects or that LSTM-derived surprisal values are not adequate proxies for human expectations during incremental filler-gap resolution.

Chapter 5

Main findings and discussion

This dissertation set out to investigate whether the acquisition of island constraints across languages and constructions can be successfully modeled using two types of approaches: symbolic cognitive models and neural network-based language models. The articles in this dissertation extended the line of research initiated by Pearl and Sprouse and Wilcox and colleagues by applying the ideas behind their modeling approaches to Norwegian data. I identified three main empirical research questions presented below, repeated from Figure 3.2 (for model-specific questions, see discussions in the corresponding articles in Part 2 of the thesis).

1. Can the models induce the (non-)island status of different domains in Norwegian?
2. What kind of evidence does the input provide, and how does it influence model generalizations?
3. Can the models generalize beyond the input — for example, across constructions?

The main findings of the five research articles paint a mixed picture of modeling successes. With respect to RQ1, both modeling approaches found conflicting evidence for the learnability of different (potential) island domains in Norwegian. Interestingly, the modeling successes were alike, whereas each model was unsuccessful in its own way. On the one hand, both types of models could pick up on the patterns of island-insensitivity pertaining to embedded questions and correctly induce the island status of subject phrases in Norwegian. On the other hand, the articles uncovered different ways in which the models could fail to approximate the target state, with the modeling failures caused by different reasons. For the symbolic model, the failures could be explained by sparsity issues found in the input, as well as over-reliance of the model on input distributions and its inability to generalize from neighboring structures. For the NLMs, the failures arguably stem from the fragile nature of learned representations, with the models not consistently encoding abstract syntactic generalizations that govern FGDs. As a result, NLMs occasionally over- and undergeneralize from the input in non human-like ways.

When it comes to RQ2, I found that 'island-violating' dependencies were present, but not abundant, in the input to the models. Specifically, I found evidence for RC- and topicalization dependencies into embedded questions — the typologically marked structure that represents one of the most prominent cases of 'island-insensitivity' found in Norwegian. Examples of such structures were found in both the Wikipedia corpus used to train the NLMs and the corpus of child- and youth-directed texts used to train the symbolic model. This direct positive evidence arguably was enough for the models to induce the non-island status of (at least) polar embedded questions for RC-dependencies in Norwegian. The corpora analyses also revealed that the input to the models is impoverished when it comes to other island-crossing dependencies, such as *wh*-dependencies into embedded questions, as well as other grammatical structures.

In relation to RQ3, I found limited evidence for the models' ability to generalize across constructions, and beyond the input more broadly (see more on this in Section 5.3 below). The symbolic model that treats embedded questions and relative clauses as separate classes for FGD-formation is inherently unable to use other FGD-types as indirect evidence. Because of this, the model came to an incorrect generalization that embedded questions are islands for *wh*-dependencies in Norwegian due to data sparsity issues in the input. On the other hand, it appears as if NLMs could use this indirect positive evidence, treating *wh*-dependencies into embedded questions on par with corresponding RC-dependencies. However, more research is needed to assess whether NLM's performance is best explained by sensitivity to underlying linguistic representations or by surface-level corpus heuristics.

Below, I discuss some of the research findings in greater detail (Section 5.1). I consider the implications and limitations of these findings for learning biases in the acquisition of filler-gap dependencies (Section 5.2), input analyses and the poverty of the stimulus challenges (Section 5.3), and the role of linking hypotheses in modeling research (Section 5.4). I conclude by summarizing directions for future research (Section 5.5).

5.1 Overview of modeling successes and failures

Both modeling approaches demonstrate notable successes in capturing core properties of FGDs. First, both types of models learn that FGDs can span multiple clauses. The symbolic model achieves this generalization by tracking n-gram probabilities over hierarchical representations, supported by examples of long-distance extraction found in the training corpus. NLMs, on the other hand, exhibit robust filled-gap and unlicensed gap effects across multiple levels of embedding, particularly when overt complementizers are absent. Second, both models correctly predict that subject phrases are islands in Norwegian, as well as in English. The symbolic model assigns low probabilities to subject island violations due to the absence of relevant structural n-grams, while NLMs show near-zero filler effects in subject island conditions, with both model types aligning with human acceptability judgments. Finally, and most importantly, the models capture some patterns of cross-linguistic variation in island constraints. On the one hand, the symbolic model shows no island effects for RC-extraction out of two types of embedded questions in Norwegian, reflecting the distributional facts of the input. On the other hand, NLMs trained on Norwegian Wikipedia correctly allow FGDs into embedded polar questions (i.e., *whether*-clauses), while English-trained models do not, mirroring native speaker intuitions. Moreover, Norwegian NLMs are able to link *wh*-fillers to gaps inside embedded adjunct questions (‘*wh*-islands’), although the models do not actively expect gaps in such environments. These findings show that some distributional patterns pertaining to FGDs, including island constraints on their distribution, are learnable with fewer (and perhaps up-to no) domain-specific biases than are assumed by older generative proposals (Chomsky, 1986, 2001; C. T. J. Huang, 1982).

Despite these successes, both model types exhibit important limitations. The symbolic model fails in cases where objectively acceptable structures are absent from the input. For example, it fails to learn that embedded subjects can be extracted in *wh*-questions (e.g., *‘Who did you say called?’*) due to the lack of direct evidence for this undoubtedly grammatical structure (see more on this in Section 5.2 below). Similarly, it treats *wh*-extraction from embedded questions as an island violation, reflecting the lack of relevant examples in its training data. When it comes to NLMs, they show reduced performance on deeply embedded FGDs when overt complementizers are present, suggesting a narrower generalization than human learners adopt. Overall, the sensitivity of NLMs to FGDs decreases with increased complexity of material intervening between the filler and the gap, reflecting the brittle nature of learned representations — a finding that echoes much previous research on FGDs (Chaves, 2020; Da Costa & Chaves, 2020; Howitt et al., 2024; Lan et al., 2024). Moreover, the cross-linguistic comparison reported in article J3 revealed that NLMs trained on English text sometimes overgeneralize beyond their training data, predicting that FGDs into embedded adjunct questions (*wh*-islands) are possible, contrary to human judgments. Together, these findings represent cases where models' performance falls short of human target state.

5.2 Generalization and nature of learning biases

The performance of both types of models is shaped by their inherent inductive biases.

The symbolic model under investigation is guided by both representational biases and biases on the inference mechanism. It operates over structured, hierarchical representations of filler-gap dependencies, such as LFG f-structure labels or phrase structure trees, keeping track of container nodes — structural elements between the filler and the gap. It then calculates probabilities over n-grams of container nodes to derive a probability distribution over observed dependencies, which it can use to determine the probability of unseen dependencies.

How do these biases support generalization from observed patterns? The decision to subset the input to specific FGD types (e.g., *wh*-dependencies) reflects a domain-specific bias, which, as argued by Pearl and Sprouse, is motivated by empirical necessity (Pearl & Sprouse, 2013b, p. 48). That is, this bias is crucial to model's generalizations, as without it, the modeled learner would not be able to arrive at the target state. The representational bias to encode FGDs as sequences of container nodes abstracts away from surface lexical content, enabling the model to generalize across structurally similar but lexically distinct sentences. The bias to track n-grams of container nodes allows the model to generalize from smaller, recurring structural units, even when full dependency paths are rare or absent in the input. Additionally, subcategorizing complement phrases with lexical or clause-type information (e.g., CP$_{that}$ vs. CP$_{whether}$ or COMP$_{nominal}$ vs. COMP$_{wh-int}$) enables the model to distinguish between different types of embedded clauses, which is crucial for learning island constraints.

In this model, the combination of a specific set of container nodes (i.e., the hard-coded f-structure labels) and the bias to learn from a subset of the input do much of the “generalization heavy lifting" for successful learning. The way syntactic information is encoded is critical as it directly affects the learner’s ability to extract generalizations. In the case of islands, in both of the formalisms there is a unique label (or label sequence) for each potential island domain, which allows the model to distinguish between island-violating and licit dependencies. If the set of container nodes was specified more coarsely — for example, if it did not distinguish between different clause types — the modeled learner would not be able to arrive at the target state.

Compared to phrase structure, LFG’s f-structure richer label set allows for more compact and semantically informative representations, but this granularity can also introduce sparsity when distinctions are not well-supported by the input data. For example, consider embedded object extraction in a sentence like *‘What did you say she ate?’*. The container node sequence differs significantly between the two formalisms: IP-VP-CP$_{null}$-IP-VP in the phrase structure vs. COMP$_{nominal}$-OBJ in LFG. These sequences differ not only in length but also in the granularity of their labels. The f-structure representation, while more compact, requires the learner to make finer-grained distinctions. This can negatively impact generalization: from this particular example, the phrase structure learner will generalize that if an object extraction is possible, then subject extraction is possible too (IP-VP-CP$_{null}$-IP). The LFG learner, on the other hand, must learn two distinct generalizations: that subjects can be extracted from embedded clauses and, independently, that objects can be extracted from embedded clauses.

Because of the finer-grained label set, a concrete sparsity issue arose in our Norwegian data: in LFG, modals introduce an additional XCOMP layer in the f-structure of a sentence. One of our test conditions involved embedded subject extraction with *wh*-dependencies (COMP$_{nominal}$-SUBJ), but the corpus lacked examples of this structure. That is, except for a single sentence where the embedded clause contained a tensed modal, yielding the path COMP$_{nominal}$-XCOMP-SUBJ. As a result, the learner had to make two separate generalizations: that subjects can be extracted from embedded clauses with and without modals. This illustrates how the success of this learning strategy depends on the granularity of the input representation and the availability of relevant evidence.

Therefore, the learner's biases also can negatively impact model performance when input data is sparse. More broadly, the modeled learner's generalization is limited to what can be decomposed into n-grams; it cannot generalize when no relevant evidence exists in the input data. The bias to treat different FGD types separately (e.g., *wh-* vs. RC-dependencies) can also be problematic. In Norwegian, it seems like generalization across dependency types is sometimes necessary to reach the target grammar (Kush et al., 2021). However, collapsing across dependency types would obscure cross-dependency variation, which is empirically attested (Kobzeva, Sant, et al., 2022). A promising model modification would then be to allow the model to learn from such *indirect positive evidence* but treat different evidence types distinctly. For example, the learner could assign lower weight to RC dependencies when calculating the probability of *wh*-dependencies than to direct evidence of the latter. This would enable partial generalization across dependencies while preserving sensitivity to cross-dependency distinctions.

In our article J2 (Kobzeva & Kush, 2025), we concluded that with a sufficiently representative corpus, both LFG and phrase structure formalisms encode domain-specific information that can support the acquisition of target generalizations in the majority of cases. To summarize, this symbolic model benefits from strong hierarchical language biases and relatively simple inference mechanisms. This enables it to learn certain patterns with minimal input, but also limits its generalization capacity in the face of sparse data. Our experiments highlight how the choice of structural formalism influences both learning outcomes and the generalization capacity of the model.

Overall, both the original Pearl & Sprouse model and our extension to Norwegian show that knowledge of island constraints can be induced from prior knowledge of FGDs, which in turn depends on structural configurations of a sentence. These findings raise broader, important questions: How is syntactic information represented in the human learner's mind? Which syntactic formalism best approximates this representation? And how does such structure arise in the learner — does it reflect innate constraints, such as structure dependence, or can it be induced from input's distributional properties? Although current linguistic theory cannot yet offer definitive answers to these questions, symbolic computational models provide a valuable testing ground for evaluating the plausibility of different learning biases.

Neural language models' biases are more difficult to characterize. Wilcox and colleagues (2023) categorize LSTM and Transformer NLMs that they test as *domain-general weakly biased* learners that have a preference for maximizing data likelihood. Following Clark and Lappin (2011), they take *domain-general* to mean that the models can learn any arbitrary type of vectorized input, not limited to language data. Following Lappin and Shieber (2007), they take *weakly biased* to mean 'uncomplicated, uniform and task-general' as opposed to *strongly biased*, which is 'articulated, non-uniform and task-specific'. Although we do not exactly know which biases NLMs posses, this classification allows us to use them as powerful domain-general learners for testing linguistic learnability (although it is important to note that this decision is based on the assumption above).

So what can such models learn from unannotated text data? Based on the experiments presented in this dissertation, it would seem that the models' architectural biases (e.g., recurrence or attention), coupled with their powerful inference mechanisms and large-scale training data, allow the models to capture many abstract properties of FGDs. In particular, the models learned the bidirectional relationship between fillers and gaps, the flexibility and unboundedness of filler-gap licensing (at least to some extent), and several island constraints. These results echo much previous work showing that the language modeling objective leads to an approximation of hierarchical representations in NLMs (Ahuja et al., 2024; Hu et al., 2020; Linzen & Baroni, 2021). Although these findings are suggestive, they are perhaps not very surprising as previous work has shown that less powerful learners can also learn that language is hierarchically structured, particularly when coupled with other domain-general pressures like a bias for simplicity (Perfors et al., 2011).

In NLMs, the exact nature of those learned representations is not yet properly understood. Here, we have minimally shown that those representations are more fragile and brittle than the representations that guide human language processing — another finding that is in line with many studies on filler-gap dependencies in particular (Bhattacharya & van Schijndel, 2020; Chaves, 2020; Da Costa & Chaves, 2020; Lan et al., 2024). For example, we found that filled-gap effects in NLMs decrease dramatically with increased structural and lexical complexity between the filler and the gap. This implies that the active gap-filling strategy may operate differently in humans and NLMs. In humans, the pressure to resolve a dependency may increase with each filled NP at a potential gap site, leading to increased filled-gap effects across increased distances between the filler and the gap (article J1, Kobzeva & Kush, 2024). However, in similar contexts, NLMs' expectations for a gap appear to diminish. This discrepancy could again be explained by the fragile nature of NLMs' learned representations — while human FGD-processing is driven by their top-down grammatical knowledge that a filler must be linked to a later gap, NLMs' surprisal-based processing is arguably best explained by a bottom-up likelihood of a particular gap given an upstream filler (K.-J. Huang et al., 2024; van Schijndel & Linzen, 2021, see more on in Section 5.4).

Another open question is whether NLMs can generalize beyond their training data. Our Wikipedia corpus analyses suggest some potential degree of generalization across dependency types — e.g., from RCs to *wh*-dependencies — at least when it comes to embedded questions. In particular, our training corpus contained no direct evidence of *wh*-dependencies into embedded questions, yet the models still exhibited filled-gap and unlicensed gap effects in such contexts. This suggests that the models may have generalized from more frequent RC-dependencies or topicalizations, consistent with the idea of shared underlying representations across FGD types, supported by very recent findings on English FGDs (Boguraev et al., 2025). However, this conclusion is at odds with findings by Howitt et al. (2024) who argued that NLMs may rely on shallow heuristics or lexical co-occurrence patterns rather than abstract syntactic representations for different types of FGDs. Some of our positive results could be partially explained by such heuristics too: for example, the relative success of NLMs on *whether*-clauses in Norwegian could be explained by the fact that complementizer *om* ('if/whether') is polysemous with a preposition ('about/around/during'), and appears in a wide range of distributional contexts. This could potentially decrease the models' surprisal when they are tested on cases of extraction from *whether*-clauses — previous studies show that homonyms can lead to what appears to be the correct generalization (Kam, Stoyneshka, Tornyova, Fodor, & Sakas, 2008), with the models being right for the 'wrong reasons' (McCoy et al., 2019). Future work therefore should explore whether these generalizations persist under controlled manipulations of the training data (e.g., filtered or augmented corpus training; Howitt et al. 2024; Leong and Linzen 2024; Patil et al. 2024). Overall, more research is needed to determine whether these generalizations are linguistically principled or driven by superficial heuristics (Boguraev et al., 2025; Howitt et al., 2024).

So where do these findings leave us when it comes to nature of learning biases required for FGD-acquisition? Recall the language acquisition model presented in Figure 2.1. Representational biases and biases on the inference mechanism are internal components that are under-determined by observable data. They represent theoretical constructs posited by a certain acquisition theory (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010). The empirical findings presented here reflect a *trade-off* between the complexity and domain-specificity of representational biases and the power of the domain-general inference procedures. Specifically, the articles show that some patterns of island-(in)sensitivity are possible to learn with both strong representational commitments and weak learning mechanisms, and weak representational commitments and powerful learning procedures. By experimentally controlling learning biases, we can *systematically manipulate* this trade-off to find the combination that provides a broader coverage of empirical facts and that is deemed most developmentally plausible. Studies conducted here highlight limitations in the models' ability to fully replicate human-like knowledge of FGDs and island constraints, so the exact combination of learning biases remains to be determined. However, they bring us one step closer to understanding which components a successful acquisition theory should incorporate.

The complementary strengths and weaknesses of symbolic and neural models point toward the value of hybrid approaches. A promising direction for future research is to integrate the interpretability and constraint-based reasoning of symbolic models with the generalization capacity of modern NLMs. Such hybrid research already seeks to integrate useful symbolic representations into NLMs (Dyer, Kuncoro, Ballesteros, & Smith, 2016; McCoy & Griffiths, 2025) or constrain their generalizations by imposing human-like extra-linguistic limitations such as working memory limitations (Mita, Yoshida, & Oseki, 2025). On the other hand, traditional symbolic models could benefit from more powerful data-driven inference procedures and linking hypotheses that would support broader generalizations (Dickson, Pearl, & Futrell, 2022).

5.3 Input analysis and poverty of the stimulus

The input analyses of the corpus of child-directed text (article J2, Kobzeva & Kush, 2025) and Wikipedia data (article J3, Kobzeva et al., 2025) have provided valuable insights into the distribution of acceptable filler-gap dependencies in Norwegian. In line with the previous corpus study (Kush et al., 2021), we found that the input to the models contained examples of typologically marked 'island-crossing' dependencies. In particular, both the children's fiction corpus from NorGramBank (Dyvik et al., 2016) and the Wikipedia data contained examples of RC- and topicalization dependencies into embedded questions, illustrated in (1) below.

(1) a. [...] en situasjon$_i$ regjeringen ikke visste hvordan de
[...] a situation government.DEF NEG knew how they
skulle håndtere ___$_i$
should handle
'[...] a situation$_i$ (that) the government did not know how they should handle ___$_i$.'[1]

b. [...] en stillhet$_i$ Naledi ikke visste hvordan hun skulle bryte
[...] a silence Naledi NEG knew how she should break
___$_i$

'[...] silence$_i$ (that) Naledi did not know how she should break ___$_i$.'[2]

However, the dependencies found in the corpora were not representative of the full range of dependencies that Norwegian speakers judge as acceptable. Notably, *wh*-dependencies into embedded questions were unattested, despite receiving high acceptability scores in controlled acceptability judgment tasks (Kobzeva, Sant, et al., 2022; Kush et al., 2018). Furthermore, the attested sentences often followed similar structural patterns: in such 'island-violating' dependencies, embedded questions were predominantly introduced by verbs *vite* 'to know', *skjønne* 'to know/understand', and *forstå* 'to understand', which together accounted for 69% of the sample. Additionally, the majority of these examples featured negation of the embedding verb (62.5%).

These findings suggest that Norwegian learners must generalize beyond their input to acquire the full range of acceptable dependencies (Kush et al., 2021). The symbolic model abstracts away from such lexical properties but faces poverty of the stimulus challenges when it comes to cross-dependency generalization and other acceptable 'island-violating' dependencies.[3] In contrast, neural language models (NLMs)

[1] Source: Strikking Wikipedia page

[2] Source: Ildlenken by Beverley Naidoo

[3] Such as dependencies into presentational relative clauses. Although we did not target

appear to exhibit a general sensitivity to island constraints across dependency types, but are likely to face POS challenges when it comes to nested embedded clauses introduced by overt complementizers, preventing the models from learning target generalizations. Similarly, the modeled symbolic learner would assign a very low probability to deeply nested structures due to the multiplication of attested n-grams, with the final probabilities in the range of probabilities for island violations (see more on this in Section 5.4).

Perhaps unsurprisingly, the models performed best on structures for which direct evidence was present in the input. Interestingly, the analysis of a 4-million-word corpus of child fiction texts revealed approximately the same number of 'island-crossing' dependencies (n=31) as a non-exhaustive search in 113 million words of Wikipedia text (n=33), potentially highlighting the conversational nature of these structures. Given that child fiction texts are more similar to child- and adult-directed speech than Wikipedia articles, it is reasonable to assume that a corpus of actual child-directed speech would contain even more relevant examples — particularly of *wh*-dependencies, as child-adult interactions typically include more questions and embedded questions than written texts (Noble, Cameron-Faulkner, & Lieven, 2018).

Because our input corpora differ qualitatively from child-directed language, it is possible that models trained on a sufficiently rich corpus of child-adult interactions would perform better. However, other features of child-directed speech — such as noise, disfluencies, and incomplete sentences — may also negatively influence model generalization, and such analyses could offer potential insights into how acquisition might proceed in a more naturalistic scenario. Given the symbolic model's sensitivity to input distribution and the data-hungriness of modern NLMs, constructing a large-enough developmentally plausible training corpus in Norwegian might be a worthwhile direction for future research.

For NLMs in particular, there is a growing body of research on English attempting to bridge this data gap by training NLMs on developmentally appropriate input from the CHILDES database (Warstadt et al., 2023). In the case of Norwegian, however, there is no large-enough collection of child-directed input, but other available sources

such dependencies in our corpus analyses, we did find a few of them when checking the Wikipedia search results manually. Such structures also consistently pop up in news articles. For example: *Og nettopp dette treet*$_i$ *er det mange bokglade berlinere som oppsøker* ___$_i$ 'And precisely this tree$_i$, there are many book-loving Berliners who seek out ___$_i$' (Source: NRK's article "Bytter bøker på trær").

could be combined to serve as a better input to the models. In an ideal scenario — where a PhD candidate has infinite time — I would construct a training corpus that is both naturalistic and large enough to support training of both symbolic and neural models. This could involve combining resources like several syntactically annotated corpora from NorGramBank (for example, adult fiction corpus), transcribed speech corpora such as the NoTa corpus and The Big Brother corpus, and whatever small corpora are available for Norwegian in the CHILDES database (Dyvik et al., 2016; Garmann, Hansen, Simonsen, & Kristoffersen, 2019; Tekstlaboratoriet, 2005, 2009). The unannotated corpora could then be syntactically parsed using NorGram computational grammar to obtain LFG-analyses (Butt, Dyvik, King, Masuichi, & Rohrer, 2002; Dyvik, 2000). While time-consuming, this effort would be worthwhile, as it would allow us to: (i) directly compare the performance of NLMs and symbolic models trained on the same data; (ii) quantify the amount of different evidence types available in the input; (iii) conduct filtered or augmented corpus training experiments to better understand the nature of learned representations in NLMs and/or to explore the role of different evidence types for symbolic models.

5.4 Linking hypotheses in modeling research

In language modeling, probability serves as a central tool for linking linguistic structure to behavioral data. This approach is widely employed in both top-down symbolic models and bottom-up neural network-based models. However, using probability as a linking hypothesis between human grammatical knowledge and model performance has limitations in capturing the full complexity of human language.[4]

One of the primary challenges in equating probability with grammaticality is that grammaticality encompasses more than just statistical likelihood. For instance, longer sentences tend to have lower

[4] The nature of human grammatical knowledge remains a topic of ongoing debate. Traditionally, linguistic theory has posited that humans possess a categorical grammar. In recent years, however, probabilistic grammars have gained traction as an alternative or complementary framework, offering robust empirical coverage — especially in light of *gradience* observed in various behavioral measures (Lau, Clark, & Lappin, 2016). One way to reconcile gradient results with categorical grammars is to incorporate extra-grammatical factors such as performance limitations and processing constraints (De Santo, 2020; Dickson et al., 2024; Hofmeister & Sag, 2010), or to integrate probabilistic components into structure formation itself (Villata & Tabor, 2022; Yang, 2008). According to Yang (2008), probabilistic and categorical grammars need not be mutually exclusive; for example, Probabilistic Context-Free Grammars (PCFGs) are categorical grammars with probabilities attached to their rules.

probabilities due to their lower frequency in corpora, yet they may still be perfectly grammatical. Conversely, short ungrammatical sentences composed of frequent elements may receive relatively high probability scores. This mismatch highlights a fundamental limitation of probabilistic models (cf. Lau et al., 2016; Phillips, 2013a). To mitigate this issue, the articles in this dissertation adopt a well-established practice in modeling research: comparing probabilities of carefully constructed minimal pairs of sentences. While this approach controls for many confounding variables, it does not fully resolve the underlying limitations of probabilistic linking hypotheses. For example, the phenomena of parasitic gaps and ATB-extraction, where specific additional gaps make an otherwise unacceptable gaps inside islands acceptable, are extremely rare in naturalistic corpora. Recently, they were shown to be difficult to learn for several modern NLMs when tested using this minimal pair setup (Lan et al., 2024). When the input to one of the models was enriched with examples of relevant structures, its performance significantly increased, emphasizing the challenges that such probabilistic learning and testing strategy faces (Lan et al., 2024).

Setting these limitations aside, there are multiple ways to compute probability. In this dissertation, I employed two primary measures: *syntactic probability*, derived from frequency distributions of syntactic structures in corpora, and neural network-induced *surprisal* calculated on a word level. As discussed in article J2 (Kobzeva & Kush, 2025), syntactic probability does not always consistently align with human acceptability judgments. While it can qualitatively capture certain patterns — such as superadditive effects in island constraints — it falls short of achieving quantitative alignment. This is because human judgments are shaped by a complex interplay of lexical, semantic, syntactic, discourse-pragmatic, and processing-related factors, almost all of which are abstracted away from in syntactic probability estimation. This abstraction inevitably leads to discrepancies between model predictions and human intuitions (cf. Kobzeva & Kush, 2025, p. 31).

In contrast, surprisal — defined as the negative log probability of a word given its context — offers a broader-coverage probabilistic linking hypothesis. NLM-induced surprisal calculated at a word level implicitly integrates lexical, semantic, and syntactic cues that influence word predictability. According to the influential surprisal theory (Hale, 2001; Levy, 2008), much of human linguistic processing can be explained in terms of predictability as measured by surprisal.

Empirical findings across a wide range of linguistic phenomena support the predictions of surprisal theory, which is reflected in various behavioral measures such as reading times, eye-tracking data, and ERP responses (Michaelov et al., 2024; Shain et al., 2024; Wilcox, Meister, et al., 2023). For instance, recent large-scale multilingual studies have demonstrated that surprisal robustly correlates with reading times across diverse languages and linguistic structures, reinforcing its cross-linguistic validity (Wilcox, Meister, et al., 2023). Overall, surprisal is a strong and generalizable predictor of processing effort. Consequently, if empirical coverage is prioritized in modeling, surprisal emerges as a more powerful and comprehensive measure than syntactic probability, as it captures a multitude of factors relevant to sentence processing. However, such comparison should be made cautiously. While using syntactic probability as a linking hypothesis requires a certain theoretical commitment (e.g., syntactic sentence analyses are best represented as phrase structure trees), surprisal itself is representation- and theory-agnostic (Futrell, Gibson, & Levy, 2020). While surprisal values reflect variance from multiple latent factors, such as syntax, semantics, pragmatics, and word frequency, it does not isolate the specific contributions of these factors to a processing phenomenon at hand. This makes it difficult to interpret the underlying causes of observed effects and to identify the mechanisms responsible for language processing (as well as acquisition). And at that, a data-driven estimate will always perform better than a theory-driven estimate (Slaats & Martin, 2025).

Recent findings show that word surprisal alone cannot account for all of the complexity of human sentence processing (K.-J. Huang et al., 2024; van Schijndel & Linzen, 2021) and some researchers claim that it cannot serve as a mechanistic theory for how humans process language (Slaats & Martin, 2025; Staub, 2025). In line with these findings and views, article J1 in this dissertation (Kobzeva & Kush, 2024) demonstrates that NLM-induced surprisal fails to capture the full magnitude of behavioral filled-gap effects, echoing earlier results on phenomena such as garden-path effects (K.-J. Huang et al., 2024; van Schijndel & Linzen, 2021). These results underscore the need for more nuanced linking hypotheses that go beyond word predictability alone, which should also benefit theory specification. While surprisal theory has convincingly shown that probabilistic inference is an important part of language processing, it is insufficient as a standalone mechanistic explanation for the full range of human language processing.

Therefore, further research is needed to integrate surprisal with other models that address lexical, syntactic, and semantic processing stages (Staub, 2025). In the domain of FGDs — which depend on structural configurations of a sentence — such research could explore alternative linking hypotheses such as syntactic surprisal that incorporates hierarchical information into NLM surprisal calculation (Arehalli, Dillon, & Linzen, 2022). An alternative approach would be to modify structural probability calculation by incorporating additional information from other linguistic levels (e.g., lexicon, Dickson et al., 2022) and/or cognitive constraints (e.g., memory constraints, De Santo, 2020; Dickson et al., 2024). I believe that these two approaches to improving our linking hypotheses represent worthwhile avenues for future research.

5.5 Future research

To summarize, the findings presented in this dissertation highlight both the promise and the limitations of current computational models in capturing the distribution of filler-gap dependencies and island constraints across languages. While both symbolic and neural models demonstrated the ability to learn certain generalizations from input data, their divergent strengths and weaknesses point to several promising directions for future research.

1. Compiling representative corpora for model training
One of the most pressing needs is the development of more representative training corpora that approximate the linguistic input available to children in Norwegian. Such corpora would enable more ecologically valid modeling and allow for direct comparisons between symbolic and neural models trained on the same data. Moreover, they would support filtered or augmented corpus training experiments to assess the role of specific evidence types in acquisition.

2. Improving linking hypotheses for symbolic models
The symbolic model used in this dissertation relies on syntactic probability as a linking hypothesis between model predictions and human judgments. While this approach captures some qualitative patterns, it fails to account for the full range of acceptability effects observed in behavioral data. Future research should explore alternative or enriched linking hypotheses that incorporate additional information (e.g., lexical items), or that integrate cognitive constraints such as memory limitations. One promising direction is to combine syntactic

probability with surprisal-based measures or to develop hybrid metrics that better reflect the multifactorial nature of acceptability.

3. Integrating symbolic structure into neural models
The complementary strengths of symbolic and neural models suggest the potential of hybrid approaches. Future work could explore how NLMs that embed symbolic representations (e.g., parse trees, dependency paths, or functional structures) fare on the acquisition task at hand. Conversely, symbolic models could benefit from incorporating data-driven inference mechanisms inspired by NLMs. Such integration may yield models that are both interpretable and capable of generalizing from sparse input, offering a more cognitively plausible account of language acquisition.

4. Investigating the internal representations of NLMs
While behavioral diagnostics (e.g., minimal pair testing) provide indirect evidence of learned generalizations in NLMs, they do not reveal how these generalizations are encoded. Future research should employ interpretability methods such as probing, causal interventions, and ablation studies to investigate whether NLMs develop unified representations of filler-gap dependencies across constructions and languages. This line of work could help adjudicate between competing accounts of NLM behavior — whether it reflects abstract syntactic knowledge or shallow heuristics — and potentially be useful in shaping more theories of island acquisition.

5. Making NLM learning conditions more human-like
Although NLMs can acquire some sophisticated generalizations, their learning conditions differ dramatically from those of human learners. Future work should explore how NLM performance changes when trained on developmentally realistic corpora, with constraints on data size, modality, and interaction.

6. Expanding the empirical scope of modeling studies
Finally, future research could extend the empirical coverage of modeling studies to include a broader range of languages, constructions, and dependency types. This includes testing models on typologically diverse languages, exploring cross-construction generalization in more detail, and potentially examining the acquisition of other island constraints. Such work would help determine the generality of current findings and refine our understanding of the interplay between input, bias, and generalization in language acquisition.

Chapter 6

Concluding remarks: Biases beyond learning

This final chapter takes a step back from the empirical focus of the preceding chapters and offers a more personal (and perhaps somewhat philosophical) reflection on broader issues that have emerged during the course of this research. While much of this dissertation has examined specific forms of bias in language acquisition research, I would like to conclude by drawing attention to a more general and, in my view, underacknowledged phenomenon: what I call *disciplinary belonging bias*. This bias became increasingly salient to me during the final stages of writing. Here I define it as the tendency for researchers' theoretical standpoint to influence how they interpret results, especially when those results are ambiguous or open to multiple explanations.

This bias is especially visible in current debates surrounding NLMs and their relevance to language acquisition and to approaches to language more broadly (Chesi, 2024; Cuskley et al., 2024; Futrell & Mahowald, 2025; Lan et al., 2024; Piantadosi, 2023; Vázquez Martínez et al., 2024; Ziv et al., 2025, a.o.). As discussed in Chapter 2, different linguistic traditions bring different assumptions to the table about what NLMs can and cannot tell us about human linguistic cognition, and I believe that those assumptions also influence result interpretation. As already noted by Linzen and Baroni (2021) (p. 206), the conclusions of studies employing NLMs will likely depend on the particular notions of competence, performance, and grammar that researchers commit to, which in turn are rooted in their disciplinary backgrounds.

Consider, for example, how similar experimental manipulations grant strikingly different conclusions depending on a theoretical lens. From a more generative perspective, as articulated by Katzir and colleagues (Lan et al., 2024; Ziv et al., 2025), NLMs at best can serve as proxies for evaluating domain-general learning theories — provided those theories are made explicit and methodologically rigorous.

When NLMs fail to acquire certain generalizations (e.g., knowledge of constraints on FGDs), this is taken as support for nativist accounts. When they succeed, however, the success is often dismissed as irrelevant. As Lan et al. (2024) elaborate on the relative success of GPT-3 on cases of across-the-board extraction, 'even if it approximates the relevant patterns, this does not indicate that a general-purpose learner would acquire the relevant knowledge from a developmentally-realistic corpus of just a few years of linguistic experience' (p. 20). This strikes me as a somewhat circular argument: NLMs are then only informative when they fail, and never when they succeed.

This position contrasts sharply with the position of researchers like Warstadt and Bowman (2022) and Wilcox, Futrell, and Levy (2023) who view NLM successes as more informative than their failures. More broadly, researchers working within usage-based or constructivist frameworks tend to view NLMs more optimistically. For example, Pannitto and Herbelot (2022) argue that LSTM-based models can validate usage-based theories of acquisition, while Goldberg (2024) extends this argument to Transformer-based LLMs. From this perspective, NLMs are not just tools for falsifying nativist claims but can be used as plausible testbeds for exploring how linguistic knowledge might emerge from statistical patterns in the input. This view is in even starker opposition with more radical generativists who dismiss NLMs entirely, arguing that they can teach us nothing about human linguistic cognition (Bolhuis et al., 2024; Chomsky et al., 2023; Moro et al., 2023).

This divergence in views on NLMs has real consequences for how we evaluate experimental evidence. In our own work (article J3), we replicated the experimental setup of Wilcox, Futrell, and Levy (2023); Wilcox et al. (2018) on learning island constraints. The authors reported model success rates of 88% on the relative metric (which compared filler effects in islands as compared to control conditions) and 65% on the absolute metric (which assessed whether filler effects were indistinguishable from 0 in island conditions), concluding that their results challenge the poverty of the stimulus argument. In J3, the tested models achieved 83% and 58% on the same metrics respectively — only marginally lower — yet we interpreted the results more cautiously, emphasizing the complexity and potential ambiguity of the findings. Is a 5–7% difference really enough to justify such divergent conclusions? Or are we witnessing the influence of disciplinary belonging on scientific interpretation?

Initially, I found this situation frustrating. If similar data can support such different conclusions, what does that say about the objectivity of our field? More recently, however, I have come to see this not as a failure of language science, but as a reflection of its sociological reality. As Hao (2025) argues, the current debates over LLMs have exposed a deeper *incommensurability* (Kuhn, 1962) between generative and LM-based approaches (i.e., usage-based approaches that see NLMs as useful tools of exploration). These paradigms operate from fundamentally different conceptual frameworks, with distinct assumptions about what counts as evidence, what needs to be explained, and how explanation should proceed. I believe that in this case, debates surrounding NLMs highlight the division in the field of linguistics that goes deeper than the actual role of such models.

These reflections have also shaped the way we approached the articles included in this dissertation. In most cases, we aimed to remain neutral, often offering multiple interpretations of ambiguous results. While this may seem unsatisfying to some, I believe it reflects the kind of epistemic humility that our field needs at the moment (Matthews, 2006). And although the current dissertation does not provide definitive support for either of the major theories of language acquisition, modeling work — particularly of the traditional symbolic kind — holds real potential to adjudicate between them through controlled experimentation.

In a very insightful chapter on the future of experimental syntax (Almeida et al., 2023), Jennifer Culbertson also highlights that disciplinary belonging bias affects us as experimenters whether we want it or not, and perhaps more than we would like to admit. There is no doubt that there are clear limits to what the empirical data can objectively tell us. She advocates for the use of experimental methods to push linguists toward generating and testing precise predictions of their hypotheses and revising theories when those predictions are not met. This call for testable theories is mirrored in many opinion pieces about NLM utility (Chesi, 2024; Ramchand, 2024; Ziv et al., 2025), and I deeply sympathize with it. As Gillian Ramchand (2024) put it, 'language scientists need to reclaim the space for real theory, and not be afraid to use computational modeling as part of their toolbox' (p. 13). However, constructing such theories is no simple task. I hope that work evaluating existing approaches to island acquisition presented in this thesis at least brings us one step closer to formulating better theories, and ultimately solving the puzzle of island learnability.

Looking ahead, I believe that the way forward lies in constructive dialogue across paradigms, grounded in shared methodological standards and a willingness to question our own assumptions. In my view, experimental work that employs computational models (alongside other methods such as artificial language learning) can not only help evaluate linguistic theories but also deepen our understanding of how language behavior interacts with broader cognitive processes. This approach is especially relevant now that computational tools at our disposal are greater than ever before. More broadly, as a field, we should remain open to new tools (including NLMs), new interpretations, and new ways of thinking about what it means to model language acquisition. And we should be honest about the biases — disciplinary, methodological, and personal — that shape our scientific judgments.

References

Abeillé, A., Hemforth, B., Winckel, E., & Gibson, E. (2020). Extraction From Subjects: Differences in Acceptability Depend on the Discourse Function of the Construction. *Cognition*, *204*, 104293. https://doi.org/10.1016/j.cognition.2020.104293

Abrusán, M. (2014). *Weak Island Semantics*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199639380.001.0001

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N. A., Goyal, N., & Tsvetkov, Y. (2024). Learning syntax without planting trees: Understanding when and why transformers generalize hierarchically. *arXiv preprint arXiv:2404.16367*.

Almeida, D., Breen, M., Brennan, J. R., Carlson, K., Chung, S., Culbertson, J., ... Yoshida, M. (2023, 03). 643the future of experimental syntax. In J. Sprouse (Ed.), *The oxford handbook of experimental syntax*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198797722.013.19

Ambridge, B., Rowland, C. F., & Pine, J. M. (2008). Is Structure Dependence an Innate Constraint? New Experimental Evidence From Children's Complex-Question Production. *Cognitive Science*, *32*(1), 222-255. https://doi.org/10.1080/03640210701703766

Arehalli, S., Dillon, B., & Linzen, T. (2022, December). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In A. Fokkens & V. Srikumar (Eds.), *Proceedings of the 26th conference on computational natural language learning (conll)* (pp. 301–313). Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.conll-1.20

Arehalli, S., & Linzen, T. (2024). Neural Networks as Cognitive Models of the Processing of Syntactic Constraints. *Open Mind*, *8*, 558–614. https://doi.org/10.1162/opmi_a_00137

Atkinson, E., Wagers, M. W., Lidz, J., Phillips, C., & Omaki, A. (2018). Developing incrementality in filler-gap dependency processing. *Cognition*, *179*, 132–149. https://doi.org/10.1016/j.cognition.2018.05.022

Baker, C. L., & McCarthy, J. J. (1981). *The Logical Problem of Language Acquisition*. MIT Press.

Bhattacharya, D., & van Schijndel, M. (2020, November). Filler-gaps that neural networks fail to generalize. In R. Fernández & T. Linzen (Eds.), *Proceedings of the 24th conference on computational natural language learning* (pp. 486–495). Online: Association for Computational Linguistics. 10.18653/v1/2020.conll-1.39

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Boguraev, S., Potts, C., & Mahowald, K. (2025). *Causal interventions reveal shared structure across english filler-gap constructions.* https://arxiv.org/abs/2505.16002

Bolhuis, J., Crain, S., Fong, S., & Moro, A. (2024, 21). Three reasons why AI doesn't model human language. *Nature*, *627*(8004), 489. 10.1038/d41586-024-00824-z

Butt, M., Dyvik, H., King, T. H., Masuichi, H., & Rohrer, C. (2002). The parallel grammar project. In *COLING-02: Grammar Engineering and Evaluation.* https://aclanthology.org/W02-1503/

Campbell, R., & Grieve, R. (1981). Royal Investigations of the Origin of Language. *Historiographia Linguistica*, *9*, 43-74. https://doi.org/10.1075/hl.9.1-2.04cam

Chater, N., Clark, A., Goldsmith, J. A., & Perfors, A. (2015). *Empiricism and Language Learnability* (First ed.). Oxford University Press.

Chater, N., Oaksford, M., Hahn, U., & Heit, E. (2010). Bayesian models of cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 811–823. https://doi.org/10.1002/wcs.79

Chaves, R. P. (2020). What don't RNN language models learn about filler-gap dependencies? In *Society for computation in linguistics* (Vol. 3, pp. 20–30). University of Massachusetts Amherst Libraries.

Chesi, C. (2024). *Is it the end of (generative) linguistics as we know it?* https://arxiv.org/abs/2412.12797

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

Chomsky, N. (1971). *Problems of knowledge and freedom: The Russell lectures.* Pantheon Books.

Chomsky, N. (1973). Conditions on transformations. In M. Halle, S. R. Anderson, & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). New York: Holt, Rinehart and Winston.

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

Chomsky, N. (1993). *Lectures on Government and Binding.* Berlin, New York: De Gruyter Mouton. https://doi.org/10.1515/9783110884166

Chomsky, N. (2001). Derivation by phase. In M. Kenstowicz (Ed.), *Ken Hale: A life in language* (pp. 1–52). Cambridge: The MIT Press.

Chomsky, N., Roberts, I., & Watumull, J. (2023). Noam Chomsky: The false promise of ChatGPT. *The New York Times*, *8*. https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgptai.html

Chowdhury, S. A., & Zamparelli, R. (2018). RNN Simulations of Grammaticality Judgments on Long-distance Dependencies. In E. M. Bender, L. Derczynski, & P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 133–144). Association for Computational Linguistics. https://aclanthology.org/C18-1012

Chowdhury, S. A., & Zamparelli, R. (2019). An LSTM adaptation study of (un) grammaticality. In T. Linzen, G. Chrupała, Y. Belinkov, & D. Hupkes (Eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 204–212). https://dx.doi.org/10.18653/v1/W19-4821

Christensen, K. K. (1982). On multiple filler-gap constructions in Norwegian. In E. Engdahl & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 77–98). Stockholm: Almquist & Wiksell.

Clark, A., & Lappin, S. (2011). *Linguistic Nativism and the Poverty of the Stimulus.* Chichester, West Sussex, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444390568

Crain, S., & Pietroski, P. (2001). Nature, nurture and universal grammar. *Linguistics and Philosophy*, *24*(2), 139–186. https://doi.org/10.1023/A:1005694100138

Culbertson, J., Smolensky, P., & Legendre, G. (2012). Learning biases predict a word order universal. *Cognition*, *122*(3), 306–329. https://doi.org/10.1016/j.cognition.2011.10.017

Cuneo, N., & Goldberg, A. E. (2023). The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition*, *240*, 105563. https://doi.org/10.1016/j.cognition.2023.105563

Cuskley, C., Woods, R., & Flaherty, M. (2024). The Limitations of Large Language Models for Understanding Human Language and Cognition. *Open Mind*, *8*, 1058–1083. https://doi.org/10.1162/opmi_a_00160

Da Costa, J. K., & Chaves, R. P. (2020). Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. *Proceedings of the Society for Computation in Linguistics*, *3*(1), 189–198.

De Santo, A. (2020). MG parsing as a model of gradient acceptability in syntactic islands. In *Proceedings of the society for computation in linguistics 2020* (pp. 59–69). https://doi.org/10.7275/srck-2j50

De Villiers, J., & Roeper, T. (1995). Relative clauses are barriers to wh-movement for young children. *Journal of child Language*, *22*(2), 389–404. https://doi.org/10.1017/S0305000900009843

De Villiers, J., Roeper, T., Bland-Stewart, L., & Pearson, B. (2008). Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics*, *29*(1), 67–103. https://doi.org/10.1017/S0142716408080041

De Villiers, J., Roeper, T., & Vainikka, A. (1990). The Acquisition of Long-Distance Rules. In L. Frazier & J. De Villiers (Eds.), *Language Processing and Language Acquisition* (pp. 257–297). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-3808-6_10

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. https://aclanthology.org/N19-1423/ 10.18653/v1/N19-1423

Dickson, N., Futrell, R., & Pearl, L. (2024). I forgot but it's okay: Learning about island constraints under child-like memory constraints. In H. A. A. AlThagafi & J. Ray (Eds.), *Proceedings of the 48th annual boston university conference on language development* (pp. 169–183). Somerville, MA: Cascadilla Press. https://www.lingref.com/bucld/48/BUCLD48-13.pdf

Dickson, N., Pearl, L., & Futrell, R. (2022). Learning constraints on wh-dependencies by learning how to efficiently represent wh-dependencies: A developmental modeling investigation with fragment grammars. *Proceedings of the Society for Computation in Linguistics*, *5*(1), 220–224. https://doi.org/10.7275/7fd4-fw49

Dunbar, E. (2019). Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language*, *95*(1), e87–e98. 10.1353/lan.2019.0013

Dyer, C., Kuncoro, A., Ballesteros, M., & Smith, N. A. (2016, June). Recurrent neural network grammars. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 conference of the north American chapter of the association for computational linguistics: Human language technologies* (pp. 199–209). San Diego, California: Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-1024

Dyvik, H. (2000). Nødvendige noder i norsk: Grunntrekk i en leksikalsk-funksjonell beskrivelse av norsk syntaks [Necessary nodes in Norwegian: Basic properties of a lexical-functional description of Norwegian syntax]. In Ø. Andersen, K. Fløttum, & T. Kinn (Eds.), *Menneske, språk og felleskap [human, language and community]* (pp. 25–45). Oslo: Novus forlag.

Dyvik, H., Meurer, P., Rosén, V., De Smedt, K., Haugereid, P., Losnegaard, G. S., … Thunes, M. (2016, May). NorGramBank: A 'deep' treebank for Norwegian. In N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 3555–3562). Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1565/

Elazar, Y., Ravfogel, S., Jacovi, A., & Goldberg, Y. (2021). Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, *9*, 160–175. https://doi.org/10.1162/tacl_a_00359

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development* (First MIT Press paperback edition ed.). Cambridge, MA: MIT Press.

Emanuilov, S. (2024, October). The transformer revolution. *UnfoldAI*. https://unfoldai.com/the-transformer-revolution/

Engdahl, E. (1982). Restrictions on unbounded dependencies in Swedish. In E. Engdahl & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 151–174). Stockholm: Almquist & Wiksell.

Erteschik-Shir, N. (1973). *On the nature of island constraints* (Unpublished doctoral dissertation). MIT.

Finlayson, M., Mueller, A., Gehrmann, S., Shieber, S., Linzen, T., & Belinkov, Y. (2021, August). Causal analysis of syntactic agreement mechanisms in neural language models. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1828–1843). Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.144

Frank, M. C. (2023a). Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, *27*(11), 990–992. https://doi.org/10.1016/j.tics.2023.08.007

Frank, M. C. (2023b). *Large language models as models of human cognition.* PsyArXiv. osf.io/preprints/psyarxiv/wxt69 10.31234/osf.io/wxt69

Freudenthal, D., Pine, J. M., Aguado-Orea, J., & Gobet, F. (2007). Modeling the Developmental Patterning of Finiteness Marking in English, Dutch, German, and Spanish Using MOSAIC. *Cognitive Science*, *31*(2), 311–341. https://doi.org/10.1080/15326900701221454

Freudenthal, D., Pine, J. M., & Gobet, F. (2006). Modeling the Development of Children's Use of Optional Infinitives in Dutch and English Using MOSAIC. *Cognitive Science*, *30*(2), 277–310. https://doi.org/10.1207/s15516709cog0000_47

Futrell, R., Gibson, E., & Levy, R. P. (2020). Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science*, *44*(3), e12814. https://doi.org/10.1111/cogs.12814

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models. *arXiv*. https://arxiv.org/abs/2501.17047

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019). Neural language models as psycholinguistic subjects: Representations of syntactic state. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 32–42). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

Gagliardi, A., Mease, T. M., & and, J. L. (2016). Discontinuous development in the acquisition of filler-gap dependencies: Evidence from 15- and 20-month-olds. *Language Acquisition*, *23*(3), 234–260. https://doi.org/10.1080/10489223.2015.1115048

Garmann, N. G., Hansen, P., Simonsen, H. G., & Kristoffersen, K. E. (2019). The phonology of children's early words: trends, individual variation and parents' accommodation in child-directed speech. *Frontiers in Communication*, *4*, 10. 10.3389/fcomm.2019.00010

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological stm in the development of vocabulary in children: A longitudinal study. *Journal of memory and language*, *28*(2), 200–213.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., ... Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*(2), 248-265. 10.1044/2016_AJSLP-15-0169

Goldberg, A. (1995). *Constructions: A Construction Grammar Approach to Argument Structure.* Chicago: The University Press.

Goldberg, A. (2006). *Constructions at work: The nature of generalization in language.* Oxford University Press.

Goldberg, A. (2024). Usage-based constructionist approaches and large language models. *Frames and Constructions*(16), 220–254. https://doi.org/10.1075/cf.23017.gol

Goodluck, H., Foley, M., & Sedivy, J. (1992). Adjunct Islands and Acquisition. In H. Goodluck & M. Rochemont (Eds.), *Island Constraints: Theory, Acquisition and Processing* (pp. 181–194). Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-017-1980-3_6

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357-364. https://doi.org/10.1016/j.tics.2010.05.004

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. Cambridge University Press.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018* (pp. 1195–1205). https://doi.org/10.18653/v1/N18-1108

Gupta, A., Kvernadze, G., & Srikumar, V. (2021, May). BERT & family eat word salad: Experiments with text understanding. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(14), 12946-12954. https://doi.org/10.1609/aaai.v35i14.17531

Haegeman, L. (1995). Root infinitives, tense, and truncated structures in dutch. *Language acquisition*, *4*(3), 205–255.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second meeting of the North American chapter of the Association for Computational Linguistics* (pp. 1–8). https://doi.org/10.3115/1073336.1073357

Hao, S. (2025). *Generative linguistics, large language models, and the social nature of scientific success.* https://arxiv.org/abs/2503.20088

Hart, B., & Risley, T. R. (1992). American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. *Developmental psychology*, *28*(6), 1096.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415. https://doi.org/10.1353/lan.0.0223

Howitt, K., Nair, S., Dods, A., & Hopkins, R. M. (2024). Generalizations across filler-gap dependencies in neural language models. In L. Barak & M. Alikhani (Eds.), *Proceedings of the 28th conference on computational natural language learning* (pp. 269–279). Association for Computational

Linguistics. https://aclanthology.org/2024.conll-1.21

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1725–1744). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Huang, C. T. J. (1982). *Logical relations in Chinese and the theory of grammar* (Unpublished doctoral dissertation). MIT.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510. https://doi.org/10.1016/j.jml.2024.104510

Jones, G., Gobet, F., Freudenthal, D., Watson, S. E., & Pine, J. M. (2014). Why computational models are better than verbal theories: The case of nonword repetition. *Developmental Science*, *17*(2), 298–310. https://doi.org/10.1111/desc.12111

Jurafsky, D., & Martin, J. H. (2024). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models.* (3rd ed.). https://web.stanford.edu/~jurafsky/slp3/ (Online manuscript released August 20, 2024)

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive science*, *32*(4), 771–787.

Kaplan, R. M., & Bresnan, J. (1995). Formal system for grammatical representation. *Formal issues in lexical-functional grammar*(47), 29.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics*, *17*, 1–12. https://doi.org/10.5964/bioling.13153

Kluender, R., & Kutas, M. (1993). Subjacency as a processing phenomenon. *Language and cognitive processes*, *8*(4), 573–633.

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2022). LSTMs Can Learn Basic Wh-and Relative Clause Dependencies in Norwegian. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 2974–2980). https://escholarship.org/uc/item/012683gb

Kobzeva, A., & Kush, D. (2024). Grammar and Expectation in Active Dependency Resolution: Experimental and Modeling Evidence from Norwegian. *Cognitive Science*, *48*(10), e13501. https://doi.org/10.1111/cogs.13501

Kobzeva, A., & Kush, D. (2025). Acquiring constraints on filler-gap dependencies from structural collocations: Assessing a computational learning model of island-insensitivity in Norwegian. *Language Acquisition*, 1–44. https://doi.org/10.1080/10489223.2024.2440340

Kobzeva, A., Sant, C., Robbins, P. T., Vos, M., Lohndal, T., & Kush, D. (2022). Comparing island effects for different dependency types in Norwegian. *Languages*, *7*(3), 195–220. https://doi.org/10.3390/languages7030197

Kodner, J., Payne, S., & Heinz, J. (2023). Why linguistics will thrive in the 21st century: A reply to piantadosi (2023). *arXiv preprint arXiv:2308.03228*. 10.48550/arXiv.2308.03228

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical learning across development: Flexible yet constrained. *Frontiers in Psychology*, *3*, 598. https://doi.org/10.3389/fpsyg.2012.00598

Kuhn, T. S. (1962). *The structure of scientific revolutions.* Chicago, IL, USA: University of Chicago Press.

Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, *36*(3), 743–779. https://doi.org/10.1007/s11049-017-9390-z

Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, *95*(3), 393–420. https://doi.org/10.1353/lan.2019.0051

Kush, D., Sant, C., & Strætkvern, S. B. (2021). Learning Island-insensitivity from the Input: A Corpus Analysis of Child- and Youth-Directed Text in Norwegian. *Glossa: A journal of general linguistics*, *6*(1), 1–50. https://doi.org/10.16995/glossa.5774

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019, June). The emergence of number and syntax units in LSTM language models. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 11–20). Minneapolis, Minnesota: Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1002

Lan, N., Chemla, E., & Katzir, R. (2024). Large Language Models and the Argument from the Poverty of the Stimulus. *Linguistic Inquiry*, 1–56. https://doi.org/10.1162/ling_a_00533

Lappin, S., & Shieber, S. M. (2007). Machine learning theory and practice as a source of insight into universal grammar. *Journal of Linguistics*, *43*(2), 393–427. https://doi.org/10.1017/S0022226707004628

Lau, J. H., Clark, A., & Lappin, S. (2016). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202-1241. https://doi.org/10.1111/cogs.12414

Laurence, S., & Margolis, E. (2001). The poverty of the stimulus argument. *British Journal for the Philosophy of Science*, *52*(2). https://doi.org/10.1093/bjps/52.2.217

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *Journal of Psycholinguistic Research*, *33*, 51–73. 10.1023/b:jopr.0000010514.50468.30

Legate, J. A., & Yang, C. D. (2002). Empirical re-assessment of stimulus poverty arguments. *The Linguistic Review*, *19*(1-2), 151–162. https://doi.org/10.1515/tlir.19.1-2.151

Legate, J. A., & Yang, C. D. (2007). Morphosyntactic learning and the development of tense. *Language Acquisition*, *14*(3), 315–344.

Leong, C. S.-Y., & Linzen, T. (2024). Testing learning hypotheses using neural networks by manipulating learning data. *arXiv*. https://doi.org/10.48550/arXiv.2407.04593

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006

Lidz, J., & Gagliardi, A. (2015). How Nature Meets Nurture: Universal Grammar and Statistical Learning. *Annual Review of Linguistics*, *1*, 333–353. https://doi.org/10.1146/annurev-linguist-030514-125236

Lidz, J., Waxman, S., & Freedman, J. (2003). What infants know about syntax but couldn't have learned: Experimental evidence for syntactic structure at 18 months. *Cognition*, *89*(3), 295–303. https://doi.org/10.1016/S0010-0277(03)00116-1

Lightfoot, D. (1991). *How to set parameters: Arguments from language change*. The MIT Press.

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). *A survey of transformers.* https://doi.org/10.48550/arXiv.2106.04554

Linzen, T. (2019). What can linguistics and deep learning contribute to each other? Response to Pater. *Language*, *95*(1), e99–e108. 10.1353/lan.2019.0015.

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics*, *7*(1), 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for*

Computational Linguistics, *4*, 521–535.

Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2022). A verb-frame frequency account of constraints on long-distance dependencies in English . *Cognition*, *222*, 104902. https://doi.org/10.1016/j.cognition.2021.104902

Liu, Y., Winckel, E., Abeillé, A., Hemforth, B., & Gibson, E. (2022). Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics*, *8*, 495–525. https://doi.org/10.1146/annurev-linguistics-011619-030319

MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2). Psychology Press.

Mahowald, K., Diachek, E., Gibson, E., Fedorenko, E., & Futrell, R. (2023). Grammatical cues to subjecthood are redundant in a majority of simple clauses across languages. *Cognition*, *241*, 105543. https://doi.org/10.1016/j.cognition.2023.105543

Matthews, D. (2006). Epistemic humility. In J. P. van Gigch (Ed.), *Wisdom, knowledge, and management: A critique and analysis of churchman's systems approach* (pp. 105–137). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-36506-0_7

McCoy, R. T., Frank, R., & Linzen, T. (2020). Does Syntax Need to Grow on Trees? Sources of Hierarchical Inductive Bias in Sequence-to-Sequence Networks. In *Transactions of the Association for Computational Linguistics* (Vol. 8, pp. 125–140). MIT Press. https://doi.org/10.1162/tacl_a_00304

McCoy, R. T., & Griffiths, T. L. (2025). Modeling rapid language learning by distilling bayesian priors into artificial neural networks. *Nature Communications*, *16*, 4676. https://doi.org/10.1038/s41467-025-59957-y

McCoy, R. T., Min, J., & Linzen, T. (2020). BERTs of a feather do not generalize together: Large variability in generalization across models with similar test set performance. In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the third blackboxnlp workshop on analyzing and interpreting neural networks for nlp* (pp. 217–227). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.blackboxnlp-1.21

McCoy, R. T., Pavlick, E., & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 3428–3448). https://doi.org/10.18653/v1/P19-1334

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, *121*(41), e2322420121. https://doi.org/10.1073/pnas.2322420121

Michaelov, J. A., Bardolph, M. D., Van Petten, C. K., Bergen, B. K., & Coulson, S. (2024). Strong Prediction: Language Model Surprisal Explains Multiple N400 Effects. *Neurobiology of Language*, *5*(1), 107–135. https://doi.org/10.1162/nol_a_00105

Mita, M., Yoshida, R., & Oseki, Y. (2025). *Developmentally-plausible working memory shapes a critical period for language acquisition.* https://arxiv.org/abs/2502.04795

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, *167*, 82-85. https://doi.org/10.1016/j.cortex.2023.07.003

Mueller, A., & Linzen, T. (2023). How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases . *arXiv preprint arXiv:2305.19905*.

Newmeyer, F. J. (2016). Nonsyntactic explanations of island constraints. *Annual Review of Linguistics*, *2*(Volume 2, 2016), 187-210. https://doi.org/10.1146/annurev-linguistics-011415-040707

Noble, C. H., Cameron-Faulkner, T., & Lieven, E. (2018). Keeping it simple: The grammatical properties of shared book reading. *Journal of Child Language*, *45*(3), 753–766.

Nygaard, L. P., & Solli, K. (2020). *Strategies for writing a thesis by publication in the social sciences and humanities*. Routledge. https://doi.org/10.4324/9780429261671

O'Connor, J., & Andreas, J. (2021). What context features can transformer language models use? In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (pp. 851–864). Association for Computational Linguistics. 10.18653/v1/2021.acl-long.70

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology*, *6*, 384. https://doi.org/10.3389/fpsyg.2015.00384

Omaki, A., & Lidz, J. (2015). Linking parser development to acquisition of syntactic knowledge. *Language Acquisition*, *22*(2), 158–192. https://doi.org/10.1080/10489223.2014.943903

Ozaki, S., Yurovsky, D., & Levin, L. (2022, February). How well do LSTM language models learn filler-gap dependencies? In A. Ettinger, T. Hunter, & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics 2022* (pp. 76–88). online: Association for Computational Linguistics. https://aclanthology.org/2022.scil-1.6/

Pannitto, L., & Herbelot, A. (2022). Can recurrent neural networks validate usage-based theories of grammar acquisition? *Frontiers in Psychology*, *13*, 741321. https://doi.org/10.3389/fpsyg.2022.741321

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language*, *95*(1), e41–e74. https://doi.org/10.1353/lan.2019.0009

Patil, A., Jumelet, J., Chiu, Y. Y., Lapastora, Wang, L., Willrich, C., & Steinert-Threlkeld, S. (2024). Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. *arXiv preprint arXiv:2405.15750*. 10.48550/arXiv.2405.15750

Pearl, L. (2008). Putting the emphasis on unambiguous: The feasibility of data filtering for learning English metrical phonology. In *BUCLD 32: Proceedings of the 32nd annual Boston University Conference on Child Language Development* (Vol. 32, pp. 390–401).

Pearl, L. (2019, 01). Fusion is great, and interpretable fusion could be exciting for theory generation: Response to Pater. *Language*, *95*, e109-e114. 10.1353/lan.2019.0017

Pearl, L. (2022). Poverty of the stimulus without tears. *Language Learning and Development*, *18*(4), 415–454. https://doi.org/10.1080/15475441.2021.1981908

Pearl, L. (2023a). Computational cognitive modeling for syntactic acquisition: Approaches that integrate information from multiple places. *Journal of Child Language*, *50*(6), 1353–1373. https://doi.org/10.1017/S0305000923000247

Pearl, L. (2023b). Modeling syntactic acquisition. In J. Sprouse (Ed.), *The Oxford Handbook of Experimental Syntax* (pp. 209–270). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198797722.013.8

Pearl, L., & Bates, A. (2022). A new way to identify if variation in children's input could be developmentally meaningful: Using computational cognitive modeling to assess input across socio-economic status for syntactic islands. *Journal of Child Language*, 1–34. https://doi.org/10.1017/S0305000922000514

Pearl, L., & Sprouse, J. (2013a). Computational models of acquisition for islands. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 109–131). Cambridge University Press. https://doi.org/10.1017/CBO9781139035309.006

Pearl, L., & Sprouse, J. (2013b). Syntactic Islands and Learning Biases: Combining Experimental Syntax and Computational Modeling to Investigate the Language Acquisition Problem. *Language Acquisition*, *20*(1), 23–68. https://doi.org/10.1080/10489223.2012.738742

Pearl, L., & Sprouse, J. (2015). Computational modeling for language acquisition: A tutorial with syntactic islands. *Journal of Speech, Language, and Hearing Research*, *58*(3), 740–753. https://doi.org/10.1044/2015_JSLHR-L-14-0362

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, *118*(3), 306–338. https://doi.org/10.1016/j.cognition.2010.11.001

Perkins, L., & Lidz, J. L. (2023). Behavioral Acquisition Methods With Infants. In J. Sprouse (Ed.), *The Oxford Handbook of Experimental Syntax* (pp. 137–170). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198797722.013.6

Phillips, C. (2006). The real-time status of island phenomena. *Language*, *82*(4), 795–823. https://doi.org/10.1353/lan.2006.0217

Phillips, C. (2013a). On the nature of island constraints II: Language learning and innateness. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 132–158). Cambridge University Press. https://doi.org/10.1017/CBO9781139035309.007

Phillips, C. (2013b). On the nature of island constraints I: Language processing and reductionistaccounts. In J. Sprouse & N. Hornstein (Eds.), *Experimental syntax and island effects* (pp. 64–108). Cambridge University Press. https://doi.org/10.1017/CBO9781139035309.005

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. In E. Gibson & M. Poliak (Eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (pp. 353–414).

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. University of Chicago Press.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The linguistic review*, *19*(1-2), 9–50. https://doi.org/10.1515/tlir.19.1-2.9

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training* (Tech. Rep.). OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Ramchand, G. (2024). *On LLMs, generative grammar, and how we need theory more than ever.* https://lingbuzz.net/lingbuzz/008643. (Preprint, LingBuzz)

Rawski, J., & Heinz, J. (2019). No free lunch in linguistics or machine learning: Response to pater. *Language*, *95*(1), e125–e135. 10.1353/lan.2019.0021

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science*, *29*(6), 1007–1028. https://doi.org/10.1207/s15516709cog0000_28

Reinhart, T. (1981). A second COMP position. In A. Belletti, L. Brandi, & L. Rizzi (Eds.), *Theory of Markedness in Generative Grammar* (pp. 518–557). Pisa: Scuola Normale Superiore.

Rizzi, L. (1982). *Issues in Italian Syntax*. De Gruyter Mouton. https://doi.org/10.1515/9783110883718.49

Ross, J. R. (1967). *Constraints on variables in syntax* (Doctoral dissertation, MIT). https://dspace.mit.edu/handle/1721.1/15166

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In J. Ronald, D. Rumelhart, G. Hinton, E. D. Rumelhart, & J. McClelland (Eds.), *Parallel distributed processing, volume 2*. MIT press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of english verbs. In J. L. McClelland, D. E. Rumelhart, & P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., & Levy, R. (2024). Large-scale evidence for logarithmic effects of word predictability on reading time. *Proceedings of the National Academy of Sciences*, *121*(10), e2307876121. https://doi.org/10.1073/pnas.2307876121

Slaats, S., & Martin, A. E. (2025). What's surprising about surprisal. *Computational Brain and Behavior*, *8*, 233–248. https://doi.org/10.1007/s42113-025-00237-9

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319. https://doi.org/10.1016/j.cognition.2013.02.013

Sprouse, J. (2007). *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge* (PhD dissertation, University of Maryland, College Park). Retrieved 2022-01-08, from https://drum.lib.umd.edu/handle/1903/7283

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 82–123.

Staub, A. (2025). Predictability in language comprehension: Prospects and problems for surprisal. *Annual Review of Linguistics*, *11*, 17-34. https://doi.org/10.1146/annurev-linguistics-011724-121517

Stepanov, A. (2007). The end of CED? Minimalism and extraction domains.

Syntax, *10*(1), 80–126.

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, *1*(3), 227–245. https://doi.org/10.1080/01690968608407062

Stromswold, K. (1995). The acquisition of subject and object wh-questions. *Language acquisition*, *4*(1-2), 5–48. https://doi.org/10.1080/10489223.1995.9671658

Suijkerbuijk, M., de Swart, P., & Frank, S. L. (2023). The Learnability of the Wh-Island Constraint in Dutch by a Long Short-Term Memory Network. In T. Hunter & B. Prickett (Eds.), *Proceedings of the society for computation in linguistics 2023* (pp. 321–331). Amherst, MA: Association for Computational Linguistics. https://aclanthology.org/2023.scil-1.28/

Syrett, K. (2023). Behavioral Acquisition Methods With Preschool-Age Children. In J. Sprouse (Ed.), *The Oxford Handbook of Experimental Syntax* (pp. 171–208). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198797722.013.7

Szabolcsi, A. (2006). Strong vs. Weak Islands. In *The Blackwell Companion to Syntax* (p. 479-531). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470996591.ch64

Szabolcsi, A., & Lohndal, T. (2017). Strong vs. Weak Islands. In *The Wiley Blackwell Companion to Syntax* (2nd ed.). John Wiley & Sons. https://doi.org/10.1002/9781118358733.wbsyncom008

Tekstlaboratoriet. (2005). *Nota - norsk talespråkskorpus [norwegian speech corpus].* https://tekstlab.uio.no/nota/english/index.html. (Transcribed speech corpus from Oslo.)

Tekstlaboratoriet. (2009). *The bigbrother corpus.* http://www.tekstlab.uio.no/nota/bigbrother/english.html. (Speech corpus from the Norwegian Big Brother TV show.)

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475. https://doi.org/10.1006/jmla.1996.0025

Valian, V. (2009). Innateness and learnability. In *Handbook of child language* (pp. 15–34). Cambridge University Press.

Valin Jr, R. D. V. (1998). The Acquisition of WH-Questions and the Mechanisms of Language Acquisition. In M. Tomasello (Ed.), *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure* (pp. 221–249). Routledge. https://doi.org/10.4324/9781315085678-9

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988. https://doi.org/10.1111/cogs.12988

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). Attention is all you need. *Conference on Neural Information Processing Systems (NIPS 2017)*, *30*.

Villata, S., & Tabor, W. (2022). A self-organized sentence processing theory of gradience: The case of islands. *Cognition*, *222*, 104943. https://doi.org/10.1016/j.cognition.2021.104943

Vázquez Martínez, H. J., Heuser, A. L., Yang, C., & Kodner, J. (2024). Evaluating the existence proof: LLMs as cognitive models of language acquisition. In J.-L. Mendivil-Giro (Ed.), *Artificial knowledge of language.* Vernon Press. https://lingbuzz.net/lingbuzz/008277

Wagers, M., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, *45*(2), 395–433. https://doi.org/10.1017/S0022226709005726

Warstadt, A., & Bowman, S. R. (2022). What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language* (pp. 17–60). CRC Press.

Warstadt, A., Cao, Y., Grosu, I., Peng, W., Blix, H., Nie, Y., . . . others (2019). Investigating BERT's knowledge of language: Five analysis methods with NPIs. *arXiv.* https://doi.org/10.48550/arXiv.1909.02597

Warstadt, A., Mueller, A., Choshen, L., Wilcox, Ciro, J., Mosquera, R., . . . others (2023). Findings of the BabyLM Challenge: Sample-efficient pretraining on developmentally plausible corpora . In *Proceedings of the babylm challenge at the 27th conference on computational natural language learning* (pp. 1–34). 10.3929/ethz-b-000650680

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, *8*, 377–392. https://doi.org/10.1162/tacl_a_00321

Werbos, P. J. (1990). Backpropagation through time: What it does and how to do it. *Proceedings of the IEEE*, *78*(10), 1550–1560.

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivations in child language. In D. Lightfoot & N. Hornstein (Eds.), *Verb movement* (pp. 305–350). Cambridge University Press.

Wilcox, E., Futrell, R., & Levy, R. (2023, apr). Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, 1–44. https://doi.org/10.1162/ling_a_00491

Wilcox, E., Levy, R., & Futrell, R. (2019a). Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP* (pp. 181–190). https://doi.org/10.18653/v1/W19-4819

Wilcox, E., Levy, R., & Futrell, R. (2019b). What Syntactic Structures block Dependencies in RNN Language Models? In *Annual Meeting of the Cognitive Science Society*. https://api.semanticscholar.org/CorpusID:166228146

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP* (pp. 211–221). https://doi.org/10.18653/v1/W18-5423

Wilcox, E., Meister, C., Cotterell, R., & Pimentel, T. (2023). Language model quality correlates with psychometric predictive power in multiple languages. In *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 7503–7511). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.emnlp-main.466

Yang, C. D. (2002). *Knowledge and learning in natural language*. Oxford University Press, USA.

Yang, C. D. (2008). The great number crunch. *Journal of Linguistics*, *44*(1), 205–228. 10.1017/S0022226707004999

Yedetore, A., Linzen, T., Frank, R., & McCoy, R. T. (2023). How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 9370–9393). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.521

Ziv, I., Lan, N., Chemla, E., & Katzir, R. (2025). *Large language models as proxies for theories of human linguistic cognition*. https://arxiv.org/abs/2502.07687

Part II

PUBLICATIONS

Article C1

Kobzeva, A., Arehalli, S., Linzen, T. & Kush, D. (2022). LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian. In J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni (Eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 44, pp. 2974–2980). URL: https://escholarship.org/uc/item/012683gb

LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian

Anastasia Kobzeva (anastasia.kobzeva@ntnu.no)
NTNU, Trondheim, Norway

Suhas Arehalli (suhas@jhu.edu)
Johns Hopkins University, Baltimore, MD, USA

Tal Linzen (linzen@nyu.edu)
New York University, New York, NY, USA

Dave Kush (dave.kush@utoronto.ca)
NTNU and University of Toronto, Toronto, ON, Canada

Abstract

One of the key features of natural languages is that they exhibit long-distance *filler-gap dependencies* (FGDs): In the sentence 'What do you think the pilot sent __?' the *wh*-filler *what* is interpreted as the object of the verb *sent* across multiple words. The ability to establish FGDs is thought to require hierarchical syntactic structure. However, recent research suggests that recurrent neural networks (RNNs) without specific hierarchical bias can learn complex generalizations about *wh*-questions in English from raw text data (Wilcox et al., 2018, 2019). Across two experiments, we probe the generality of this result by testing whether a long short-term memory (LSTM) RNN model can learn basic generalizations about FGDs in Norwegian. Testing Norwegian allows us to assess whether previous results were due to distributional statistics of the English input or whether models can extract similar generalizations in languages with different syntactic distributions. We also test the model's performance on two different types of FGDs: *wh*-questions and relative clauses, allowing us to determine if the model learns abstract generalizations about FGDs that extend beyond a single construction type. Results from Experiment 1 suggest that the model expects fillers to be paired with gaps and that this expectation generalizes across different syntactic positions. Results from Experiment 2 suggest that the model's expectations are largely unaffected by the increased linear distance between the filler and the gap. Our findings provide support for the conclusion that LSTM RNN's ability to learn basic generalizations about FGDs is robust across dependency type and language.

Keywords: Filler-Gap Dependencies, Neural Language Models, Norwegian, Relative Clauses, Embedded Questions

Introduction

Natural languages exhibit Filler-Gap Dependencies (FGDs) in which *filler* phrases are interpreted at later *gap* positions. Embedded questions like (1) are a type of FGD: the *wh*-filler *what* is interpreted as though it occupied the gap in the direct object position of the verb *sent* (marked with an underscore). Relative clauses (RCs) like (2) are also FGDs that include a relative pronoun (*that*) or a null operator as the filler and the head of the RC (*the present*), which is interpreted in the gap position.

(1) I know **what** the pilot sent __ to his family.

(2) I heard about **the present** that the pilot sent __ to his family.

Establishing an FGD requires abstract generalizations and representations. The well-formedness of an FGD must be described in terms of syntactic relations between the filler, the gap, and other elements in a hierarchical syntactic structure. FGDs are also potentially unbounded in length (3), which suggests that they cannot be adequately described in terms of linear predictability.

(3) I know **what** the guy from the airport said Mary saw that the pilot sent __ to his family after landing.

Despite the fact that FGDs require abstract generalizations over hierarchical representations, recent findings by Wilcox and colleagues (2018, 2019) suggest that Recurrent Neural Networks (RNNs, Elman, 1990), which are inherently sequence models without built-in biases for representing hierarchical structure, can learn FGDs and associated constraints on them. Specifically, the authors argue that Long Short-Term Memory (LSTM) RNNs (Hochreiter & Schmidhuber, 1997) that are trained with a generic language modeling objective on unannotated English text implicitly learn the distribution of acceptable FGDs in English. Their results indicate that the LSTMs could represent dependencies between fillers and gaps in multiple syntactic positions, maintain this relationship over large spans of text, and even obey complex constraints that govern where FGDs cannot be established. These results go in line with previous studies where LSTMs showed impressive results on linguistic processing tasks that require structurally-mediated dependencies, such as subject-verb agreement (Linzen et al., 2016; Gulordava et al., 2018) or auxiliary inversion (McCoy et al., 2018).

The results of Wilcox and colleagues (2018, 2019) are intriguing, but our ability to draw strong conclusions from this work about the general ability of LSTM RNNs is limited by the scope of previous experiments. First, past experiments have only investigated FGDs in English, which leaves open the question of whether the models could achieve similar success on input from languages with different distributional characteristics. Second, previous experiments only investigated one type of FGDs: *wh*-questions. It is unclear whether the success of past models should be attributed to idiosyncratic properties of (the distribution of) *wh*-questions or to a general ability of LSTMs to learn abstract generalizations about FGDs of any type.

We address this gap by exploring the ability of LSTM models to learn two types of FGDs in Norwegian: *wh*-questions and relative clauses. Norwegian is like English in that it permits FGDs across various syntactic positions, which facilitates close comparison. However, the morphosyntax of Nor-

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

wegian differs from English in a number of respects, such that the distribution of cues to syntactic structure varies between the languages. For example, Norwegian is a V2 language that makes extensive use of fronting, which means that the mapping from surface word order to grammatical role is sometimes less obvious than in English. Norwegian also lacks morphological cues that might help learn syntactic dependencies, such as subject-verb agreement.

Testing RC dependencies in addition to *wh*-questions can also shed light on how abstract or general the LSTM's representations of FGDs are by testing whether success depends on specific overt lexical contingencies. *Wh*-words provide relatively unambiguous, superficial cues to the presence of a later gap. In some RCs, however, the cues are superficially ambiguous. In English, RCs can be introduced by the complementizer *that*, as seen in (2). But the complementizer *that* is also used in declarative complement clauses, where it does not license a gap (4). It also has other uses (e.g., determiner).

(4) I heard **that** the pilot sent the present to his family.

In Norwegian, the relative pronoun *som* is used in RCs as in (5). Similar to relative pronouns in English, *som* is ambiguous: it can be used as a comparative operator as in *Han er like høy som meg* 'He is as tall as me'.

(5) Jeg hørte om gaven som piloten sendte __ til
I heard about present.DEF REL pilot.DEF sent __ to
familien sin etter landing
family.DEF his after landing
'I heard about the present that the pilot sent to his family after landing.'

Such superficially ambiguous cues to the presence of a gap could potentially hinder (Gulordava et al., 2018) or improve the model's performance (Kam et al., 2008) on recognizing FGDs.

We now turn to our experiments. Experiment 1 explored whether an LSTM model can learn that fillers can be associated with gaps in different syntactic positions. Experiment 2 tested whether the model's representation of FGDs is robust to intervening material by manipulating the linear distance between the filler and the gap. To preview our results, we find that the model can represent both *wh*- and RC FGDs across different syntactic positions and can represent the FGDs across intervening material.

Methods

Language models

We trained an LSTM RNN with a language modeling objective. Such language models take a sequence of words as an input, transform it into a vector, and predict the most probable next word in that sequence using a softmax classifier over the model's vocabulary. Our model was trained on 113 million tokens of Norwegian Bokmål Wikipedia dump (Bokmål is one of the two written standards of Norwegian). Following (Gulordava et al., 2018), the model was a 2-layer LSTM with 650 hidden units in each layer and a vocabulary size of most frequent 50 000 tokens. It was trained for 40 epochs and achieved a perplexity of 30.4 on the validation set. We also trained a 5-gram model - a simple statistical model that can represent local dependencies between words within a 5-words window. This model was trained on the same corpus with Knesser-Ney smoothing and achieved a perplexity of 133.5 on the validation set. We primarily use this model as a baseline model.

Dependent variable

We investigate the model's syntactic generalizations about FGDs by looking at *surprisal*, which is the inverse log probability that the model assigns to a word given the previous context. Surprisal shows to what extent a word is unexpected given the model's probability distribution. Surprisal has been shown to correlate with incremental processing difficulty during human sentence processing (Hale, 2001; Levy, 2008).

Measuring filler-gap dependencies

Following Wilcox and colleagues (2018), we created our experimental items using a 2x2 factorial design that manipulated the presence of a filler and the presence of a gap in a sentence as in (6).

(6) She knows...
a. that the priest revealed the secret -FILLER, -GAP
b. *that the priest revealed __ -FILLER, +GAP
c. *what the priest revealed the secret +FILLER, -GAP
d. what the priest revealed __ +FILLER, +GAP
...in front of the guests at the party.

According to this factorial design, there should be an interaction between the presence of a filler and the presence of a gap, such that grammatical sentences with either no FGD (6-a) or a licensed FGD (6-d), should have lower surprisal values compared to ungrammatical sentences that contain an unlicensed gap (6-b), or a filler with no gap (6-c). To test for an interaction, we ran linear mixed-effects regression models with surprisal as a response variable, sum-coded conditions as predictors, and by-item random slopes (Barr et al., 2013).

When presenting our experimental results, we will collapse across two out of the four conditions by looking at pairwise differences between +FILLER and -FILLER conditions, which we call *filler effects*. There are two separate filler effects: a *filled gap effect* (Stowe, 1986, -GAP conditions) and an *unlicensed gap effect* (+GAP conditions).

The filled gap effect provides a measure of whether the presence of the filler triggers an expectation for an upcoming gap (in the earliest possible position). A filled gap effect is measured by comparing surprisal at NPs in the grammatical -FILLER, -GAP condition to the same NPs in the corresponding +FILLER, -GAP condition. If the model expects a gap after seeing a filler, it should assign a higher surprisal value to an NP in a potential gap position than it assigns to the same NP in a sentence without a filler (e.g., compare surprisal values at *the secret* in (6-c) v. (6-a)). Filled gap effects should manifest as positive differences in surprisal.

The unlicensed gap effect measures how 'surprised' the model is to find a gap in a sentence without a filler. The effect is calculated by comparing surprisal in the immediate post-gap region (i.e. *in front of* in (6)) in the +FILLER +GAP and -FILLER +GAP conditions. If the model knows that gaps must be licensed by a filler, surprisal in the post-gap region should be lower in +FILLER sentences than in -FILLER sentences. The unlicensed gap effect (the surprisal difference between conditions (6-d)-(6-b)) should be negative in such cases.

Experiment 1: Flexibility of filler-gap licensing

In Experiment 1 we test whether the models learn that fillers can license gaps in different syntactic positions. Following Wilcox and colleagues' methodology (2018), we tested both *wh*- and RC FGDs with gaps in subject, direct object, and oblique (complement of a prepositional phrase) positions in Norwegian. We present the materials and the results for each dependency type in turn.

Wh-dependencies

We created 20 test items according to a factorial design that crossed the 2x2 design in (6) with a factor that manipulated whether the gap was in subject, direct object, or oblique position as in (7), resulting in 12 conditions and 240 test sentences. Verbs were either ditransitive or transitive and accompanied by a prepositional phrase that could host a gap in oblique sentences. When the gap occurs in direct or oblique object positions in Norwegian, the structure of the sentence is the same as in English (7-b), (7-c). However, when the gap is in subject position, an expletive relative pronoun *som* is required in front of the gap in Norwegian (7-a), which could serve as an additional cue to the model for identifying the FGD.

(7) Hun vet... 'She knows...'

a. SUBJECT GAP

hvem som __ avslørte hemmeligheten foran
who REL __ revealed secret.DEF in front of
gjestene på festen
guests.DEF at party.DEF

'who revealed the secret in front of the guests at the party.'

b. DIRECT OBJECT GAP

hva presten avslørte __ foran gjestene på
what priest.DEF revealed __ in front of guests.DEF at
festen
party.DEF

'what the priest revealed __ in front of the guests at the party.'

c. OBLIQUE GAP

hvem presten avslørte hemmeligheten foran
who priest.DEF revealed secret.DEF in front of
__ på festen
__ at party.DEF

'who the priest revealed the secret in front of __ at the party.'

Figure 1 shows filler effects (differences between +FILLER and -FILLER conditions) measured in bits of surprisal (on the y-axis) by sentence region and gap position. Filled gap effects are measured at argument NPs in -GAP conditions (orange lines). Unlicensed gap effects are measured in the regions immediately following the gap for +GAP conditions (blue lines). Figure 2 compares the filled gap and unlicensed gap effects at each region of interest from the LSTM model to the baseline 5-gram model.

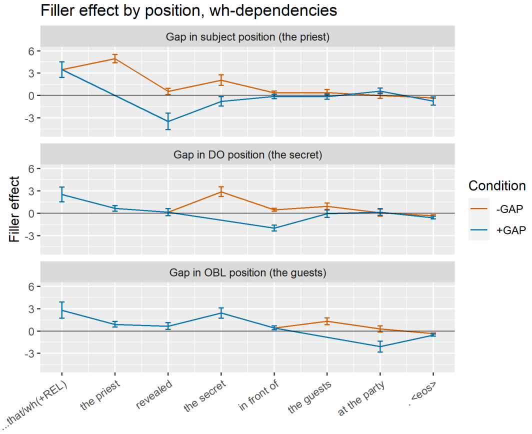


Figure 1: Filler effects for *wh*-dependencies by sentence region and gap position. Region labels are given in English for presentation purposes. Error bars are 95% confidence intervals across test items.

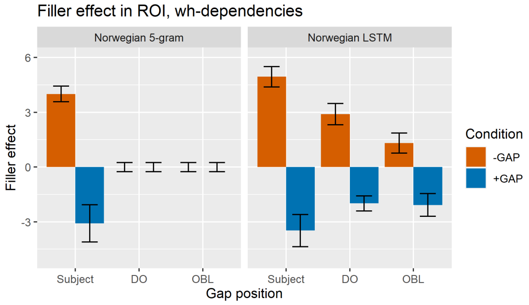


Figure 2: Filler effect for *wh*-dependencies by position.

Visual inspection of the figures suggests that the LSTM exhibits filled gap effects at all three argument positions, as evidenced by the positive surprisal differences at *the priest*, *the secret*, and *the guests*. It also appears that filled gap effects persist throughout the sentence if a gap has not been identified in an earlier position: filled gap effects are observed at the DO region after a filled subject position (top panel Figure 1) and in OBL position after a filled DO position (middle panel Figure 1). These results suggest that the model behaves like an active parser, positing a gap at every possible site after encountering a filler in the preceding context (although the effect is notably smaller in the positions following the first filled

NP position). This could be interpreted as evidence that the presence of a filler sets up an expectation for a gap in general, not in a particular syntactic position. Figure 2 also shows that the size of the filled gap effect varies by position, with subject positions inducing the largest filled gap effects, followed by direct object, and then oblique position. The baseline 5-gram model showed a filled gap effect in subject position, but nowhere else.

The model also appears to recognize unlicensed gaps in subject, DO and OBL position, as evidenced by the negative surprisal differences at *revealed*, *in front of*, and *at the party*. Once again, effects appear to be strongest in subject position, however unlicensed gap effects in DO and OBL position are comparable in size. As with the filled gap effect, the 5-gram model only exhibited an unlicensed gap effect in subject position.

Statistical analysis revealed significant interactions at all the positions tested (p <0.001 in all cases) for the LSTM. For the 5-gram model, the interaction was only significant in subject position. The fact that the 5-gram model exhibits both effects in subject position indicates that there were sentences in the training set that contained a filler and a corresponding subject gap within a 5-word window. We suspect that the apparent filled gap effects were driven by two highly frequent bigrams: *hvem som* 'who REL' and *hva som* 'what REL', where the filler is immediately adjacent to the expletive relative pronoun *som* that signals a subject gap in embedded questions. The large unlicensed gap effect can be attributed to the absence of n-grams containing the declarative complementizer *at* and the relative pronoun *som*.

Our results suggest that the LSTM model learned that *wh*-fillers can be linked to subject, object, and OBL positions in Norwegian and that gaps in these positions must be licensed by a preceding filler. Thus we replicate Wilcox and colleagues' basic findings in Norwegian. We now turn to the second part of the Experiment 1 that tested RC dependencies.

RC dependencies

The experimental items for *wh*-dependencies were modified to create sentences with RC dependencies as follows: Main-clause verbs, like *høre* 'heard' in (8), were followed by a PP headed either by *fra* 'from' (in -FILLER) sentences or *om* 'about' (in +FILLER) sentences. PPs contained either the indefinite *noen* 'someone' or *noe* 'something'. In +FILLER sentences, the embedded clause was an RC, headed by the indefinite, followed by the relative pronoun *som*. In -FILLER sentences, the embedded clause was a complement of the main clause verb (*hørte*), followed by a PP with the indefinite *fra noen* 'from someone' and the declarative complementizer *at* 'that'. As above, the experiment manipulated the presence of a filler, the presence of a gap, and syntactic position. (8) illustrates the four SUBJECT conditions from a single item.

(8) Hun hørte... 'She heard...'

a. +FILLER, +GAP

om noen som __ avslørte hemmeligheten foran gjestene på festen
about someone REL __ revealed secret.DEF in front of guests.DEF at party.DEF

'about someone who __ revealed the secret in front of the guests at the party.'

b. +FILLER, -GAP

om noen som presten avslørte
about someone REL priest.DEF revealed
hemmeligheten foran gjestene på festen
secret.DEF in front of guests.DEF at party.DEF

'about someone who the priest revealed the secret in front of the guests at the party.'

c. -FILLER, +GAP

fra noen at __ avslørte
from someone that revealed secret.DEF
hemmeligheten foran gjestene på festen
in front of guests.DEF at party.DEF

'from someone that __ revealed the secret in front of the guests at the party.'

d. -FILLER, -GAP

fra noen at presten avslørte
from someone that priest.DEF revealed
hemmeligheten foran gjestene på festen
secret.DEF in front of guests.DEF at party.DEF

'from someone that the priest revealed the secret in front of the guests at the party.'

Filler effects for the LSTM model are presented in Figure 3 by sentence region and gap position. Filled gap and unlicensed gap effects for each gap position for the LSTM and the 5-gram model are in Figure 4. Overall, the qualitative pattern of effects for RC dependencies is almost identical to the pattern found with *wh*-FGDs.[1]

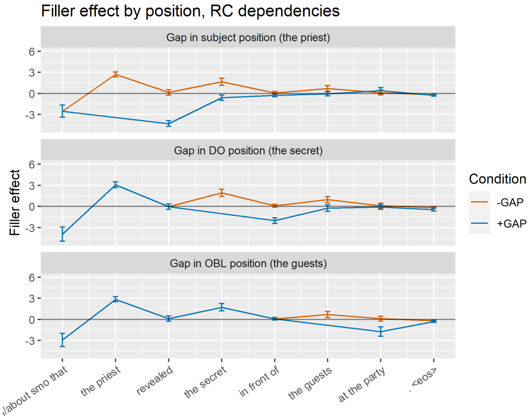


Figure 3: Filler effect for RC dependencies by sentence region and gap position.

[1] Negative difference scores in the region preceding the subject NP 'the priest' largely reflect the fact that embedding verbs like *å høre* 'to hear' are more commonly followed by the preposition *om* 'about' than by the preposition *fra* 'from'. As a result, our -FILLER sentences contained less frequent collocations in the matrix clause than +FILLER sentences, contributing to baseline surprisal differences. These differences, though, are orthogonal to our comparisons of interest.

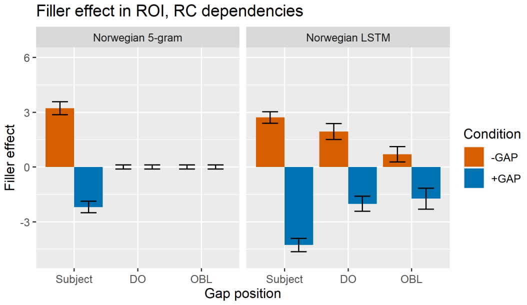


Figure 4: Filler effect for RC dependencies by position.

As with *wh*-FGDs, the LSTM model exhibits clear filled gap effects (-GAP conditions) at each potential gap position and it can distinguish between licensed and unlicensed gaps (+GAP conditions) at each position tested. Once again, the 5-gram model exhibits filled gap and unlicensed gap effects, but only in subject position. Statistical analysis confirmed a significant interaction between the presence of a filler and the presence of a gap at all three positions for the LSTM model and in subject position for the 5-gram model (p <0.001 in all cases). As with *wh*-FGDs, filled gap effects are largest in subject position and decline in size across the sentence. The unlicensed gap effect is largest in subject position, but the size does not differ between DO and OBL positions. Interestingly, the filled gap effect in subject position was larger for *wh*-FGDs (>4.5 bits) than with RCs (≈ 3 bits), though the opposite was true of the unlicensed gap effect.

Experiment 2: Distance between the filler and the gap

In Experiment 2, we manipulated the linear distance between the filler and the gap to test whether the network's representation of the dependency is robust to intervening material that is irrelevant to the FGD. We manipulated distance between the filler and the gap by varying the length of a phrase modifying a subject that came between the filler and the gap, as in (Wilcox et al., 2018). As in Experiment 1, we also manipulated the presence of the filler, the presence of the gap, and the position of the gap. However, in Experiment 2 we only investigated gaps in direct object or oblique position. As in Experiment 1, we measure the size of filled gap effects and unlicensed gap effects and test whether the interaction is significant. If the model can ignore the intervening material, we expect a significant interaction between the presence of the filler and the gap at both DO and OBL positions irrespective of modifier length. If the model's ability to represent the FGD is sensitive to the intervening material, we expect a three-way interaction between the presence of the filler, the presence of the gap, and modifier length.

Wh-dependencies

We began with 20 test items crossing the presence of the filler, the presence of the gap and gap position. We crossed these items with a four-level factor controlling modifier length: No modifier as in (9-a), short modifier (2-4 words), medium modifier (5-8 words) as in (9-b), and long modifier (8-12 words), distributed across the four modifier conditions, resulting in 640 test sentences. In their original materials (Wilcox et al., 2018) used modifiers that were composed either of PPs or RCs. Our modifiers only contained PPs and conjunctions. We chose not to use RCs in our modifiers so as not to introduce any verbs that could be misinterpreted as potential gap sites between our filler and gap positions.

(9) a. NO MODIFIER
Jeg vet hva piloten sendte __ til familien sin
I know what pilot.DEF sent __ to family.DEF his
etter landing
after landing
'I know what the pilot sent __ to his family after landing.'

b. MEDIUM MODIFIER
Jeg vet hva piloten [med den blå hatten og
I know what pilot.DEF with the blue hat.DEF and
kappen] sendte __ til familien sin etter landing
coat.DEF sent __ to family.DEF his after landing
'I know what the pilot [in the blue hat and coat] sent __ to his family after landing.'

Filler effects are presented in Figure 5 by modifier and position. Not pictured are the results from the 5-gram model which showed no effects across all conditions.

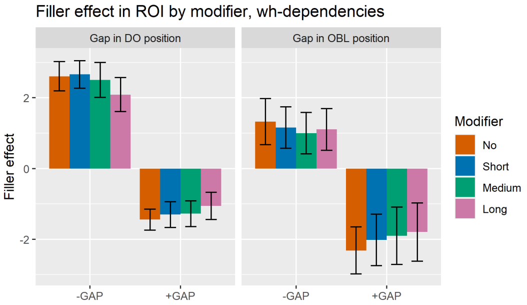


Figure 5: Filler effect for *wh*-dependencies by modifier.

As in Experiment 1, the model learned the bidirectional relationship between the presence of a filler and the presence of a gap by exhibiting filled gap effects and unlicensed gap effects in both DO and OBL position (p's <0.001). Filled gap effects were larger in DO position than in OBL position, but unlicensed gap effects were larger in OBL position. There was no significant effect of modifier length on filler-gap licensing.

RC dependencies

Materials for *wh*-FGDs were modified to create test items with RCs as in Experiment 1. Filled gap and unlicensed gap effects for RC dependencies are presented in Figure 6 by modifier and position. The 5-gram model yielded no effects.

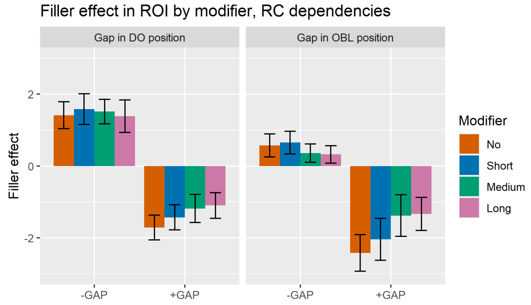


Figure 6: Filler effect for RC dependencies by modifier.

As with *wh*-dependencies, there was a significant two-way interaction between the presence of the filler and the presence of the gap for both positions (p's <0.001). For DO conditions, there was a significant three-way interaction between the presence of a filler, a gap and modifier length ($\beta = 0.05$, $t = 2.37$, $p = 0.018$) mostly driven by a modest diminishment in the size of the unlicensed gap effect as modifier length increased. A significant three-way interaction was also observed for OBL conditions ($\beta = 0.09$, $t = 3.44$, $p < 0.001$), once again driven mostly by smaller unlicensed gap effects with longer modifiers. Despite the decrease in size, however, unlicensed gap effects are still robust across modifier length. Once again we observed that filled gap effects were rather small at the OBL position compared to the DO position.

Taken together the results of Experiment 2 suggest that the model has strong expectations for gaps in DO position with *wh-* and RC dependencies alike. Expectations for a gap in OBL position are less robust, as observed in Experiment 1, but modifier length appears to have little effect on gap expectations. The model appears to recognize unlicensed gaps in both DO and OBL position with *wh-* and RC dependencies and although the size of the effect diminishes slightly with modifier length in RC dependencies, the length of intervening material does not consistently attenuate the model's ability to detect unlicensed gaps. Overall, there seems to be an asymmetry in how the model represents the bidirectional relationship between fillers and the gaps: Unlicensed gap effects are robust and may even increase in size towards towards the end of a sentence, while filled gap effects decrease dramatically between DO and OBL position. The decrease in the size of the filled gap effect suggests that the model has weaker expectations for an RC gap in OBL position than in DO position.

Conclusions and future work

In this paper we have shown that an LSTM RNN model was able to learn two basic properties of FGDs in Norwegian: flexibility in gap position (Experiment 1) and robustness to intervening material (Experiment 2). The model appears to generalize these properties over two dependency types: *wh-* and RC dependencies. Taken together with the results of (Wilcox et al., 2018, 2019), our results provide convergent evidence that general-purpose models without pre-defined language bias can learn basic syntactic generalizations about the distribution of acceptable FGDs across different languages.

The results presented here are promising but they do not conclusively establish that the models have a robust understanding of the distribution of FGDs in Norwegian. We identify two ways in which the test materials can be modified in order to further explore the robustness of the model's generalizations. First, Experiment 2 tested the effect of *linear* distance between the filler and the gap by manipulating the length of a subject modifier phrase as in (Wilcox et al., 2018). The experiment does not establish that the model understands that FGDs are structurally unbounded, as it did not manipulate *hierarchical* distance between the filler and the gap. (Wilcox et al., 2019) showed how hierarchical distance affects the models' abilities to detect filled and unlicensed gaps in English by manipulating layers of embedding, as in (10). Future work will test the effect of hierarchical distance on Norwegian FGD licensing.

(10) I know what [the postman said [the newspaper reported [the priest revealed __ at the party]]].

Second, in both *wh-* and RC dependencies that we tested, the gaps were licensed by an *overt* lexical item. In our *wh*-FGDs the overt licensor is the *wh*-word. In our RCs the licensor was the overt relative pronoun *som*. Not all grammatical FGDs, however, require overt lexical licensing. For example, RCs without overt relative pronouns or complementizers are possible in both English and Norwegian, as shown below:

(11) a. I saw the present [$_{RC}$ the pilot sent __].
b. Jeg så gaven [$_{RC}$ piloten sendte __].
I saw present.DEF pilot.DEF sent

Testing whether the models could successfully identify licit gaps in such RCs would help determine whether the model could recognize structural cues to FGDs, or whether it was limited to lexically-signalled dependencies.

In addition to the questions mentioned above, future work will explore whether LSTMs can learn about *islands*. Islands are environments that block formation of FGDs (Ross, 1967; Chomsky et al., 1977; Huang, 1982). Wilcox et al. report that RNNs learn that *wh*-FGDs are not allowed in some island environments in English - or at least that filler-gap licensing is attenuated inside of island environments. The generality of these results should be tested in other languages. Moreover, Norwegian represents a particularly interesting case with respect to the acquisition of island constraints, because Norwegian (like other Mainland Scandinavian languages like Swedish and Danish) is argued to only exhibit sensitivity to a subset of islands that languages like English are sensitive to (Maling & Zaenen, 1982; Engdahl, 1997; Kush et al., 2021). It will be interesting to see whether RNNs can learn a different set of island constraints from different input.

Acknowledgments

Our experiments were conducted using the resources of the NTNU IDUN computing cluster (Själander, Jahre, Tufte, & Reissmann, 2019). We thank Ingrid Bondevik and Charlotte Sant for their help in creating the stimuli in Norwegian.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, *68*(3), 255–278.

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

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211.

Engdahl, E. (1997). Relative clause extractions in context. *Working Papers in Scandinavian Syntax*, *60*, 51–79.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*.

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics.*

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Huang, C. T. J. (1982). *Logical relations in Chinese and the theory of grammar* (PhD dissertation). MIT.

Kam, X.-N. C., Stoyneshka, I., Tornyova, L., Fodor, J. D., & Sakas, W. G. (2008). Bigrams and the richness of the stimulus. *Cognitive science*, *32*(4), 771–787.

Kush, D., Sant, C., & Strætkvern, S. B. (2021). Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: a journal of general linguistics*, *6*(1), 1–50. Retrieved 2022-01-08, from `https://doi.org/10.16995/glossa.5774` doi: 10.16995/glossa.5774

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*, 521–535.

Maling, J., & Zaenen, A. (1982). A phrase structure account of Scandinavian extraction phenomena. In P. Jacobson & G. K. Pullum (Eds.), *The nature of syntactic representation* (pp. 229–282). Dordrecht: Springer Netherlands. Retrieved 2022-01-08, from `https://doi.org/10.1007/978-94-009-7707-57` doi: 10.1007/978-94-009-7707-57

McCoy, R. T., Frank, R., & Linzen, T. (2018). Revisiting the poverty of the stimulus: Hierarchical generalization without a hierarchical bias in recurrent neural networks. *arXiv preprint arXiv:1802.09091*.

Ross, J. R. (1967). *Constraints on variables in syntax* (PhD dissertation, MIT). Retrieved 2022-01-08, from `https://dspace.mit.edu/handle/1721.1/15166`

Själander, M., Jahre, M., Tufte, G., & Reissmann, N. (2019). *EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure.*

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, *1*(3), 227–245.

Wilcox, E., Levy, R., & Futrell, R. (2019). Hierarchical representation in neural language models: Suppression and recovery of expectations. *arXiv preprint arXiv:1906.04068*.

Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2018). What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Article C2

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2023). Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian. In T. Hunter & B. Prickett (Eds.), *Proceedings of the Society for Computation in Linguistics* (Vol. 6, pp. 175–185). DOI: 10.7275/qb8z-qc91

Neural Networks Can Learn Patterns of Island-insensitivity in Norwegian

Anastasia Kobzeva
Norwegian University of Science and Technology
anastasia.kobzeva@ntnu.no

Suhas Arehalli
Johns Hopkins University
suhas@jhu.edu

Tal Linzen
New York University
linzen@nyu.edu

Dave Kush
University of Toronto
dave.kush@utoronto.ca

Abstract

Recent research suggests that Recurrent Neural Networks (RNNs) can capture abstract generalizations about filler-gap dependencies (FGDs) in English and so-called *island* constraints on their distribution (Wilcox et al., 2018, 2021). These results have been interpreted as evidence that it is possible, in principle, to induce complex syntactic knowledge from the input without domain-specific learning biases. However, the English results alone do not establish that island constraints were induced from distributional properties of the training data instead of simply reflecting architectural limitations independent of the input to the models. We address this concern by investigating whether such models can learn the distribution of acceptable FGDs in Norwegian, a language that is sensitive to fewer islands than English (Christensen, 1982). Results from five experiments show that Long Short-Term Memory (LSTM) RNNs can (i) learn that Norwegian FGD formation is unbounded, (ii) recover the island status of temporal adjunct and subject islands, and (iii) learn that Norwegian, unlike English, permits FGDs into two types of embedded questions. The fact that LSTM RNNs can learn cross-linguistic differences in island facts therefore strengthens the claim that RNN language models can induce the constraints from patterns in the input.

1 Introduction

Human linguistic knowledge is complex and abstract, yet children master language relatively easily and quickly through exposure to their native language(s). A major debate centers around whether acquiring such knowledge requires complex domain-specific learning biases or whether it can be induced from the input using domain-general learning routines. We contribute to this debate by investigating whether Recurrent Neural Networks (RNNs), which are weakly biased language models, can induce complex knowledge of filler-gap dependencies and constraints on them from the input in Norwegian.

Filler-Gap Dependencies (FGDs) are contingencies between a displaced filler phrase and a later gap position where the filler is interpreted (denoted with __ throughout the paper). There are different types of FGDs. (1-a) is a *wh*-FGD where the filler *wh*-word is interpreted as the direct object of the verb *forged*. (1-b) is a Relative Clause (RC) FGD where the filler, the head of the RC, *painting*, is interpreted as the direct object of *forged* within the RC.

(1) a. They found out what the dealer forged __ using a new technique.
b. They found the painting that the dealer forged __ using a new technique.

FGDs have been the subject of extensive research because they require complex hierarchical generalizations about sentence structure to be interpreted. For example, establishing the RC FGD in (1-b) requires (i) identifying the head of the RC as a filler corresponding to a later empty NP position; (ii) knowing that *forged* requires a direct object; (iii) identifying the gap by recognizing the absence of an object next to *forged*, and (iv) associating the filler with the gap to form a dependency. There is a bidirectional relationship between the filler and the gap: fillers require gaps to be interpreted, and gaps require fillers to be properly licensed. This relationship can be established across a potentially unbounded structural distance as in (2).

(2) She knows what he thought they found out the dealer forged __ using a new technique.

FGDs are also constrained. Certain environments, called *islands* (Ross, 1967), block FGD formation. Various structures have been identified as islands. For example, embedded questions (3-a), sentential subjects (3-b), and adjuncts (3-c) are generally considered island domains in English.

(3) a. *What did he wonder [whether the dealer forged __]?
b. *What is [that the dealer forged __] extremely likely?
c. *What does the dealer worry [if they find out __]?

How do learners acquire island constraints? Nativist approaches hold that acquisition of islands would be impossible without innate domain-specific learning biases due to the induction problem known as the Poverty of the Stimulus (PoS; e.g., Chomsky 1986; Crain and Pietroski 2001). According to this argument, the input to the learner lacks direct evidence that islands exist. The input is therefore compatible with conflicting hypotheses about whether islands should be in the adult target state. The fact that learners nevertheless converge on the same set of island constraints has led the proponents of the nativist approach to suggest that innate domain-specific learning biases guide learners to the conclusion (for example, Subjacency Condition, Chomsky 1973).

Empiricist approaches, on the other hand, claim that the input is sufficiently rich to support learning island constraints when coupled with domain-general learning biases (Clark and Lappin, 2010). This position has recently gained support from neural network simulations. Wilcox and colleagues suggest that RNNs (and other autoregressive neural models) can capture the abstract generalizations governing *wh*-FGDs in English, as well as the associated island constraints (2018; 2019b; 2019a; 2021). They claim that this result militates against the PoS argument that islands cannot be induced from the input without domain-specific biases.

Wilcox and colleagues' results are suggestive, but they do not fully establish that the models 'learn' islands from the input. An alternate explanation is that the results are artifacts. Under this possibility, RNNs do not pursue FGDs into islands in English because the models are simply incapable of representing syntactic dependencies into island environments irrespective of the input they receive (either because the domains are too complex or because of some other unknown limitation inherent to the RNN architecture). One way of ruling out this explanation is to test the models' performance on a language that has a different set of island constraints. If the models can learn to pursue FGDs in another language into domains that are islands in English, that would constitute additional evidence that the models are inducing islands from the input.

To this end, we explore whether RNNs can learn the distribution of acceptable FGDs and island constraints in Norwegian – a language that differs from English in the set of domains that are islands. To preview our results, the models can learn that temporal adjuncts and subject phrases are islands in Norwegian, but that embedded questions are not (*wh*-islands). These results suggest that weakly-biased RNNs can capture patterns of island-insensitivity in Norwegian, thus providing empirical evidence that this pattern of cross-linguistic variation can be learned from the input.

2 Island constraints in Norwegian

Norwegian is similar to English in several respects when it comes to FGDs. Norwegian allows long-distance dependencies with gaps in various syntactic positions. Norwegian also exhibits sensitivity to some of the same islands that English does. FGDs into temporal adjuncts (4) or subject phrases (5) are unacceptable in Norwegian like English (Bondevik et al., 2021; Kush et al., 2019, 2018; Kobzeva et al., 2022b).

(4) *Hva spiste du kake [da han spiste __]?
What ate you cake when he ate __
*'What did you eat cake when he ate __?'

(5) *Hva har [brevet om __] skapt problemer?
What has letter.DEF about __ created problems
*'What has the letter about __ created problems?'

On the other hand, Norwegian allows FGDs into environments that are considered islands in English, such as Embedded Questions (EQs, Christensen 1982; Maling and Zaenen 1982). RC FGDs into embedded constituent questions like (6) are found in written corpora of Norwegian (Kush et al., 2021) and native speakers rate various types of FGD into EQs as acceptable in judgment studies (Kobzeva et al., 2022b).

(6) Vi var redde for noe vi ikke visste [hva __ var].
We were afraid of smth we NEG knew what __ was.
'We were afraid of something we did not know what __ was.'

This distribution of FGDs in Norwegian makes it a good testing ground for exploring whether RNNs can induce a set of islands that is different from what is observed in English. Recent research shows that RNNs can capture basic generalizations about *wh*- and RC FGDs in Norwegian: they learn that fillers can license gaps in different syntactic positions and across increased linear distance between the filler and the gap (Kobzeva et al., 2022a). Here we expand on this line of research by testing whether RNNs can learn that FGDs like (6) are acceptable in Norwegian, while simultaneously ruling out FGDs like (4) and (5). We do so by testing whether the models are less likely to expect FGDs in potential island environments relative to control sentences without island structures. We also test the robustness of the result by testing two more models with the same architecture but different initializations.

We ran five experiments. Experiment 1 tested whether the models learn that Norwegian FGDs are unbounded by seeing if they can successfully associate fillers and gaps across multiple embedded clauses. Establishing this basic result is a prerequisite for testing islands, which typically require cross-clausal dependencies. Experiments 2 and 3 tested if the models can learn that temporal adjunct clauses and complex subject phrases are islands in Norwegian, as in English. Finally, Experiments 4 and 5 tested if RNNs can learn that FGDs into embedded questions are possible in Norwegian. Experiments 1-4 evaluate the models performance on Norwegian only, while Experiment 5 directly compares *wh*-FGDs in Norwegian and English.

3 Method

3.1 Language models

We trained Long Short-Term Memory (LSTM) RNNs (Hochreiter and Schmidhuber, 1997) to take a sequence of words as input and compute a probability distribution of the next word over the model's vocabulary. We trained three such models with different random initializations following the procedure described in (Gulordava et al., 2018), using the code provided by the authors[1]. Each model was a 2-layer LSTM with 650 hidden units in each layer, trained for 40 epochs on 113 million tokens of Norwegian Wikipedia (in the Bokmål written standard) with a vocabulary size of 50000 most frequent words. The models achieved perplexities between 30.05 and 30.3 on the validation set.

3.2 Dependent measure

We test how the models would fare as incremental language processors by looking at *surprisal*, which measures how (un)predictable a word is given a specific prompt using the models' probability distribution. We measure the surprisal values by computing the negative log of the predicted conditional probability from the models' softmax layer.

3.3 Measuring FGDs

Wilcox et al. (2018) introduced a 2×2 factorial design for measuring FGDs inspired by psycholinguistic paradigms. The design independently manipulates the presence of a filler and the presence of a gap as in (7).

(7) They found out...
a. that the dealer forged the art -FILLER, -GAP
b. *what the dealer forged the art +FILLER, -GAP
c. *that the dealer forged __ -FILLER, +GAP
d. what the dealer forged __ +FILLER, +GAP
...using a new technique.

When both the filler and the gap are absent (7-a) or present (7-d), the sentences are grammatical. When either the filler or the gap is absent, (7-b) and (7-c), the sentences are ungrammatical. We measure *filler effects* – how the presence of a filler affects surprisal – in two different pairwise comparisons. *Filled gap effects* are measured by comparing surprisal associated with an NP in -GAP conditions. *Unlicensed gap effects* are measured by comparing surprisal associated with a gap in the +GAP conditions. We discuss each type of filler effect in more detail below.

3.3.1 Filled gap effects

In behavioral studies, filled gap effects are regarded as support for the *active gap-filling* strategy: after encountering a filler, the processor actively predicts a gap without waiting for the actual gap site. Stowe (1986) observed a slow-down in self-paced reading times at the direct object *us* in (8-b), which contains the filler *who*, compared to the same word in a corresponding sentence without a filler (8-a). The slow-down reflects a violated expectation: seeing a filler caused the processor to predict a gap in object position.

(8) a. My brother wanted to know if Ruth will bring *us* home to Mom at Christmas.
b. My brother wanted to know who Ruth will bring *us* home to __ at Christmas.

We test whether the models exhibit similar filled gap effects. We measure the surprisal difference between the ungrammatical +FILLER, -GAP condition as in (7-b) and the grammatical -FILLER, -GAP condition in (7-a) at the region of the filled NP (*the*

[1] https://github.com/facebookresearch/colorlessgreenRNNs

art in (7)). If seeing a filler sets up an expectation for a gap in object position, the NP should be more surprising in (7-b) than in (7-a), resulting in a *positive* surprisal difference.

Crucially, humans do not exhibit filled gap effects inside island environments (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006), indicating that the active prediction of gaps is suspended where they are impossible. Following the same logic, if the models show sensitivity to island constraints, we expect to see no filled gap effects inside islands.

3.3.2 Unlicensed gap effects

Unlicensed gap effects provide a measure of how 'surprised' the model is to encounter a gap without a filler to license it. We measure these effects as a difference in surprisal between the grammatical +FILLER, +GAP (7-d) condition and ungrammatical -FILLER, +GAP (7-c) condition at the region following the gap (*using a new technique* in (7)). If a presence of a gap without a licensing filler is surprising to the models, the unlicensed gap effect should manifest as a negative difference between low surprisal in the post-gap region in (7-d) and high surprisal in (7-c).

Unlicensed gap effects show if the models recognize gaps as licit inside certain syntactic environments. Whereas filled gap effects measure the models' expectation for an upcoming gap, unlicensed gap effects arguably should reflect the models' understanding of grammaticality, as sentences with illicit gaps are ungrammatical (and, unlike filled gaps, cannot be 'rescued' by establishing another gap site later in a sentence). Analogous to filled gap effects, unlicensed gap effects should be close to zero in island environments if the models can derive their island status from their training data.

3.4 Statistical analysis

Following standard practice in psycholinguistics, statistical analysis was performed using mixed-effect linear regression models with sum-coded fixed effects of FILLER (0.5 for +FILLER, -0.5 for -FILLER) and CONDITION (0.5 for CONTROL and -0.5 for ISLAND except for Experiments 1 and 4, see details below). We fit the statistical models on differences in surprisal between +FILLER, -FILLER conditions with these fixed effects and a maximal random effect structure (Barr et al., 2013). We ran separate models for filled gap effects in the filled NP region and for unlicensed gap effects in the post-gap region. If a model failed to converge, we reduced the random effect structure until convergence was reached. Model formulas are presented in Appendix A.

4 Experiments

4.1 Experiment 1: Unboundedness

It is important to establish whether LSTMs can represent FGDs across hierarchical distance before testing island environments, as they involve cross-clausal dependencies. Therefore, in Experiment 1 we tested how increased hierarchical distance between the filler and the gap influences models' representations of FGDs. To do that, we manipulated the number of clausal embeddings between the filler and the gap (from 1 to 5 layers of clausal embedding, as illustrated in (9)). We created 30 items by crossing the factors FILLER and GAP in (7) with NUMBER OF LAYERS, resulting in a 2 × 2 × 5 design. Test sets were created for *wh-* and RC FGDs (600 test sentences per dependency type).

(9) a. 1 LAYER (+FILLER, +GAP)
Hun vet hva selgeren forfalsket __ ved hjelp
She knows what dealer.DEF forged __ with help
av moderne teknologi.
of modern technology.
'She knows what the dealer forged __ using modern technology'.

b. 5 LAYERS (+FILLER, +GAP)
Hun vet hva han trodde de fant ut
She knows what he thought they found out
avisen rapporterte politiet visste
newspaper.DEF reported police.DEF knew
selgeren forfalsket __ ved hjelp av moderne
dealer.DEF forged __ with help of modern
teknologi.
technology.
'She knows what he thought they found out the newspaper reported the police knew the dealer forged __ using modern technology'.

We tested all three models on all of the items, and we present the results averaged across the models for both dependency types together. Overall, filler effects decrease as layers of embedding increase (Figure 1). For *wh*-dependencies (blue bars), there was a significant reduction in both the filled gap effect and the unlicensed gap effect already at two layers of embedding, which was also true for every layer thereafter (p's <0.05 in all cases). For RC dependencies (orange bars), there was a significant reduction in filled gap effects at three layers (p <0.05), and in unlicensed gap effects at two layers (p's <0.001) of sentential embedding, as well as for every layer thereafter (p's <0.001 in all cases).

Tables with statistics summary can be found in Appendix A.

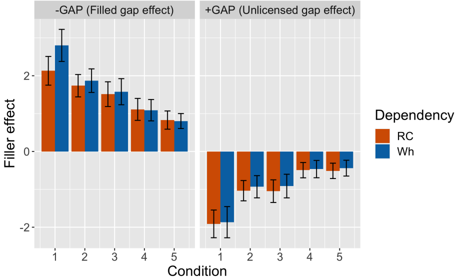


Figure 1: Unboundedness experiment: Filler effects by the number of embeddings for both dependency types. Bars represent an average over three models, error bars represent 95% confidence intervals.

Despite the reduction in filler effects as a function of the number of sentential embeddings, the filler effects remain above zero even at the largest hierarchical distance. This suggests that the models have learned that FGD formation is unbounded and have the basic representational capacity required for testing FGDs inside islands.

4.2 Islands shared between Norwegian and English

Experiments 2 and 3 tested FGDs into constituents that are islands in Norwegian (just as in English) – subjects and temporal adjunct clauses – to see if the models' expectations for FGDs are attenuated within the two environments in Norwegian, as previously seen in English (Wilcox et al., 2018, 2021).

4.2.1 Experiment 2: Subject island

Fillers cannot be associated with gaps inside a subject phrase, like the gap inside the prepositional phrase attached to the subject in (10). Such sentences are rated as unacceptable by English speakers, and the same pattern is found in Norwegian (11-b). We compare the island condition in (11-b) to an NP-subject extraction as in (11-a).

(10) *The newspaper reported what [the agreement with __] will strengthen the political interaction after the elections.

(11) a. SUBJECT CONTROL (+FILLER, +GAP)

Avisen rapporterte hva som __ vil
Newspaper.DEF reported what REL __ will
forsterke det politiske samspillet etter
strengthen the political interaction.DEF after
valget.
election.DEF

'The newspaper reported what __ will strengthen the political interaction after the election.'

b. SUBJECT ISLAND (+FILLER, +GAP)

*Avisen rapporterte hva [avtalen
Newspaper.DEF reported what agreement.DEF
med __] vil forsterke det politiske
with __ will strengthen the political
samspillet etter valget.
interaction.DEF after election.DEF

'*The newspaper reported what the agreement with __ will strengthen the political interaction after the election.'

We created 30 items according to a 2 × 2 × 2 design that crossed the factors FILLER and GAP in (7) with a third factor: CONDITION (CONTROL, ISLAND). Again we created separate sets of sentences for *wh-* and RC FGDs (240 total test sentences per dependency type). The results of this experiment are presented in Figure 2.

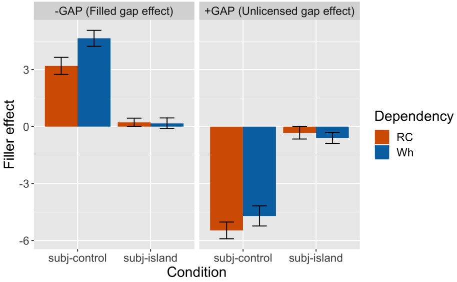


Figure 2: Subject island experiment: Filler effects by gap position for both dependency types.

Filled gap effects (Figure 2 left panel) were large in the control condition, but were significantly reduced in the island condition: statistical analysis revealed a main effect of CONDITION for both dependency types (both p's <0.001). The same pattern was found for unlicensed gap effects (Figure 2 right panel). For both dependency types, there was a significant effect of CONDITION (p's <0.001 in both cases). These results show that the models exhibit reduced filler effects within subject islands, which is in line with behavioral acceptability data from native Norwegian speakers.

4.2.2 Experiment 3: Adjunct island

Adjuncts are said to block FGD formation, which explains the unacceptability of (12): The filler *what* cannot be associated with the gap inside the adjunct *when*-clause. Norwegian, like English, does not allow gaps inside temporal adjuncts (Bondevik et al., 2021; Bondevik and Lohndal, 2023).

(12) *What were the voters excited [when the politician visited __ last week]?

We created 30 items according to a 2 × 2 × 3 design that crossed FILLER, GAP, and CONDITION for each dependency type (360 test sentences per dependency). CONDITION had three levels that determined the location of a direct object gap. In the LINEAR CONTROL (13-a) and STRUCTURAL CONTROL (13-b) the gap was not embedded in an island, whereas in ADJUNCT ISLAND (13-c), the gap was embedded inside a temporal adjunct (headed by *mens 'while', da 'when', etter at 'after'* and *før 'before'*). In the linear control condition (13-a), first used in (Wilcox et al., 2018), the filler and gap are in the same clause, but the linear distance between them is comparable to the distance in (13-c). In the structural control condition (13-b), our novel addition to the design, the filler and the gap are separated across two clauses, making the *structural* distance between the filler and the gap comparable to (13-c). We included these control conditions in order to estimate the independent effects of linear distance and structural distance on the model's performance, so as to better isolate island effects.

(13) a. LINEAR CONTROL (+FILLER, +GAP)
Jeg husker hva politikeren med godt
I remember what politician.DEF with good
omdømme besøkte __ forrige uke.
reputation visited __ last week.
'I remember what the politician with a good reputation visited __ last week.'

b. STRUCTURAL CONTROL (+FILLER, +GAP)
Jeg husker hva avisen rapporterte at
I remember what newspaper.DEF reported that
politikeren besøkte __ forrige uke.
politician.DEF visited __ last week.
'I remember what the newspaper reported that the politician visited __ last week.'

c. ADJUNCT ISLAND (+FILLER, +GAP)
*Jeg husker hva velgerne var begeistret
I remember what voters.DEF were excited
da politikeren besøkte __ forrige uke.
when politician.DEF visited __ last week.
'*I remember what the voters were excited when the politician visited __ last week.'

We defined two contrasts for analysis: CONTROL contrast compared effect size between the two control conditions (linear vs. structural). ISLAND contrast compared effects between the structural control and the adjunct island condition.

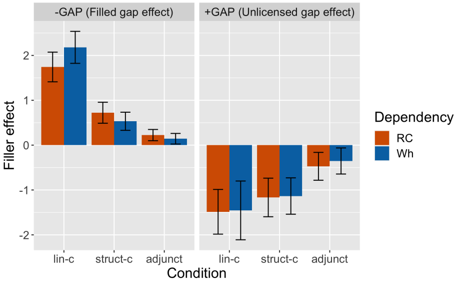


Figure 3: Adjunct island experiment: Filler effects by condition for both dependency types. Control conditions are `lin-c` and `struct-c`.

The results of the experiment are presented in Figure 3. Filled gap effects for both dependency types (left panel) were largest in the linear control condition, significantly larger than in the structural control condition (CONTROL contrast p's <0.001). Filled gap effects were in turn significantly larger in the structural control condition than in the adjunct island condition (ISLAND contrast p's <0.001), where filled gap effects were close to zero.

The same qualitative pattern was observed with unlicensed gap effects for both dependency types (right panel). Unlicensed gap effects were larger in the linear control condition compared to the structural control, and in the structural control condition compared to the island condition (p's <0.001 in all cases). Therefore, the models show reduced filler effects inside temporal adjuncts in Norwegian. However, the average filler effects are not 0 in the adjunct island condition, suggesting that the models might not treat them as full islands.[2] Norwegian shows some variation in adjunct island effects, with extraction from conditional adjuncts rated higher than from temporal and reason-adjuncts (Bondevik et al., 2021; Bondevik and Lohndal, 2023). The result obtained here could be explained by the models' sensitivity to this variation (and potential overgeneralization).

4.3 Islands contrasting English and Norwegian

The results of Experiments 2 and 3 suggest that the models learn that subjects and temporal adjuncts are islands in Norwegian, similar to the conclusions

[2] On around 65% of the trials, the models show filled-gap effects greater than zero, while unlicensed gap effects are less than zero on around 70% of the trials. However, the effects are mostly small, under 1 bit of surprisal 90% of the time.

made for English by Wilcox et al.. Experiments 4 and 5 test whether the models can learn that embedded questions (EQs) are not islands in Norwegian. We test two types of EQs in Norwegian: 1) interrogative EQs, and 2) *whether*-EQs.

4.3.1 Experiment 4: Interrogative EQ

According to Kush et al. (2021), the most common type of extraction from EQs (in a children's fiction corpus) includes a subject gap inside an interrogative EQ as in (14).

(14) Vi var redde for noe vi ikke visste [hva __ var].
We were afraid of smth we NEG knew what __ was.
'We were afraid of something we did not know what __ was.'

We chose to first test such EQs because we reasoned that they were likely the most frequent in the model's training data. We created 30 items that crossed FILLER, GAP, and CONDITION for each dependency type (240 test sentences per dependency). CONDITION controlled whether the embedded clause was an EQ (15-b) or a declarative complement (15-a) control.[3]

(15) a. DECLARATIVE CONTROL (+FILLER, +GAP)
Han sa hvem som sjåføren glemte at __ skulle hentes i sentrum den dagen.
He said who REL driver.DEF forgot that __ should be.picked.up in center.DEF that day.DEF.
'He said who$_i$ the driver forgot (that) __$_i$ should be picked up in the center that day.'

b. WH-ISLAND (+FILLER, +GAP)
Han sa hvem som sjåføren glemte hvor __ skulle hentes __ den dagen.
He said who REL driver.DEF forgot where __ should be.picked.up that day.DEF.
'He said who$_i$ the driver forgot where$_k$ __$_i$ should be picked up __$_k$ that day.'

We expected clear filled gap effects and unlicensed gap effects in the declarative clauses. If the models recognize that interrogative EQs are not islands in Norwegian, the filled gap effects and unlicensed gap effects in the EQ sentences should be comparable to their declarative counterparts, or at least greater than zero.

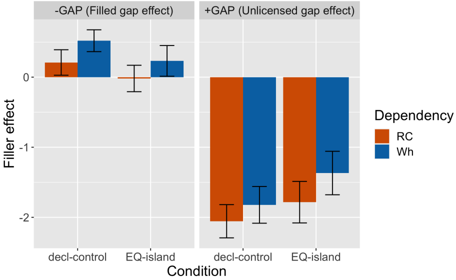


Figure 4: Interrogative EQ island experiment: Filler effects by condition for both dependency types.

Figure 4 shows that filled gap effects were small or close to 0 across all conditions and dependency types, while unlicensed gap effects were large. Statistical analysis revealed a main effect of CONDITION for both filled gap effects and unlicensed gap effects with *wh*-dependencies (p's <0.01). With RC dependencies, the same was true for the filled gap effect (p <0.05, orange bars on the left panel). For the unlicensed gap effect with RC dependencies, the effect of CONDITION was not significant (p <0.1). Importantly, despite the significant effect of CONDITION in three out of four cases tested, both filled gap effects and unlicensed gap effects in the island condition were comparable to the declarative control, suggesting that the models treat EQs and embedded declarative clauses similarly with respect to FGD formation in Norwegian.

4.3.2 Experiment 5: *Whether*-EQ

In Experiment 4, we tested FGDs into interrogative EQs with gaps in subject position. However, previous research in English has not tested interrogative EQs and has instead focused on FGDs into polar EQs, *whether*-islands. For example, Wilcox et al. tested *whether*-islands with gaps in object position in English. An example of +FILLER, +GAP, ISLAND condition from their *whether*-island experiment is presented in (16).

(16) *I know what my brother said whether our aunt devoured __ at the party.

In order to facilitate more direct cross-linguistic comparison, and to test the robustness of the result of Experiment 4, we decided to run an experiment comparing FGDs into *whether*-EQs in English and Norwegian side by side. To do so, we slightly modified the 24 English items from (Wilcox et al., 2018) and created 24 novel items following the same tem-

[3] The direct translation of (15-b) would be ungrammatical in English due to *that*-trace effects. Norwegian exhibits some variation in *that*-trace effects; theoretical and experimental work shows that it mostly allows subject gaps after *that* (Lohndal, 2009; Kush and Dahl, 2020). We return to this issue in the Discussion.

plate, resulting in 48 items total. We then translated them into Norwegian. As the original (Wilcox et al., 2018) items did not include RC dependencies, we restricted dependency types to *wh*-FGDs in this experiment. We compared the performance of the Gulordava model (used by Wilcox et al., 2018) on English stimuli and the performance of one of the Norwegian models (used by Kobzeva et al., 2022a). The results are presented in Figure 5.

Overall, filler effects are smaller in English (light blue bars) than in Norwegian (dark blue bars; main effect of LANGUAGE, p <0.001). The pattern of island sensitivity also differs. In Norwegian, robust filled gap effects were observed in both declarative control and *whether*-island environments, while in English, no filled gap effect was observed inside a *whether*-island (left panel). Statistical analysis confirmed a significant CONDITION × LANGUAGE interaction for filled gap effects (p <0.01). Similar differences were observed for unlicensed gap effects (right panel): In Norwegian, unlicensed gap effects are equally large in declarative complements and *whether*-islands, whereas there is no unlicensed gap effect inside a *whether*-island in English compared to the declarative control (CONDITION × LANGUAGE p <0.05).

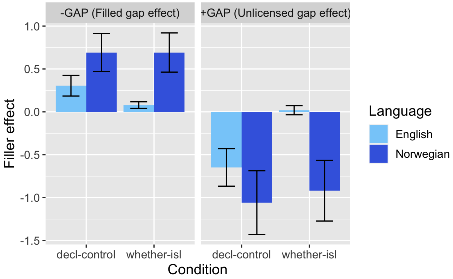


Figure 5: *Whether*-island experiment (with *wh*-dependencies): Comparison of filler effects in English and Norwegian.

Taken together with the fact that the architecture of the English and the Norwegian model was the same, and that they were trained using the same hyper-parameter combination for the same number of epochs on input data that were comparable in size and genre, these results suggest that RNNs can come to different conclusions about the status of *whether*-islands based on different language input. This provides further evidence for the claim, made in Wilcox et al., that autoregressive language models can learn the distribution of FGDs in a language from their input.

5 Discussion

In this paper, we tested LSTMs' ability to establish FGDs in Norwegian by looking at filled gap effects and unlicensed gap effects. Experiment 1 found non-zero filled gap effects and unlicensed gap effects across multiple layers of embedding suggesting that the models learn that FGDs are unbounded. Experiments 2 and 3 showed that filled gap effects and unlicensed gap effects are significantly reduced inside subject phrases and temporal adjuncts, suggesting that the models learned that these domains are islands in Norwegian, mirroring previous findings for English (Wilcox et al., 2018, 2019a,b, 2021).

Broadly speaking, results from Experiments 4 and 5 suggest that the models can learn that embedded questions are not island environments in Norwegian. In both Experiment 4 and 5, we found large unlicensed gap effects in Norwegian interrogative EQs and in Experiment 5 we observed filled gap effects inside Norwegian *whether*-EQs. Taken together, the results are consistent with the conclusion that LSTM RNNs can learn cross-linguistic differences in island facts from different language input. We do not know whether the model's generalization was derived from actual examples of FGDs into embedded questions in the training data, or whether the model learned the distribution indirectly. We cannot verify that in this case that the models learned from direct evidence, but it is plausible that such evidence would be available in the Wikipedia corpus given that FGDs into embedded questions are found (in relatively small numbers) in other corpora (such as the child fiction corpus investigated by Kush et al., 2021).

One potentially surprising finding was the asymmetry in filled and unlicensed gap effects between Experiments 4 and 5. In Experiment 4, filled gap effects were not robust in subject position, but unlicensed gap effects were. In Experiment 5, both filled gap effects and unlicensed gap effects were observed in object position. We take this effect to mean that the model was not actively pursuing embedded subject gaps in our stimuli. There are various possible interpretations for this effect. One possibility is that the model avoids gaps after overt material in left edge of a clause (a kind of *that-trace* effect, see Lohndal, 2009). Another possibility is that embedded subject gaps were not frequent enough in the training data to establish strong expectations for them.

We do not take the fact that filled gap effects are absent in some EQs as evidence against the models being able to establish FGDs into EQs. Even in the absence of filled gap effects, unlicensed gap effects show that the models can still recognize gaps in EQs as licit in Norwegian. We think that unlicensed gap effects provide a better indication of what the models have learned is possible. In other words, the two effects measure different aspects related to an FGD: While filled gap effects measure active expectation/prediction for a gap inside a particular structural configuration (i.e. whether the models think that a gap is *likely* in a given position), unlicensed gap effects reflect whether the models 'understand' that FGDs are in principle possible in that configuration. We suggest that future work using this paradigm should keep this dissociation in mind when interpreting results: Learning what a possible FGD is, does not necessarily entail active expectation in RNN language models.

One outstanding question is how well the model's active gap-filling behavior mirrors how actual humans would process these sentences. Native English speakers do not actively pursue gaps inside islands (Stowe, 1986; Traxler and Pickering, 1996; Phillips, 2006). In this regard, the English models of Wilcox et al. mimic human behavior. It is unknown whether native Norwegian speakers suspend active gap-filling inside islands, but pursue active gap-filling inside structures like EQs, that are not islands in their language. Future work should test the alignment between the model's performance and human behavior.

6 Conclusion

In this study, we tested whether LSTMs, an RNN architecture without language-specific bias, can learn two types of filler-gap dependencies in Norwegian in several (potential) island environments. We found evidence that the models can pick up patterns of island-insensitivity when it comes to embedded questions in Norwegian, while still inducing island effects in subject and adjunct islands. Our results also show that RNNs are sensitive to differences in the distribution of FGDs in English and Norwegian, suggesting that the input to the models must provide enough evidence for the diverging patterns. Our results lead us to reassess the importance of domain-specific learning biases in acquiring island constraints from the input.

Acknowledgements

All experimental materials and analysis scripts are available in the project's OSF repository. Our models were trained with the resources from NTNU's IDUN computing cluster (Själander et al., 2019). We thank the members of NTNU's EyeLands Lab (Øyelab) for their comments on an earlier version of this paper. We also thank Sigurd Farstad Iversen for his help in creating the stimuli in Norwegian.

References

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Ingrid Bondevik, Dave Kush, and Terje Lohndal. 2021. Variation in adjunct islands: The case of Norwegian. *Nordic Journal of Linguistics*, 44(3):223–254.

Ingrid Bondevik and Terje Lohndal. 2023. Extraction from finite adjunct clauses: an investigation of relative clause dependencies in norwegian. *Glossa: a journal of general linguistics*, 8(1).

Noam Chomsky. 1973. Conditions on transformations. In Morris Halle, Stephen R. Anderson, and Paul Kiparsky, editors, *A Festschrift for Morris Halle*, pages 232–286. Holt, Rinehart and Winston, New York.

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

Kirsti Koch Christensen. 1982. On multiple filler-gap constructions in Norwegian. In Elisabet Engdahl and Eva Ejerhed, editors, *Readings on unbounded dependencies in Scandinavian languages*, pages 77–98. Almquist & Wiksell, Stockholm.

Alexander Clark and Shalom Lappin. 2010. *Linguistic Nativism and the Poverty of the Stimulus*. John Wiley & Sons.

Stephen Crain and Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and philosophy*, 24(2):139–186.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of NAACL 2018*, pages 1195–1205.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Anastasia Kobzeva, Suhas Arehalli, Tal Linzen, and Dave Kush. 2022a. LSTMs Can Learn Basic Wh- and Relative Clause Dependencies in Norwegian. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Anastasia Kobzeva, Charlotte Sant, Parker T. Robbins, Myrte Vos, Terje Lohndal, and Dave Kush. 2022b. Comparing island effects for different dependency types in Norwegian. *Languages*, 7(3):195–220.

Dave Kush and Anne Dahl. 2020. L2 transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research*, pages 1–32.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2018. Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, 36(3):743–779.

Dave Kush, Terje Lohndal, and Jon Sprouse. 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, 95(3):393–420.

Dave Kush, Charlotte Sant, and Sunniva Briså Strætkvern. 2021. Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: a journal of general linguistics*, 6(1):1–50.

Terje Lohndal. 2009. Comp-t effects: Variation in the position and features of C. *Studia Linguistica*, 63(2):204–232.

Joan Maling and Annie Zaenen. 1982. A phrase structure account of Scandinavian extraction phenomena. In Pauline Jacobson and Geoffrey K. Pullum, editors, *The Nature of Syntactic Representation*, pages 229–282. Springer Netherlands, Dordrecht.

Colin Phillips. 2006. The real-time status of island phenomena. *Language*, pages 795–823.

John Robert Ross. 1967. *Constraints on variables in syntax*. PhD dissertation, MIT.

Magnus Själander, Magnus Jahre, Gunnar Tufte, and Nico Reissmann. 2019. EPIC: An energy-efficient, high-performance GPGPU computing research infrastructure.

Laurie A Stowe. 1986. Parsing wh-constructions: Evidence for on-line gap location. *Language and cognitive processes*, 1(3):227–245.

Matthew J Traxler and Martin J Pickering. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35(3):454–475.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations. In *Proceedings of the 2019 ACL Workshop BlackboxNLP*, pages 181–190.

Ethan Wilcox, Roger Levy, and Richard Futrell. 2019b. What syntactic structures block dependencies in RNN language models? *arXiv preprint arXiv:1905.10431*.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN Language Models Learn about Filler-Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*, pages 211–221.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2021. Using computational models to test syntactic learnability. *Linguistic Inquiry*, pages 1–88.

A Results of Statistical Tests

The levels of significance used in the tables below: + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. The statistics are presented separately for filled gap effects (FGE) and unlicensed gap effects (UGE) by each dependency type and experiment. The response variable *s* in `lmer` formulas is the difference in surprisal between +FILLER, -FILLER conditions.

1. Unboundedness

s ~lyrs + (1+lyrs | model) + (1+lyrs | item)

	FGE, *wh*-dependencies		
	Est.	S.E.	t
(Intercept)	2.801	0.304	9.221***
layers2	−0.931	0.220	−4.240*
layers3	−1.223	0.204	−5.980***
layers4	−1.711	0.246	−6.959***
layers5	−1.997	0.219	−9.104***
	UGE, *wh*-dependencies		
(Intercept)	−1.867	0.147	−12.681***
layers2	0.936	0.099	9.488***
layers3	0.954	0.099	9.671***
layers4	1.402	0.099	14.212***
layers5	1.427	0.099	14.465***
	FGE, RC dependencies		
(Intercept)	2.131	0.194	10.971***
layers2	−0.394	0.281	−1.402
layers3	−0.617	0.237	−2.598*
layers4	−1.019	0.203	−5.024***
layers5	−1.301	0.233	−5.593**
	UGE, RC dependencies		
(Intercept)	−1.912	0.192	−9.954***
layers2	0.877	0.161	5.447***
layers3	0.864	0.156	5.557***
layers4	1.419	0.166	8.564***
layers5	1.400	0.158	8.885***

2. Subject island

s ~cond + (1+cond | model) + (1+cond | item)

	FGE, *wh*-dependencies		
	Est.	S.E.	t
(Intercept)	2.411	0.255	9.459***
condition	4.476	0.335	13.368***
	UGE, *wh*-dependencies		
(Intercept)	−2.658	0.255	−10.437***
condition	−4.098	0.488	−8.390***
	FGE, RC dependencies		
(Intercept)	1.713	0.132	12.944***
condition	2.970	0.254	11.697***
	UGE, RC dependencies		
(Intercept)	−2.895	0.223	−13.008***
condition	−5.147	0.383	−13.455***

3. Adjunct island

s ~cntrs + (1+cntrs | model) + (1+cntrs | item)

	FGE, *wh*-dependencies		
	Est.	S.E.	t
(Intercept)	0.952	0.127	7.476***
controlCntrs	2.457	0.232	10.609***
islandCntrs	1.618	0.221	7.323***
	UGE, *wh*-dependencies		
(Intercept)	−0.981	0.208	−4.714***
controlCntrst	−0.948	0.298	−3.182**
islandCntrst	−1.255	0.273	−4.602***
	FGE, RC dependencies		
(Intercept)	0.896	0.136	6.593***
controlCntrst	1.692	0.200	8.454***
islandCntrst	1.344	0.234	5.755***
	UGE, RC dependencies		
(Intercept)	−1.042	0.187	−5.569***
controlCntrst	−0.889	0.201	−4.423***
islandCntrst	−1.139	0.178	−6.382***

4. Interrogative EQ

s ~cond + (1+cond | model) + (1+cond | item)

	FGE, *wh*-dependencies		
	Est.	S.E.	t
(Intercept)	0.376	0.081	4.647***
condition	0.288	0.107	2.690**
	UGE, *wh*-dependencies		
(Intercept)	−1.595	0.260	−6.142***
condition	−0.454	0.153	−2.961**
	FGE, RC dependencies		
(Intercept)	0.095	0.080	1.189
condition	0.228	0.100	2.271*
	UGE, RC dependencies		
(Intercept)	−1.920	0.220	−8.707***
condition	−0.272	0.152	−1.795+

5. *Whether*-EQ

*s ~condition*language + (1+condition | item)*

	FGE		
	Est.	S.E.	t
(Intercept)	0.617	0.074	8.388***
condition	0.109	0.102	1.074
language	0.700	0.102	6.880***
condition:language	−0.625	0.204	−3.073**
	UGE		
(Intercept)	−0.652	0.099	−6.570***
condition	−0.354	0.132	−2.690**
language	−0.676	0.127	−5.346***
condition:language	0.627	0.253	2.477*

Article J1

Kobzeva, A. & Kush, D. (2024). Grammar and Expectation in Active Dependency Resolution: Experimental and Modeling Evidence from Norwegian. *Cognitive Science, 48*(10), e1350. DOI: 10.1111/cogs.13501

COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 48 (2024) e13501

ISSN: 1551-6709 online
DOI: 10.1111/cogs.13501

Grammar and Expectation in Active Dependency Resolution: Experimental and Modeling Evidence from Norwegian

Anastasia Kobzeva,[a] Dave Kush[b,c]

[a]*Department of Language and Literature, Norwegian University of Science and Technology*
[b]*Department of Language Studies, University of Toronto*
[c]*Department of Linguistics, University of Toronto*

Received 8 March 2024; received in revised form 10 September 2024; accepted 20 September 2024

Abstract

Filler-gap dependency resolution is often characterized as an active process. We probed the mechanisms that determine where and why comprehenders posit gaps during incremental processing using Norwegian as our test language. First, we investigated why active filler-gap dependency resolution is suspended inside *island* domains like embedded questions in some languages. Processing-based accounts hold that resource limitations prevent gap-filling in embedded questions across languages, while grammar-based accounts predict that active gap-filling is only blocked in languages where embedded questions are grammatical islands. In a self-paced reading study, we find that Norwegian participants exhibit filled-gap effects inside embedded questions, which are not islands in the language. The findings are consistent with grammar-based, but not processing, accounts. Second, we asked if active filler-gap processing can be understood as a special case of probabilistic ambiguity resolution within an *expectation-based* framework. To do so, we tested whether word-by-word surprisal values from a neural language model could predict the location and magnitude of filled-gap effects in our behavioral data. We find that surprisal accurately tracks the location of filled-gap effects but severely underestimates their magnitude. This suggests either that mechanisms above and beyond probabilistic ambiguity resolution are required to fully explain active gap-filling behavior or that surprisal values

Correspondence should be sent to Anastasia Kobzeva, Department of Language and Literature, Norwegian University of Science and Technology, ISL NTNU, Edvard Bulls veg 1, 7049 Trondheim, Norway. E-mail: anastasia.kobzeva@ntnu.no

derived from a long short-term memory neural language model are not good proxies for humans' incremental expectations during filler-gap resolution.

Keywords: Active gap-filling; Filled-gap effects; Island constraints; Neural language models; Surprisal; Norwegian

1. Introduction

Sentence comprehension routinely requires processing long-distance *filler-gap* dependencies. For example, understanding the relative clause (RC) in (1) involves linking a filler, *the experiment*, to a later gap site (marked with an underscore) so that it can be interpreted as the object of the verb *conducted*.

(1) I took part in the experiment$_i$ [that the research group in psycholinguistics conducted __$_i$].

Studies on diverse languages show that incremental processing of filler-gap dependencies is *active*: Comprehenders eagerly posit gaps in upcoming positions before they can be sure of the gaps' true location (Aoshima, Phillips, & Weinberg, 2004; Atkinson, Wagers, Lidz, Phillips, & Omaki, 2018; Frazier & Flores d'Arcais, 1989; Lee, 2004; Stowe, 1986; Traxler & Pickering, 1996; Wagers, Borja, & Chung, 2015). *Filled-gap effects* are one piece of evidence for active gap-filling. A well-known example comes from Stowe (1986), where reading times (RTs) at the direct object *us* were compared between a condition in which participants had to resolve a filler-gap dependency (2-a) and a condition in which they did not (2-b).

(2) a. My brother wanted to know who$_i$ Ruth will bring us home to __$_i$ at Christmas.
b. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

Participants took longer to read the direct object *us* in (2-a), indicating that they initially analyzed the filler *who* as the object of *bring* and were surprised when that analysis was disconfirmed.

In this paper, we explore how to best characterize where and why comprehenders pursue active gap-filling. As our primary focus, we consider why active gap-filling appears to be suspended inside certain domains. *Islands* are constituents that block filler-gap dependency formation (Ross, 1967). *Embedded questions* (3-a) and *subject phrases* (4-a) are considered islands in English. As seen in (3-b) and (4-b), filler-gap dependencies that cross into these domains are unacceptable.

(3) a. I know [who conducted this experiment.]
b. *That's the experiment$_i$ that I know [who conducted __$_i$.]
(4) a. [The failure of the experiment] disappointed the scientists.
b. *That's the experiment$_i$ that [the failure of __$_i$] disappointed the scientists.

Studies on filler-gap processing in English suggest that comprehenders immediately recognize island boundaries (Kluender & Kutas, 1993; McKinnon & Osterhout, 1996) and avoid actively positing gaps in islands (Omaki & Schulz, 2011; Omaki et al., 2015; Phillips, 2006; Stowe, 1986; Traxler & Pickering, 1996; Wagers & Phillips, 2009). Evidence for the suspension of active gap-filling comes from the absence of both filled-gap effects and plausibility-mismatch effects inside island environments.

There is disagreement about why active gap-filling is suspended in island domains. Some researchers argue that filler-gap dependencies into islands are blocked by grammatical constraints and that knowledge of those constraints guides active gap-filling routines (Phillips, 2006; Wagers & Phillips, 2009). We refer to this as the *grammar-based* view. An alternative holds that islands are not grammatical in origin but instead arise because of limitations on domain-general resources like memory usage during incremental processing (Deane, 1991; Hofmeister & Sag, 2010; Kluender & Kutas, 1993; Pritchett, 1991). Under this account, active gap-filling is suspended in islands because it is simply too taxing for the language processor to pursue. We refer to these proposals as *simple processing* accounts.

If both approaches can account for empirical data, processing accounts would be preferred on parsimony grounds as grammar-based theories require more domain-specific knowledge. It is difficult to tease grammar-based and simple processing explanations apart in English (though see Phillips, 2006), but there are other languages for which the two accounts make divergent predictions. In some languages like Norwegian, which we investigate in this paper, the set of domains that are considered islands differs from the set in English. Insofar as variation in islands exists across languages, grammar-based accounts predict variation in active gap-filling routines: if a particular domain is not an island for the language in question, active gap-filling should be operative in that domain in that language. By contrast, processing-based accounts predict greater cross-linguistic uniformity in active gap-filling: if a particular domain (such as embedded questions) is an island in English, then active gap-filling should be suspended in analogous domains in other languages.

To test these predictions, we investigate whether native Norwegian speakers actively posit gaps inside embedded polar questions ("whether"-questions). In English, embedded polar questions are generally considered islands, as evidenced by the fact that native English-speaking participants judge as unacceptable sentences with *wh* and RC dependencies that cross into "whether"-questions (Pham, Covey, Gabriele, Aldosari, & Fiorentino, 2020; Sprouse, Wagers, & Phillips, 2012; Sprouse, Caponigro, Greco, & Cecchetto, 2016). An example of a *wh*-dependency that crosses into an embedded polar question is given in (5).

(5) *What$_i$ do you wonder [whether John bought __$_i$]?

Proponents of simple processing accounts have argued that embedded questions are islands because their increased complexity overburdens the processor (Hofmeister & Sag, 2010; Kluender & Kutas, 1993). An embedded polar question, for example, requires that the processor consider at least two alternatives: the affirmative and the negative version of the embedded proposition. Holding these alternatives in memory is thought to interfere with the maintenance of an unintegrated filler in sentences like (5), thereby resulting in an island effect (Hofmeister & Sag, 2010).

Unlike in English, native Norwegian speakers have been shown to accept filler-gap dependencies into embedded questions in formal judgment studies (Kush, Lohndal, & Sprouse, 2018, 2019; Kobzeva, Sant et al., 2022; Kush & Dahl, 2022) and produce them in elicitation tasks (Kush, Dahl, & Lindahl, 2024). Naturally occurring examples are also attested in corpora (Kush, Sant, & Strætkvern, 2021). Examples are given in (6). In both examples, an RC head *signalet* ("the signal") or *unge* ("child") is linked to a subject gap inside an embedded question.

(6) a. Det var signalet$_i$ som sjømennene ikke visste [$_{\text{EQ}}$ hva __$_i$ betydde.]
It was signal.DEF REL sailors.DEF NEG knew what meant
lit. "That was the signal$_i$ that the sailors didn't know what$_k$ __$_i$ meant __$_k$."
≈ "That was the signal that the sailors didn't know the meaning of." (Kush et al., 2024)

b. Han var en sånn unge$_i$ som du ikke skjønner [$_{\text{EQ}}$ om __$_i$ er lei seg …]
He was a such child REL you NEG understand whether is upset SELF
lit. "He was the kind of child$_i$ that you don't know if __$_i$ is upset…"
≈ "He was the kind of child that you don't know if *he* is upset" (Kush et al., 2021)

Most work on active gap-filling into islands has investigated whether comprehenders posit gaps inside *subject islands* (Omaki & Schulz, 2011; Phillips, 2006; Traxler & Pickering, 1996), but one recent eye-tracking study suggests that active gap-filling is similarly suspended inside English embedded questions (Cokal & Sturt, 2022). Taking this result at face value, we ask whether Norwegian active gap-filling behaves differently. In Experiment 1, a self-paced reading study, we specifically test whether Norwegian participants actively posit gaps in embedded questions as they do in embedded declarative clauses. To preview our results, we find filled-gap effects in both embedded questions and embedded declarative clauses, suggesting that our Norwegian-speaking participants actively posit gaps in both domains.

Experiment 1 establishes where Norwegians actively posit gaps but leaves open the relatively independent questions of how active gap-filling is implemented and how the incremental parser chooses which grammatical analyses to favor or pursue. Historically, researchers have cast active gap-filling in terms of predictive structure building within a serial processing framework (Fodor, 1978; Omaki et al., 2015; Stowe, 1986; Wanner & Maratsos, 1978). According to this view, when processing an input string, comprehenders are assumed to pursue one analysis at a time and to eagerly adopt analyses that allow syntactic commitments to be discharged as quickly as possible. When holding an unintegrated filler, the drive to complete the dependency quickly leads the processor to predictively commit to an analysis in which the corresponding gap occupies the nearest upcoming position allowed by the grammar. Difficulty ensues when the bottom-up input contradicts the predicted analysis, and the language processor must *revise/reanalyze*.

Although it has not been widely discussed in the literature, modern *expectation-based* processing models (Hale, 2001; Levy, 2008; Wilcox, Pimentel, Meister, Cotterell, & Levy, 2023) offer an alternative way of thinking about "active" dependency resolution. Expectation-based models posit that incremental processing, by default, involves generating expectations about plausible continuations in the face of partial input. Expectations are taken to be probabilistic:

instead of committing to a single analysis or continuation, processors entertain multiple analyses in parallel, weighted relative to their experiential probabilities. No analyses are "actively" committed to, they are simply assigned greater or lesser probability mass. The degree of processing difficulty associated with a particular input correlates with how unlikely that input is given the current probability distribution and how drastically probability mass must be reallocated across the set of parses under consideration in response to that input. Typically, disambiguation difficulty is thought to be captured by word predictability, or the negative log probability of a word in context, also referred to as its *surprisal* (Demberg & Keller, 2008; Hale, 2001; Levy, 2008; Smith & Levy, 2013).

An expectation-based explanation of filled-gap effects is potentially appropriate because filler-gap resolution can be seen as a paradigm case of incremental ambiguity resolution: identifying the true location of a gap requires choosing the correct parse of a sentence from among possible continuations where the gap is in different locations. Under an expectation-based account, comprehenders should posit gaps in syntactic positions where they have encountered them in the past, and the probabilities assigned to such gaps should track their frequency in the input. According to these models, filled-gap effects should arise when continuations with gaps in the next upcoming position are ranked as more probable than continuations where the next position is filled by another NP. The size of the effect should be proportional to the surprisal of the word that signals that the gap is located elsewhere.

In this paper, we consider whether an expectation-based explanation can adequately account for active gap-filling by assessing whether surprisal values accurately predict the location and the magnitude of empirical filled-gap effects we observed in our behavioral data. Under the strongest interpretation of the expectation-based account, filled-gap effects should be entirely reducible to surprisal. To our knowledge, no previous work has tested the predictive psychometric power of surprisal for filled-gap effects, but our work can be seen as contributing to an emerging strand of research that evaluates the qualitative and quantitative fits of word surprisal to specific incremental sentence processing phenomena like garden-path effects (Huang et al., 2024; van Schijndel & Linzen, 2021). Assessing expectation-based models requires having proxies for participants' surprisal values. Such values can be estimated using different statistical language models (Wilcox, Gauthier, Hu, Qian, & Levy, 2020). Recent research has shown that neural language model-derived surprisal is a good broad-coverage predictor of reading behavior (Hoover, Sonderegger, Piantadosi, & O'Donnell, 2023; Shain, 2019; Wilcox et al., 2023), and other studies suggest that neural language models used for these purposes successfully represent the distribution of acceptable gaps in the languages they have been trained on, including Norwegian (Kobzeva, Arehalli, Linzen, & Kush, 2022; Kobzeva, Arehalli, Linzen, & Kush, 2023; Wilcox, Levy, Morita, & Futrell, 2018). To this end, Experiment 2 evaluates how well surprisal values from a recurrent neural language model predict filled-gap effects in the same experimental sentences as were presented to human participants in Experiment 1. To anticipate our conclusion, we find that surprisal successfully predicts the location of filled-gap effects but does a poor job of predicting the degree of processing difficulty associated with filled-gap effects. Inasmuch as there is a large residual difference between surprisal-derived estimates of processing difficulty and the size of the empirical filled-gap effects, it appears that filled-gap effects cannot be reduced to word predictability alone, at least as measured by surprisal from a long short-term memory (LSTM) model. In the General Discussion, we take up how the results could be reconciled with expectation-based models or more traditional models.

2. Experiment 1: Self-paced reading

We tested whether native Norwegian participants actively posit gaps, by looking for filled-gap effects inside two types of complement clauses: embedded declaratives and embedded questions. As we discuss in more detail below, we had participants read sentences in which a filler's true gap site was inside a sentence-final oblique prepositional phrase. Our experiment tested for filled-gap effects at two positions along the path to the true gap site: at the embedded subject and embedded object positions. All three positions are acceptable positions for gaps inside embedded questions, as the naturally occurring examples in (7) indicate.

(7) a. SUBJECT GAP
Har du henrettet noen$_i$ du har tvilt på om __$_i$ var skyldig?
Have you executed someone you have doubted on whether was guilty
lit. "Have you executed someone$_i$ you doubted whether __$_i$ was guilty?"

b. OBJECT GAP
…som gir oss kunnskap$_i$ vi ennå ikke vet hvordan vi skal bruke __$_i$ …
…which gives us knowledge we still NEG know how we shall use …
≈"…which gives us knowledge$_i$ we still don't know how to use __$_i$ …"

c. OBLIQUE GAP
…ofte er det småting$_i$ jeg ikke skjønner hvorfor han lyver om __$_i$.
…often is it small.things I NEG understand why he lies about
lit. "…it's often small things$_i$ that I don't understand why he lies about __$_i$."

2.1. Participants

One hundred forty-eight participants took part in the experiment. Forty-four self-declared native Norwegian speakers were recruited through Prolific and took the experiment online. They were paid 6 GBP for their participation. One hundred four participants were recruited at the Norwegian University of Science and Technology (NTNU) in Trondheim and completed the experiment in person as part of an hour-long testing session that consisted of two experiments (the present experiment, followed by an unrelated acceptability judgment study). The testing session was held on campus with the experimenter present, and the participants were paid 400 NOK for the whole session.

Prior to the experiment, the participants filled in an informed consent form and a background questionnaire. Out of 148 participants, 94 (64%) reported being in the 18–24 age group, 41 (28%) were between 25 and 34 years old, 5 (3%) were between 35 and 44, 7 (5%) between 45 and 54, and 1 was 65 years or older. The participants specified their dialectal background by choosing 1 out of 11 larger dialectal areas. The grouping of the dialects was based on Mæhlum and Røyneland (2012), with an addition of *Bergensk* "Bergen dialect." All of the 11 dialects were represented in the sample, with the majority of participants reporting *Østlandsk* "Eastern Norwegian" as their dialect (66 participants).

2.2. Materials

Twenty-four experimental items were created according to a 2 × 2 design[1] illustrated in (8). Sentences consisted of a main clause containing a subject NP (*Anna*), a verb, and a second NP (*læreren*). An RC was attached to the second NP making that NP a filler. The RC was always composed of two clauses, such that the second clause was embedded below the first. Other aspects of the RC were determined by the factors CLAUSE and DISTANCE. CLAUSE manipulated whether the second clause in the RC was a *Declarative* clause headed by the complementizer *at* "that" or an *Embedded question* headed by the complementizer *om* "whether/if." DISTANCE manipulated the position of the filler's true gap site. In *Short* conditions, the gap was in the subject position in the first clause immediately following the relative pronoun *som*. In *Long* conditions, the gap was in the second clause inside an oblique prepositional phrase, which always followed the direct object.[2]

(8) a. *Short, Declarative*
Anna snakket om læreren som __ visste at rektoren skjelte ut den late studenten foran klassen.
Anna talked about teacher.DEF REL knew that principal.DEF scolded the lazy student.DEF in.front.of class.DEF
"Anna talked about the teacher who knew that the principal scolded the lazy student in front of the class."

b. *Long, Declarative*
Anna snakket om læreren som hun visste at rektoren skjelte ut den late studenten foran __.
Anna talked about teacher.DEF REL she knew that principal.DEF scolded the lazy student.DEF in.front.of
"Anna talked about the teacher who she knew that the principal scolded the lazy student in front of."

c. *Short, Embedded Question*
Anna snakket om læreren som __ ville vite om rektoren skjelte ut den late studenten foran klassen.
Anna talked about teacher.*def* REL wanted to.know whether principal.DEF scolded the lazy student.DEF in.front.of class.DEF
"Anna talked about the teacher who wanted to know whether the principal scolded the lazy student in front of the class."

d. *Long, Embedded Question*
Anna snakket om læreren som hun ville vite om rektoren skjelte ut den late studenten foran __.
Anna talked about teacher.DEF REL she wanted to.know whether principal.DEF scolded the lazy student.DEF in.front.of
"Anna talked about the teacher who she wanted to know whether the principal scolded the lazy student in front of."

We identified two critical regions for potential filled-gap effects inside the lowest clause in the RC: the subject NP (*rektoren*, “the principal”) and the first two words of the embedded object NP (*den late*, “the lazy”), which were presented as a single region. Filled-gap effects should manifest as longer RTs in *Long* conditions compared to their *Short* counterparts. We expected filled-gap effects in the *Declarative* conditions, which would manifest as longer RTs in (8-b) in the critical regions compared to (8-a). If participants exhibit similar differences between (8-d) and (8-c), we can conclude that they actively fill gaps inside embedded questions.

2.3. Procedure

Twenty-four target items were distributed across four experimental lists according to a Latin Square design and interspersed among 48 filler sentences in a pseudorandom order such that target items from the same condition never appeared in immediate succession. Filler sentences were roughly matched with the experimental items in length and varied in complexity. They included both “simple” sentences without any filler-gap dependencies, and more complex sentences with a variety of dependency types (embedded questions, RCs, and cataphoric dependencies). The experiment was built using Open Sesame (Mathôt, Schreij, & Theeuwes, 2012; Mathôt & March, 2022) and hosted on a JATOS server (Lange, Kühn, & Filevich, 2015) at NTNU. Participants read all sentences in a stationary/centered region-by-region noncumulative window self-paced reading paradigm (Just, Carpenter, & Woolley, 1982) using the space bar to advance. After each item, participants were presented with a yes-no comprehension question. Correct answers were counterbalanced across items. Online participants used their own computers and took approximately 20 min on average to complete the task. At the start of the experiment, online participants were instructed to turn off any potential distractions (mobile phones, music, television, etc.) and to close all browser windows/applications other than the window running the experiment. Participants tested in person at NTNU also used their own computers and took 25 min on average to complete the experiment.

2.4. Analysis

Before analysis, the data from five participants (two online, three in-person) who reported themselves to be non-native speakers of Norwegian were excluded. Thereafter we excluded six additional participants (one online, five in-person) whose average comprehension question accuracy was 75% or lower. After exclusion, 137 participants remained for analysis. Reaction times under 100 ms and above 3000 ms were excluded (0.7% of all data points). Trials on which a participant answered the comprehension question incorrectly (10.1% of all trials) were also excluded from the analysis. Log-transformed RTs were analyzed with linear mixed-effects models using the *lme4* package (Bates, Mächler, Bolker, & Walker, 2015) in R (R Core Team, 2021). We adopted the standard threshold for significance ($\alpha = 0.05$) and computed p-values using the Satterthwaite approximation of the degrees of freedom in the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017). All models included sum-coded fixed effects of DISTANCE (0.5 *Long*, -0.5 *Short*), CLAUSE (0.5 *Question* and −0.5 *Declarative*), and their interaction. When choosing random effects, we first fit the maximal model, which

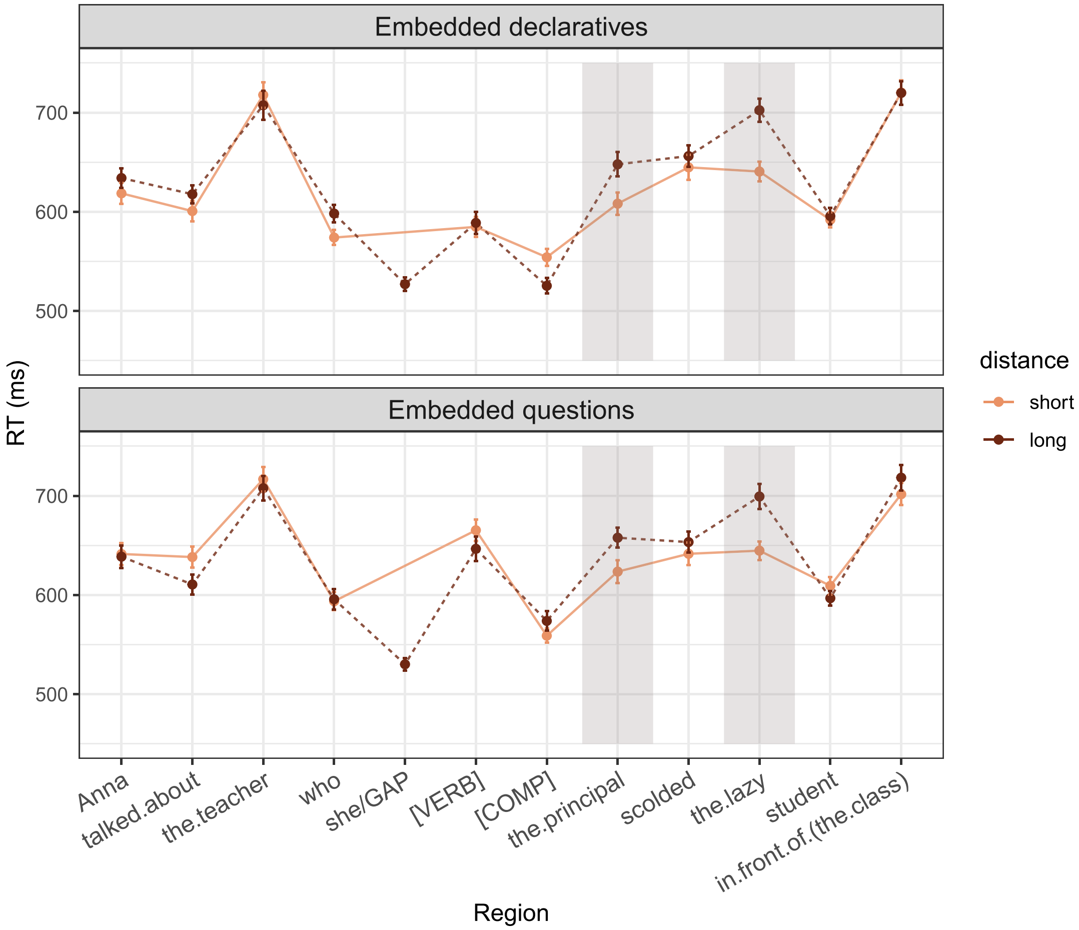


Fig. 1. Empirical RTs by sentence region split by CLAUSE type with *Embedded declaratives* on the top panel and *Embedded questions* on the lower one. Error bars represent the standard error of the mean corrected for between participant variance (Bakeman & McArthur, 1996).

included by-item and by-subject random intercepts and random slopes for all fixed effects and their interaction (Barr, Levy, Scheepers, & Tily, 2013). When the maximal model did not converge or resulted in a singular fit, we simplified the random effects structure following recommendations in Matuschek, Kliegl, Vasishth, Baayen, and Bates (2017). We specify the structure of the final models for each region in footnotes below. Planned comparisons were performed using the *emmeans* package (Lenth, 2023).

2.5. *Results and discussion*

The results are presented in Fig. 1, with the critical regions shaded in gray. Descriptive statistics for RTs in the critical regions are presented in Table 1. In the critical subject region ("the principal"), RTs were longer on average in *Long* conditions than in *Short* conditions,

Table 1
Descriptive statistics for mean RTs corrected for between participant variance in regions that differed between *Embedded declaratives* and *Embedded questions* (EQ), and in the critical regions

Condition	*Short, Declaratives*	*Long, Declaratives*	*Short, EQ*	*Long, EQ*
Region	Mean RT (*SD*)	Mean RT (*SD*)	Mean RT (*SD*)	Mean RT (*SD*)
[VERB]	585 (119)	589 (132)	665 (129)	647 (144)
[COMP]	554 (99)	525 (92)	559 (82)	574 (116)
Subject (the principal)	608 (131)	648 (145)	624 (134)	658 (118)
Obj. det + adj. (the lazy)	641 (117)	702 (137)	645 (108)	699 (149)
Object noun (student)	592 (87)	596 (95)	609 (102)	597 (86)

Table 2
Summary of linear mixed-effects models fit on log-transformed RTs from Experiment 1, split by critical region

Predictor	Critical Region					
	Subject (the Principal)			Object Adjective (the Lazy)		
	Estimate	*S.E.*	*t*	Estimate	*S.E.*	*t*
DISTANCE	**0.04***	0.02	2.06	**0.07*****	0.02	4.14
CLAUSE	**0.04***	0.01	2.61	0	0.01	0.11
DISTANCE × CLAUSE	0	0.03	−0.12	0	0.03	0

Note. $^{*}p < .05$, $^{***}p < .001$

consistent with a filled-gap effect. This numerical trend was observed in embedded declarative clauses (average raw RTs: 648 ms vs. 608 ms) and embedded questions (average raw RTs: 658 ms vs. 624 ms). Similar filled-gap effects are seen at the critical object adjective region ("the lazy") two words downstream: longer average RTs were observed in the *Long* condition compared to the *Short* condition in embedded declarative clauses (average raw RTs: 702 ms vs. 641 ms) and embedded questions (average raw RTs: 699 ms vs. 645 ms).

The output of the linear mixed-effects models is presented in Table 2. At the embedded subject region, the linear mixed-effects model[3] confirmed a main effect of DISTANCE ($p < .05$), but no DISTANCE × CLAUSE interaction. Pairwise comparisons revealed that the numerical trend toward a filled-gap effect was marginally significant in embedded declarative clauses ($t = -1.76, p = .08$) and embedded questions ($t = -1.65, p = .10$). There was also a significant effect of CLAUSE at the embedded subject ($p < .05$) such that subject phrases took longer on average to read in embedded questions than in embedded declaratives.

The linear mixed-effects model[4] confirmed a significant main effect of DISTANCE at the critical object adjective region ($p < .001$), but no other significant effects. Pairwise comparisons showed that filled-gap effects were significant in both embedded declarative clauses ($t = -3.25, p = .002$) and embedded questions ($t = -3.27, p = .002$).

We observed strong evidence for filled-gap effects in object position and slightly weaker evidence for filled-gap effects in subject position in both embedded declarative clauses and embedded questions. Taken together, the results suggest that participants actively predicted gaps in both kinds of embedded clauses, suggesting that from a processing perspective, embedded questions are not islands in Norwegian, in line with judgment studies. Our results argue against simple processing-based accounts that treat embedded questions as islands cross-linguistically due to inherent complexity (Hofmeister & Sag, 2010; Kluender & Kutas, 1993). Instead, the language processor must be sensitive to grammatical generalizations specific to a particular language.

3. Experiment 2: Modeling

To assess whether filled-gap effects can be reduced to surprisal as predicted by an expectation-based account of active dependency resolution, we used a recurrent neural language model with an LSTM architecture (Hochreiter & Schmidhuber, 1997) to estimate the word-by-word surprisal for our experimental stimuli. In line with past studies on processing difficulty in garden-path sentences (Huang et al., 2024; van Schijndel and Linzen, 2021), we explore whether surprisal differences between conditions can predict both the location and the magnitude of filled-gap effects found in the behavioral data.

3.1. Language model

To derive surprisal values, we used an LSTM language model trained on 113 million tokens of the Norwegian Bokmål Wikipedia corpus, as described in Kobzeva, Arehalli et al. (2022) and Kobzeva et al. (2023). The model was trained using the architecture and the procedure described in Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018). It had two layers of 650 hidden units each and a vocabulary comprising the 50,000 most frequent tokens in the Norwegian Bokmål Wikipedia corpus. It was trained for 40 epochs with a batch size of 128, dropout rate of 0.2, and learning rate of 20.0. After training, the model achieved a perplexity of 30.4 on the validation set. The model has been independently shown to predict gaps for corresponding fillers across at least two levels of hierarchical embedding and to distinguish acceptable gap positions from unacceptable gap positions, including inside embedded polar questions (Kobzeva et al., 2023), making it appropriate for our use.

3.2. Converting surprisal into predicted RTs

It has been argued that there is a linear relationship between the surprisal assigned to a word and the difficulty of processing that word as measured by RTs (Levy, 2008; Smith & Levy, 2008, 2013; Shain, 2019; Wilcox et al., 2023). Given this linear relationship, one can derive *surprisal-predicted* RTs—or the portion of overall RTs attributable to surprisal—by multiplying surprisal values by some linear coefficient. Van Schijndel and Linzen (2021) call such coefficients *conversion factors*. Conversion factors can be estimated statistically from empirical RTs within a given experiment (Huang et al., 2024; Smith & Levy, 2013; van Schijndel and Linzen, 2021).

Following the method outlined in van Schijndel and Linzen (2021) and Huang et al. (2024), we estimated the surprisal-to-reading time conversion factor for our test sentences. To do so, we fit linear mixed-effects models to raw empirical RTs from our *filler sentences*.[5] We fit two statistical models to investigate the contribution of surprisal under different assumptions. Both statistical models contained simple predictors of surprisal, mean unigram frequency (calculated from the training corpus in tokens per million and log-transformed), length in characters, the current region's linear position, and the interaction of length and frequency for every region considered in the statistical model.[6] The statistical models differed in how many regions outside the current region were taken into account when predicting RTs. The first statistical model, which we call the SPILLOVER model, included predictors for the current region and the two preceding regions. The inclusion of preceding regions was motivated by past modeling studies (Huang et al., 2024; Monsalve, Frank, & Vigliocco, 2012; Smith & Levy, 2013; van Schijndel and Linzen, 2021), which have consistently found that preceding words' surprisal values influence RTs of subsequent words. In this regard, this model accounts for potential *spillover effects* (Mitchell, 1984) that are commonly observed in self-paced reading studies. When fitting the SPILLOVER model, we excluded all observations for the first two regions of each sentence because they lacked predictors from preceding regions as well as the data from the last region of each sentence as it often displays wrap-up effects (Huang et al., 2024; Smith & Levy, 2013). The second statistical model, the NO-SPILLOVER model, only included predictors for the current region. The decision to fit this model was justified by the observation that filled-gap effects in the behavioral experiment above were clearly localized to our predicted critical regions. The regression equations for the two different statistical models are given below:

(9) a. SPILLOVER MODEL:
$\text{RT} \sim \text{Surprisal}_{R_i} + \text{Surprisal}_{R_{i-1}} + \text{Surprisal}_{R_{i-2}} + \text{Log-Freq}_{R_i} + \text{Log-Freq}_{R_i-1} + \text{Log-Freq}_{R_2} + \text{Length}_{R_i} + \text{Length}_{R_i-1} + \text{Length}_{R_i-2} + \text{Log-Freq} \times \text{Length}_{R_i} + \text{Log-Freq} \times \text{Length}_{R_i-1} + \text{Log-Freq} \times \text{Length}_{R_i-2} + \text{Region number}_i + (1 \mid \text{participant})$

b. NO-SPILLOVER MODEL:
$\text{RT} \sim \text{Surprisal}_{R_i} + \text{Log-Freq}_{R_i} + \text{Length}_{R_i} + \text{Log-Freq} \times \text{Length}_{R_i} + \text{Region number}_i + (1 \mid \text{participant})$

The outputs of the two statistical models are presented in Table 3. In order to calculate the conversion factors, we retained any regression coefficients that were significant at the level of $p < .01$. In the SPILLOVER model, surprisal for the current region and the previous region were significant predictors of RTs in filler sentences. Following the procedure of van Schijndel and Linzen (2021), we used these two coefficients, which we label δ_i and δ_{i-1}, respectively, to calculate the predicted, spillover-adjusted surprisal effect for the current region $\hat{S}_i$ as in (10-a), where S_i is the summed surprisal value for region i. The SPILLOVER model estimates the relevant conversion factors to be $\delta_{i-1} = 2.91$ and $\delta_i = 3.24$. Summed together, this yields a spillover-adjusted conversion factor of around 6 ms per one bit of surprisal. Conversion factors from previous studies in English (Huang et al., 2024; Smith & Levy, 2013; van Schijndel and Linzen, 2021; Wilcox et al., 2023) have ranged from 2 to 4 ms per bit of surprisal, which means that our SPILLOVER model would potentially predict larger effects of surprisal

Table 3
Output of the linear mixed-effects models fit on empirical raw RTs of the filler items. The SPILLOVER model included predictors for the current region and two preceding regions. The NO-SPILLOVER model only included predictors for the current region. Estimates used for deriving conversion coefficients are in bold

Predictor	Spillover Model			No-Spillover Model		
	Estimate	*S.E.*	*t*	Estimate	*S.E.*	*t*
Region number	−9.49***	0.81	−11.64	−9.39***	0.68	−13.78
Surprisal$_{R_i}$	**3.24*****	0.67	4.84	**2.87*****	0.65	4.43
Surprisal$_{R_{i-1}}$	**2.91*****	0.68	4.32			
Surprisal$_{R_{i-2}}$	−0.15	0.63	−0.24			
Log-Freq$_{R_i}$	10.04***	1.50	6.68	9.47***	1.45	6.55
Log-Freq$_{R_{i-1}}$	2.66	1.50	1.78			
Log-Freq$_{R_{i-2}}$	−5.13***	1.53	−3.35			
Length$_{R_i}$	22.00***	1.06	20.73	21.59***	1.04	20.66
Length$_{R_{i-1}}$	9.02***	1.05	8.55			
Length$_{R_{i-2}}$	−0.58	1.06	−0.55			
Log-Freq × Length$_{R_i}$	−1.45***	0.17	−8.37	−1.35***	0.17	−7.97
Log-Freq × Length$_{R_{i-1}}$	−0.47**	0.17	−2.74			
Log-Freq × Length$_{R_{i-2}}$	0.55**	0.18	2.99			

Note. $^{**}p < .01$, $^{***}p < .001$.

than found in these studies. The single coefficient in the NO-SPILLOVER model was $\delta_i = 2.87$, closer to previous estimates. Thus, the surprisal effect for the NO-SPILLOVER model was calculated as in (10-b).

(10) a. SURPRISAL EFFECT, SPILLOVER: $\hat{S}_i = \delta_i \mathrm{S}_i + \delta_{i-1} \mathrm{S}_{i-1}$
b. SURPRISAL EFFECT, NO-SPILLOVER: $\hat{S}_i = \delta_i \mathrm{S}_i$

We now turn to how the two alternatives in (10) predict RTs for our critical regions, as well as the degree of qualitative and quantitative alignment between the LSTM-predicted RTs and human RTs.

3.3. Modeling results

The surprisal-predicted effects on RTs for our test sentences are presented in Fig. 2, with SPILLOVER-predicted RTs and NO-SPILLOVER-predicted RTs plotted side by side. As in Experiment 1, filled-gap effects are defined as longer RTs in *Long* conditions compared to their *Short* counterparts in the critical or spillover regions. Since we observed filled-gap effects directly on the critical subject and object regions, we analyzed predicted RTs in those regions. The analysis used linear mixed-effects models with fixed effects of DISTANCE, CLAUSE, and their interaction. All models contained a by-item random intercept and random slopes for DISTANCE and CLAUSE. The output for all models is presented in Table 4.

For SPILLOVER-predicted RTs, there was an apparent filled-gap effect at the critical subject region in embedded declaratives but not in embedded questions (DISTANCE × CLAUSE inter-

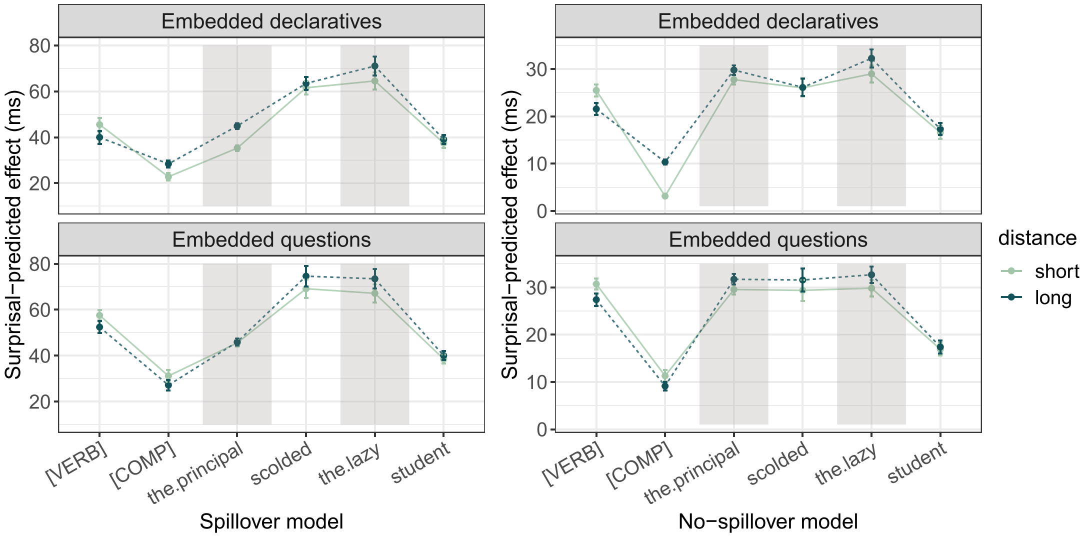


Fig. 2. Region-by-region RTs as predicted by surprisal estimates from the SPILLOVER-adjusted model (left panel) and NO-SPILLOVER model (right panel). Error bars represent 95% confidence intervals of region means.

action $p < .001$). We qualify the effect as "apparent" because the difference between the two conditions is already present in the pre-critical complementizer region. Given the conversion method used, the difference could be driven—at least in part—by spillover from the previous region. We also note that although there was no filled-gap effect immediately at the subject in embedded questions, an effect emerged at the following verb, which could also be driven by spillover. At the critical object region, there was a main effect of DISTANCE ($p < .01$),

Table 4
Output of the linear mixed-effects models fit on predicted RTs for both regions of interest

Predictor	Critical Region					
	Subject (the Principal)			Object Adjective (The Lazy)		
	Estimate	*S.E.*	*t*	Estimate	*S.E.*	*t*
			SPILLOVER MODEL			
DISTANCE	5.06***	0.47	10.79	6.46***	0.82	7.90
CLAUSE	5.69***	1.20	4.74	2.52**	0.80	3.16
DISTANCE × CLAUSE	−9.28***	0.77	−12.00	−0.18	1.17	−0.16
			NO-SPILLOVER MODEL			
DISTANCE	2.10***	0.23	8.96	3.07***	0.45	6.81
CLAUSE	1.87**	0.51	3.66	0.63	0.34	1.85
DISTANCE × CLAUSE	0.13	0.33	0.39	−0.43	0.47	−0.92

Note. $^{**}p < .01$, $^{***}p < .001$.

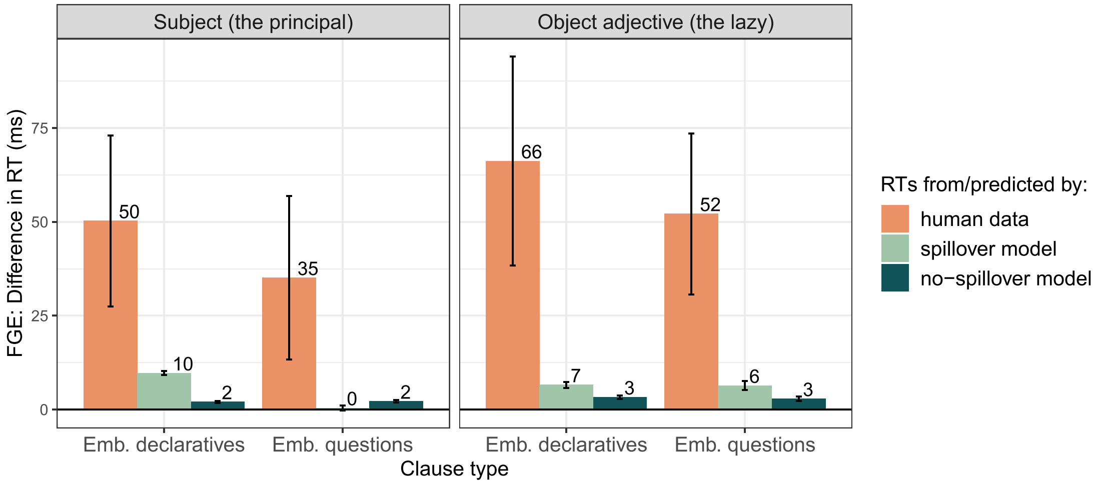


Fig. 3. Difference in filled-gap effects at both critical regions as predicted by the SPILLOVER and NO-SPILLOVER language models compared to empirical filled-gap effects from Experiment 1. Error bars represent 95% confidence intervals of by-item means.

but the DISTANCE × CLAUSE interaction was not significant ($p = .88$), suggesting comparable filled-gap effects between embedded clause types.

Looking at NO-SPILLOVER-predicted RTs, we see significant main effects of DISTANCE in the subject and object regions (both $ps < .001$), with no hint of an interaction. Similar to human RTs, surprisal values predict comparable filled-gap effects between embedded declarative clauses and embedded questions in both subject and object positions.

Before moving on to quantitative comparison, we discuss the insights that the effects found in NO-SPILLOVER-predicted RTs provide regarding the contribution of the previous regions. Regarding the surprisal effect at the embedded subject in declarative clauses, we see a relatively large baseline difference in the complementizer region immediately preceding the subject. This suggests that the apparent filled-gap effect at the subject was amplified by orthogonal spillover from the complementizer region. Spillover may also help explain the absence of a significant subject filled-gap effect in embedded questions: The surprisal differences in the preceding complementizer region go in the opposite direction. Thus, conversion without spillover confirms that the LSTM assigns higher surprisal values to the critical regions themselves in *Long* conditions than in *Short* conditions, which suggests that the misaligned predictions in spillover-adjusted surprisal effects were driven by surprisal differences that were already present in the preceding regions.

Fig. 3 compares the numerical size of the filled-gap effects from Experiment 1 to the predicted filled-gap effects using the SPILLOVER and NO-SPILLOVER conversion methods. As can be seen in the figure, surprisal-based predictions vastly underestimate the size of the empirical filled-gap effects. Empirical filled-gap effects ranged in size from roughly 35–65 ms, whereas surprisal effects ranged from 3 to 10 ms. Approximating the magnitude of the empirical effects would require LSTM-derived surprisal effects on the order of 6–11 bits of surprisal assuming the generous SPILLOVER conversion factor of 6 ms/bit or 13–23 bits of surprisal with the smaller NO-SPILLOVER conversion factor of 2.8 ms/bit.

3.4. Discussion

Stepping back, we see that neural language model-based surprisal values exhibit relatively good qualitative—but poor quantitative—alignment with human behavioral patterns. Qualitatively, the model predicts filled-gap effects in the correct positions in both embedded declaratives and embedded questions. Misalignment was only observed when predicted RTs were adjusted for potential spillover effects.

SPILLOVER-predicted RTs align with human judgment in the critical object region. However, they deviate from the human pattern in the subject region by failing to predict a filled-gap effect in embedded questions. The filled-gap effects at the subject position found in human data were not delayed in embedded questions. There is also reason to suspect that the differences between the declarative conditions at the subject should not be treated as a true filled-gap effect: this difference already appears in the pre-critical complementizer region, which affects the difference in the critical region. It would, therefore, appear to be an orthogonal spillover effect from a difference between the conditions at the complementizer. The suspicion is confirmed by the large differences seen at the complementizer in the NO-SPILLOVER model. NO-SPILLOVER-predicted RTs fare better with respect to qualitative alignment with human RTs, predicting small (2 ms) filled-gap effects at both the subject and object regions in both embedded clause types. Despite qualitative alignment, LSTM-derived surprisal values fail to match human RTs quantitatively, severely underestimating the magnitude of filled-gap effects, in line with previous studies on garden-path effects (Huang et al., 2024; van Schijndel and Linzen, 2021).

4. General discussion

In this paper, we investigated (i) whether native Norwegian speakers exhibit filled-gap effects inside embedded questions and (ii) if surprisal values derived from an LSTM language model trained on Norwegian text could predict the location and magnitude of filled-gap effects. The first question was designed to tease apart predictions of grammar-based and simple processing accounts of why active gap-filling is suspended inside potential island domains cross-linguistically. The second question was intended to test whether explanations for well-established psycholinguistic effects could be subsumed under a general surprisal-based account of processing difficulty. In response to the first question, we found filled-gap effects at subject and object positions inside embedded questions comparable to effects in embedded declaratives. To the second question, we found that surprisal was reasonably successful at predicting where filled-gap effects would arise, but it systematically and drastically underestimated the true size of filled-gap effects. Below, we consider the implications of each finding in turn.

4.1. Active gap-filling in embedded questions

As discussed in the Introduction, simple processing accounts predict that active gap-filling should be suspended inside embedded questions because of inherent processing complexity (Hofmeister & Sag, 2010; Kluender & Kutas, 1993). In contrast, grammar-based accounts predict that active-gap filling should only be suspended in embedded questions in languages like English whose grammar categorizes embedded questions as islands. In Norwegian, embedded questions are not island domains, so active filling should proceed unimpeded. The filled-gap effects that we observed inside embedded questions indicate that participants posited gaps actively inside embedded questions, as in embedded declarative clauses. Thus, our results are consistent with grammar-based accounts.

We have, at a minimum, shown that active gap-filling occurs inside Norwegian embedded polar questions. One question that remains is whether active gap-filling inside embedded questions is only seen in languages where such dependencies are clearly acceptable, or whether similar effects would be observed in languages where filler-gap dependencies into embedded questions are ruled out or at least degraded. If embedded questions are grammatical islands in English, the strong interpretation of grammar-based accounts predicts that there should be no active gap-filling into embedded questions in English. A complication with drawing this conclusion, though, is that dependencies into embedded questions are occasionally attested in colloquial English and that speaker intuitions suggest that such dependencies do not always feel as unacceptable as other island violations. This is especially true of particularly embedded polar questions headed by *whether* and *if*, analogous to the Norwegian structures we tested here (though judgments from formal acceptability judgment studies seem to affirm the presence of island effects even in these constructions; see Sprouse et al. 2012). As yet, only a few studies have directly addressed the question of active gap-filling inside embedded questions in English, and these studies have provided conflicting or inconclusive results. In an eye-tracking study, Cokal and Sturt (2022) measured how having an unresolved filler (*magazine* in (11)) affected participants' responses to gaps or pronouns inside prepositional phrases (after *about*) depending on whether the gap/pronoun was inside a declarative complement clause or an embedded question.

(11) a. Pronoun/that complement
This is the magazine$_i$ that Jane said that the hairdresser had talked **about it** before going to the salon.

b. Gap/that complement
This is the magazine$_i$ that Jane said that the hairdresser had talked **about** __$_i$ before going to the salon.

c. Pronoun/embedded question
This is the magazine$_i$ that Jane wondered whether the hairdresser had talked **about it** before going to the salon.

d. Gap/embedded question
This is the magazine$_i$ that Jane wondered whether the hairdresser had talked about **about** __$_i$ before going to the salon.

The authors found significantly longer regression path duration on the spillover region (*before*) in (11-a) than in (11-b), indicating a filled-gap effect caused by the pronoun. They failed to find a corresponding difference between (11-c) and (11-d), which they interpreted as an absence of a filled-gap effect. Based on these findings, the authors concluded that participants suspended active gap-filling inside the embedded *whether*−question. However, the authors' use of a resumptive pronoun to fill the object gap position complicates interpretation. The results are compatible with an alternate interpretation under which participants actively posit the foot of the dependency in the object position, but do not discriminate between a gap and a resumptive pronoun as an acceptable completion. Thus, the results are equivocal.

Villata, Tabor, and Sprouse (2020) used forced-choice and maze tasks to test if participants posit gaps inside English embedded *whether*-questions. The researchers found that when forced to choose a continuation after the object *solved* in sentences like (12), participants preferred the option that entailed a gap (*before*) over an alternative continuation that did not (*the*). The authors interpreted this preference as evidence that participants could fill gaps inside the embedded question.

(12) Which puzzle did you wonder whether the candidate solved ...{*before* | *the*}

We agree that the findings show that participants are able to complete the filler-gap dependency when presented with an explicit choice. But they do not, to our mind, conclusively demonstrate that English-speaking participants actively fill gaps in these domains. The findings are equally consistent with the conclusion that participants only posited a gap when they were explicitly presented with the option. Thus, while their findings are suggestive, determining whether there is clear evidence for cross-linguistic differences in gap-filling inside embedded *whether*-questions, or embedded questions more generally, remains to be tested empirically.

While we have interpreted our results as supporting grammar-based accounts, we consider what it would take to accommodate the findings within a processing framework. How processing models could account for our results depends, in part, on whether *whether*-questions block active filling. On the one hand, if active filling is possible in English *whether*-questions, proponents of simple processing accounts could maintain that embedded polar questions headed by *whether/if* and *om* are less taxing to process than other embedded questions. One reason polar questions might be easier to process is that the question words do not themselves trigger an additional filler-gap dependency. If that were the case, we might expect that active gap-filling would be stopped in argument or adjunct embedded questions (headed by *who/what* or *where/when/how*, respectively) in both languages. Given that such embedded questions are not islands in Norwegian either (see (7-b) above), we suspect that active gap-filling would not be suspended at least in that language. On the other hand, if there is a cross-linguistic difference between Norwegian and English with respect to active filling in polar questions, a simple processing account would have to hold that polar questions are inherently less complex or costly in Norwegian than in English. At present, we see little reason to suppose that this could be the case, given the high degree of structural similarities between the two languages and the standard ways of conceptualizing complexity.

Finally, we do not wish to suggest that grammar alone determines whether comprehenders can actively or successfully compute filler-gap dependencies. It is clear that memory limitations or interference can negatively impact dependency resolution (Gibson & Thomas, 1999; Keshev & Meltzer-Asscher, 2019). Moreover, semantic and pragmatic considerations seem to be at play. Anecdotally, the acceptability of structurally identical sentences with dependencies into embedded questions seems to vary depending on the selection of embedding verbs, arguments, and other factors. At present, these influences are not well understood but should be systematically analyzed.

4.2. Expectation-based explanations of active dependency resolution

Our findings suggest that filled-gap effects cannot be reduced to a simple effect of surprisal derived by an LSTM language model. Surprisal was able to predict the location of filled-gap effects, which suggests that filled-gaps are indeed less predictable than actual gaps. However, a full reduction of filled-gap effects to surprisal must meet a more stringent requirement than merely localizing effects: Values should fully account for the magnitude of those effects.

A strong interpretation of our results is that filled-gap effects cannot be reduced to surprisal or be understood as a simple consequence of expectation-based processing. Under this interpretation, a supplemental mechanism is required to explain filled-gap effects. Following the (often implicit) assumptions of early work on active gap-filling within a serial parsing framework, a candidate mechanism would be *reanalysis*. A reanalysis mechanism has been independently argued to account for processing difficulty associated with garden-path effects (Fodor & Inoue, 1994; Gorrell, 1995; Huang et al., 2024; Jurafsky, 1996; Pritchett, 1988; Sturt, 1997, 1996; van Schijndel and Linzen, 2021), where the process usually involves finding an alternative structure for the sentence that is drastically different from the initial parse, as in the case of the classic main verb/reduced-RC ambiguity (Bever, 1970). When it comes to filled-gap effects, the amount of structural reanalysis required would be considerably smaller; the parser would only need to change the initial representation by replacing a predicted gap with an observed phrase (and posit the gap in a subsequent position, which may be relatively predictable). While our results are consistent with a reanalysis-based interpretation of filled-gap effects, it is important to point out that (i) they do not provide direct positive support for such an explanation and (ii) violated expectations could, in principle, cause processing difficulty for other reasons in a serial processing account.

A narrower interpretation would be that the problem lies not with surprisal per se, but rather with the specific values derived from the class of LSTM language models we used. According to this line of thinking, the LSTM-derived surprisal values are a poor proxy for true human expectations as reflected in RTs (Oh, Clark, & Schuler, 2022), which would in fact be sufficient to correctly predict the effect size. We consider some reasons why our LSTM-derived surprisal estimates might underestimate empirical RTs.

The misalignment between LSTM-derived surprisal and human RTs could have arisen because the Wikipedia corpus we used to train our model was not representative of our participants' experience. For example, the corpus might lack examples of long-distance dependencies into embedded declaratives and embedded questions entirely. We can, however, rule out this extreme possibility as Kobzeva et al. (submitted) show that a non-exhaustive search of the Wikipedia training corpus uncovered 33 examples of filler-gap dependencies into embedded questions, 10 of which were examples into embedded polar questions. Below are two attested examples.

(13) Det kan dreie seg om misjonærer$_i$ som man er usikker på [$_{\text{EQ}}$ om faktisk __$_i$ kom dit] …
It can revolve around missionaries REL one is uncertain whether actually came there …
lit. "It can revolve around missionaries$_i$ who one is uncertain whether __$_i$ actually came there …"
≈ "These may be the missionaries who one is uncertain whether **they** actually arrived there …"[7]

(14) …og å gi ham komplimenter$_i$ han er usikker på om han fortjener __$_i$.
…and to give him compliments he is uncertain whether he deserves
"…and to give him compliments$_i$ that he is uncertain if he deserves __$_i$."[8]

Of course, our training corpus may underrepresent the relative frequency of the constructions in question relative to everyday experience. There is reason to believe that the estimated frequencies would not differ substantially if other written corpora were used: for example, Kush et al. (2021) found that long-distance filler-gap dependencies into embedded declaratives and embedded questions were relatively rare in child and adolescent literature. However, we cannot rule out the possibility that surprisal values derived from a corpus that included dialogue and more informal daily writing alongside the text from Wikipedia would provide better estimates.

Differences between LSTM-derived surprisal and human expectations could potentially arise if the model is just worse at representing long-distance filler-gap dependencies than humans are. Consistent with this possibility, Kobzeva et al. (2023) found that although the LSTM can predict gaps in embedded environments, surprisal values decreased with each additional level of embedding. It seems unlikely that human expectations for gaps would be similarly attenuated with increased embedding. Even though this factor probably contributes to the differences somewhat, we note that the surprisal values that the same model assigns to filled-gap effects in shorter distance dependencies usually did not exceed 3–4 bits of surprisal, which is still too small to account for the size of our effects given the conversion factors that we used.

Finally, our model might underpredict filled-gap effect sizes because it underestimates effects of more abstract *syntactic* predictability (i.e., is the object of the upcoming verb likely to be a *gap* or an NP) or expectations at other levels of analysis. The model we used estimated the predictability of individual words in a context. Arehalli, Dillon, and Linzen (2022) show that such *lexical* surprisal values correlate with but do not fully account for, *syntactic* surprisal that estimates the probability of the phrasal category of the next word. Incorporating LSTM-derived estimates of syntactic surprisal as predictors into our calculation of effect size could improve the model's performance. We leave testing this possibility to future work but note that Arehalli et al. (2022) found that even when such predictors were taken into account, LSTM-derived surprisal values still substantially undershot empirical garden-path effect sizes.

5. Conclusion

We found that Norwegian participants actively fill gaps inside two types of embedded clauses: embedded declaratives and embedded questions. Our results rule out simple processing accounts that hold that active gap-filling is sometimes suspended inside potential island domains due to inherent complexity. We also tested whether word predictability (as measured by surprisal values from an LSTM neural language model) can account for the processing difficulty posed by violated expectations as measured by filled-gap effects. Our results show that RTs derived from surprisal predict the direction and location of filled-gap effects, but severely underestimate their magnitude. This suggests either that the cost of filled-gap effects cannot be reduced to simple effects of predictability or that LSTM-derived surprisal values are not adequate proxies for human expectations during incremental filler-gap resolution.

Acknowledgments

We thank the participants of the 29th Architectures and Mechanisms for Language Processing Conference (AMLaP 2023) and the members of the AcqVA group at NTNU for their comments on earlier versions of this work. We also thank Tal Linzen, Suhas Arehalli, and Brian Dillon for their feedback on the manuscript, and Ingrid Bondevik for collaboration on behavioral data collection. Dave Kush is supported in part by funding from the Social Sciences and Humanities Research Council of Canada. All materials and analysis scripts are available in the following OSF repository: https://osf.io/me5xh/?view_only=47a61519a6f94ff389ad43974bedb456.

Notes

[1] Test items were written by a native Norwegian-speaking research assistant according to a template or written by the authors and checked by the same assistant for grammaticality. All materials were subsequently checked for grammaticality and parseability by three different native Norwegian-speaking members of the NTNU community.

[2] The full set of experimental items, alongside filler sentences and analysis scripts can be found at the following OSF repository: https://osf.io/me5xh/?view_only=47a61519a6f94ff389ad43974bedb456.

[3] Model formula: logRT $\sim$ DISTANCE×CLAUSE + (1+DISTANCE+CLAUSE | participant) + (1+DISTANCE | item).

4 Model formula: logRT ∼ DISTANCE×CLAUSE + (1+DISTANCE | participant) + (1+DISTANCE | item).

5 Not all the tokens from the filler sentences were in the LSTM model's vocabulary, so we excluded 11 fillers that had three or more unknown tokens, resulting in 37 filler sentences remaining for the analysis. Fillers with one or two unknown tokens were slightly modified to include only in-vocabulary tokens.

6 Unlike van Schijndel and Linzen (2021), we did not include terms for entropy or entropy reduction in any region, since the authors found them to be much poorer predictors of garden-path effects than surprisal (cf. Aurnhammer & Frank, 2019, for a similar conclusion for next word entropy).

7 'Liste over kinamisjonærer tilhørende jesuittordenen' Wikipedia page

8 'Knøttene' Wikipedia page

References

Aoshima, S., Phillips, C., & Weinberg, A. (2004). Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language*, *51*(1), 23–54.

Arehalli, S., Dillon, B., & Linzen, T. (2022). Syntactic surprisal from neural models predicts, but underestimates, human processing difficulty from syntactic ambiguities. In *Proceedings of CONLL 2022* (pp. 301–313). Stroudsburg, PA: Association for Computational Linguistics.

Atkinson, E., Wagers, M. W., Lidz, J., Phillips, C., & Omaki, A. (2018). Developing incrementality in filler-gap dependency processing. *Cognition*, *179*, 132–149.

Aurnhammer, C., & Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia*, *134*, 107198.

Bakeman, R., & McArthur, D. (1996). Picturing repeated measures: Comments on Loftus, Morrison, and others. *Behavior Research Methods, Instruments & Computers*, *28*, 584–589.

Barr, D. J., Levy, R., Scheepers, C., & 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., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bever, T. (1970). The cognitive basis for linguistic structure. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley

Cokal, D., & Sturt, P. (2022). The real-time status of strong and weak islands. *PLoS ONE*, *17*(2), e0263879.

Deane, P. (1991). Limits to attention: A cognitive theory of island constraints. *Cognitive Linguistics*, *2*, 1–63.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Fodor, J. D. (1978). Parsing strategies and constraints on transformations. *Linguistic Inquiry*, *9*(3), 427–473.

Fodor, J. D., & Inoue, A. (1994). The diagnosis and cure of garden paths. *Journal of Psycholinguistic Research*, *23*, 407–434.

Frazier, L., & Flores d'Arcais, G. B. (1989). Filler driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language*, *28*(3), 331–344.

Gibson, E., & Thomas, J. (1999). Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, *14*(3), 225–248.

Gorrell, P. (1995). *Syntax and parsing*, Cambridge Studies in Linguistics, volume 76. New York: Cambridge University Press.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the North American Chapter of the Association for Computational Linguistics 2018* (pp. 1195–1205). Stroudsburg, PA: Association for Computational Linguistics.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics* (pp. 1–8). Stroudsburg, PA: Association for Computational Linguistics.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415.

Hoover, J. L., Sonderegger, M., Piantadosi, S. T., & O'Donnell, T. J. (2023). The plausibility of sampling as an algorithmic theory of sentence processing. *Open Mind*, *7*, 350–391.

Huang, K.-J., Arehalli, S., Kugemoto, M., Muxica, C., Prasad, G., Dillon, B., & Linzen, T. (2024). Large-scale benchmark yields no evidence that language model surprisal explains syntactic disambiguation difficulty. *Journal of Memory and Language*, *137*, 104510.

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science*, *20*(2), 137–194.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, *111*(2), 228.

Keshev, M., & Meltzer-Asscher, A. (2019). A processing-based account of subliminal wh-island effects. *Natural Language & Linguistic Theory*, *37*, 621–657.

Kluender, R., & Kutas, M. (1993). Bridging the gap: Evidence from ERPs on the processing of unbounded dependencies. *Journal of Cognitive Neuroscience*, *5*(2), 196–214.

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2022). LSTMs can learn basic Wh- and relative clause dependencies in Norwegian. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol, 44, pp. 2974–2980). Austin, TX: Cognitive Science Society.

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (2023). Neural networks can learn patterns of island-insensitivity in Norwegian. In *Proceedings of the Society for Computation in Linguistics* (Vol. 6, pp. 175–185). Amherst, MA: Association for Computational Linguistics.

Kobzeva, A., Arehalli, S., Linzen, T., & Kush, D. (Manuscript submitted for publication). Learning filler-gap dependencies with neural language models: Testing island sensitivity in Norwegian and English.

Kobzeva, A., Sant, C., Robbins, P. T., Vos, M., Lohndal, T., & Kush, D. (2022). Comparing island effects for different dependency types in Norwegian. *Languages*, *7*(3), 195–220.

Kush, D., & Dahl, A. (2022). L2 transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research*, *38*(2), 315–346.

Kush, D., Dahl, A., & Lindahl, F. (2024). Filler–gap dependencies and islands in L2 English production: Comparing transfer from L1 Norwegian and L1 Swedish. *Second Language Research*, *40*(3), 739–763.

Kush, D., Lohndal, T., & Sprouse, J. (2018). Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory*, *36*(3), 743–779.

Kush, D., Lohndal, T., & Sprouse, J. (2019). On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language*, *95*(3), 393–420.

Kush, D., Sant, C., & Strætkvern, S. B. (2021). Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: A Journal of General Linguistics*, *6*(1), 1–50.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

Lange, K., Kühn, S., & Filevich, E. (2015). "Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLoS ONE*, *10*(6), e0130834.

Lee, M.-W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *Journal of Psycholinguistic Research*, *33*, 51–73.

Lenth, R. V. (2023). *emmeans: Estimated marginal means, aka least-squares means*. R package version 1.8.7.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Mæhlum, B., & Røyneland, U. (2012). *Det norske dialektlandskapet: Innføring i studiet av dialekter* [The Norwegian dialect landscape: An introduction to the study of dialects]. Oslo, Norway: Cappelen Damm Akademisk.

Mathôt, S., & March, J. (2022). Conducting linguistic experiments online with OpenSesame and OSWeb. *Language Learning*, *72*(4), 1017–1048.

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*, 314–324.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

McKinnon, R., & Osterhout, L. (1996). Constraints on movement phenomena in sentence processing: Evidence from event-related brain potentials. *Language and Cognitive Processes*, *11*(5), 495–524.

Mitchell, D. C. (1984). An evaluation of subject-paced reading tasks and other methods for investigating immediate processes in reading 1. In *New Methods in Reading Comprehension Research* (pp. 69–90). London: Routledge.

Monsalve, I. F., Frank, S. L., & Vigliocco, G. (2012). Lexical surprisal as a general predictor of reading time. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, (pp. 398–408). Avignon, France: Association for Computational Linguistics.

Oh, B.-D., Clark, C., & Schuler, W. (2022). Comparison of structural parsers and neural language models as surprisal estimators. *Frontiers in Artificial Intelligence*, *5*, 777963.

Omaki, A., Lau, E. F., Davidson White, I., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology*, *6*, 384.

Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. *Studies in Second Language Acquisition*, *33*(4), 563–588.

Pham, C., Covey, L., Gabriele, A., Aldosari, S., & Fiorentino, R. (2020). Investigating the relationship between individual differences and island sensitivity. *Glossa: A Journal of General Linguistics*, *5*(1), 1–17.

Phillips, C. (2006). The real-time status of island phenomena. *Language*, *82*(4), 795–823.

Pritchett, B. L. (1988). Garden path phenomena and the grammatical basis of language processing. *Language*, *64*(3), 539–576.

Pritchett, B. L. (1991). Subjacency in a principle-based parser. In *Principle-based parsing: Computation and psycholinguistics* (pp. 301–345). Berlin: Springer.

R Core Team (2021),. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ross, J. R. (1967). *Constraints on variables in syntax* [Doctoral dissertation]. MIT, Cambridge, MA.

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. In Burstein, J., Doran, C., & Solorio, T. (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 4086–4094). Minneapolis, MN: Association for Computational Linguistics.

Smith, N. J., & Levy, R. (2008). Optimal processing times in reading: A formal model and empirical investigation. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 30, pp. 595-600). Red Hook, NY: Curran Associates Inc.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.

Sprouse, J., Caponigro, I., Greco, C., & Cecchetto, C. (2016). Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, *34*, 307–344.

Sprouse, J., Wagers, M., & Phillips, C. (2012). A test of the relation between working-memory capacity and syntactic island effects. *Language*, 82–123.

Stowe, L. A. (1986). Parsing wh-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227–245.

Sturt, P. (1996). Monotonic syntactic processing: A cross-linguistic study of attachment and reanalysis. *Language and Cognitive Processes*, *11*(5), 449–494.

Sturt, P. (1997). *Syntactic reanalysis in human language processing* [Doctoral dissertation]. University of Edinburgh. College of Science and Engineering.

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475.

van Schijndel, M., & Linzen, T. (2021). Single-stage prediction models do not explain the magnitude of syntactic disambiguation difficulty. *Cognitive Science*, *45*(6), e12988.

Villata, S., Tabor, W., & Sprouse, J. (2020). Gap-filling in syntactic islands: Evidence for island penetrability from the maze tasks. In *The 33rd Annual CUNY Conference on Human Sentence Processing* (pp. 19–21). Ann Arbor, MI: Society for Human Sentence Processing.

Wagers, M., Borja, M. F., & Chung, S. (2015). The real-time comprehension of wh-dependencies in a wh-agreement language. *Language*, *91*(1), 109–144.

Wagers, M., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, *45*(2), 395–433.

Wanner, E., & Maratsos, M. P. (1978). An ATN approach to comprehension. In *Linguistic theory and psychological reality* (pp. 119–161). Cambridge, MA: MIT Press.

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (pp. 1707–1713). Austin, TX: Cognitive Science Society.

Wilcox, E. G., Levy, R., Morita, T., & Futrell, R. (2018). What do RNN language models learn about filler-gap dependencies? In *Proceedings of the 2018 EMNLP Workshop Blackbox NLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 211–221). Brussels, Belgium: Association for Computational Linguistics.

Wilcox, E. G., Pimentel, T., Meister, C., Cotterell, R., & Levy, R. P. (2023). Testing the Predictions of surprisal theory in 11 languages. *Transactions of the Association for Computational Linguistics*, *11*, 1451–1470.

Article J2

Kobzeva, A. & Kush, D. (2025). Acquiring Constraints on Filler-Gap Dependencies from Structural Collocations: Assessing a Computational Learning Model of Island-insensitivity in Norwegian. *Language Acquisition.* DOI: 10.1080/10489223.2024.2440340

LANGUAGE ACQUISITION
https://doi.org/10.1080/10489223.2024.2440340

Routledge
Taylor & Francis Group

Acquiring constraints on filler-gap dependencies from structural collocations: Assessing a computational learning model of island-insensitivity in Norwegian

Anastasia Kobzeva[a] and Dave Kush[b]

[a]Norwegian University of Science and Technology (NTNU); [b]University of Toronto

ABSTRACT
Children induce complex syntactic knowledge from their native language input. A long-standing discussion focuses on types of learning biases that help them arrive at correct generalization and solve induction problems posed by impoverished input. Studies employing computational models for learning specific language phenomena serve as testing grounds for evaluating types of biases required for successful acquisition. Recent work by Pearl & Sprouse (2013b) demonstrates that a distributional learner that tracks trigrams over structurally annotated input can acquire *wh*-filler-gap dependencies and island constraints on them in English. Though intriguing, it is unclear yet whether a similar distributional learning model is a viable mechanism for learning island facts in other languages given the possibility of cross-linguistic variation. In this study, we explore whether a distributional learner can acquire *wh*- and relative clause filler-gap dependencies and island constraints in Norwegian from child-directed annotated text. We find that the proposed learning strategy can capture some patterns of island-insensitivity in Norwegian while failing to learn others due to a lack of relevant data in the input. Our findings suggest that given limited input data, a simple n-gram-based distributional learning over structured representations may not be sufficient to fully recover human-like knowledge of filler-gap dependency relations and island constraints cross-linguistically.

ARTICLE HISTORY
Received 28 Nov 2023
Accepted 28 Nov 2024

1. Introduction

Children induce complex syntactic knowledge from their native language input. A long-standing question is how much language acquisition relies on domain-general knowledge and abilities versus domain-specific components. One standard view among generative linguists is that the input lacks (sufficient) direct evidence to guarantee the acquisition of certain facts through domain-general knowledge and procedures alone. According to such Poverty of the Stimulus (POS) arguments (Chomsky 1986; Crain & Pietroski 2001; Laurence & Margolis 2001), domain-specific linguistic knowledge and mechanisms scaffold the acquisition process by helping children navigate the hypothesis space of possible grammars (see Pearl 2022 for a useful review). In contrast, there are those who hold that the input is not as impoverished as has been previously assumed (Pullum & Scholz 2002; Clark & Lappin 2010; Chater et al. 2015) and that the relevant complex grammatical facts can be acquired using domain-general knowledge and procedures alone.

Most theorists agree that it would be preferable to minimize the number of domain-specific components that are postulated. However, the quantity and complexity of necessary domain-specific content are determined, in part, by what domain-general components are available to

CONTACT Anastasia Kobzeva anastasia.kobzeva@ntnu.no Department of Language and Literature, Norwegian University of Science and Technology (NTNU), Edvard Bulls veg 1, Trondheim 7049, Norway.

This article has been corrected with minor changes. These changes do not impact the academic content of the article.

the child. Therefore, formulating specific hypotheses about domain-specific content that is necessary (if any) can be done relative to explicit proposals for what kinds of domain-general learning procedures humans actually use and what types of representations those procedures operate over. Case studies of specific language phenomena serve as testing grounds.

In this article, we use the acquisition of *island constraints* on filler-gap dependencies to investigate the above issues. Islands have been central in syntactic theorizing since their discovery by Ross (1967). As we discuss below, island constraints represent a classic POS problem and are taken by some to motivate complex domain-specific biases (Chomsky 1973, 1986; Huang 1982). More recently, Pearl & Sprouse proposed an explicit computational learning algorithm that can induce the "knowledge" of islands in English by tracking trigrams over syntactically parsed child input (Pearl & Sprouse 2013a, 2013b; Pearl & Bates 2022). Importantly, although the learning algorithm does not entirely eliminate the need for domain-specific assumptions, it reduces the quantity and complexity of domain-specific knowledge relative to traditional proposals.

We continue this general line of research by asking whether the distributional learning strategy introduced by Pearl & Sprouse can be used to learn island constraints on two types of filler-gap dependencies—*wh*-dependencies and relative clause (RC) dependencies—in Norwegian. Norwegian represents an excellent case for testing whether variable patterns of judgments can be induced from the input, as it allows dependencies into environments that are considered islands in English. We test whether a modeled learner that tracks n-grams over structured input can capture patterns of island-insensitivity in Norwegian, while inducing the island status of other domains. Additionally, we train our modeled learner on input encoded in a different representational formalism than used by Pearl & Sprouse, which allows us to test whether the success of the modeled learner is formalism-dependent.

The article is structured as follows. The rest of this section introduces island constraints on filler-gap dependencies and provides a short overview of computational models proposed to capture islands. It also motivates our focus on evaluating one computational learning model: the structural trigram-based distributional learner of Pearl & Sprouse (2013b). Section 2 describes the modeled learner of Pearl & Sprouse in detail. In Section 3, we describe how the idea behind the modeled learner can be applied to Norwegian data annotated using the lexical functional grammar (LFG) formalism and how differences between the formalisms allow us to evaluate bigram- and trigram-based learner implementations. Section 4 presents the modeling results alongside human acceptability data for four potential island environments in Norwegian. To preview our results, we find that while the proposed learning strategy can capture some patterns of island-insensitivity in Norwegian when direct evidence is present in the input, it fails to learn other important patterns because the relevant dependencies were unattested in our training corpus. We also find that the LFG formalism has the representational capacity to discriminate between "island-violating" and licit dependencies in the same manner as the phrase structure representation used in previous work and that there is a trade-off between formalism complexity and n-gram window size. Section 5 discusses the modeling results and their implications for the acquisition of island constraints in general and this distributional learning strategy in particular. Section 6 concludes.

1.1. *Filler-gap dependencies and island constraints*

One of the core features of human languages is that they allow dependencies between two non-adjacent items in a sentence. One example is filler-gap dependencies (FGDs) where a displaced *filler* phrase is interpreted in a position later in the sentence. An example of an FGD is the *wh*-question in

(1), where the filler *wh*-word 'what' is interpreted as a direct object of the verb 'bought.' The canonical object position is empty, leaving a *gap*, which we mark with an underscore throughout the article.[1]

(1) What does Olav think that Astrid bought __?

Apart from *wh*-questions like (1), constructions like relative clauses (2-a) and topicalizations (2-b) involve long-distance FGDs as well.

(2) a. Olav likes the dress that Astrid bought __.

b. That dress, Astrid bought __.

Two important generalizations characterize FGD formation. First, FGDs are potentially *unbounded* (Chomsky 1965): Fillers can be related to gaps across an arbitrary linear or hierarchical distance (subject to memory limitations), as demonstrated in (3).

(3) What dress did Nora say (that) Olav thought (that) Tor knew (that) Astrid bought __?

Second, though unbounded, FGDs are nevertheless constrained. Attempting to associate gaps in certain environments with fillers outside those environments results in unacceptability. Following Ross (1967), constituents that block FGDs are called *islands*. In a language like English, many different structures have been identified as islands, including subject phrases (4-a), embedded questions (4-b), adjuncts (4-c), and relative clauses (4-d).

(4) a. *What is [that Astrid bought __] extremely unlikely?

b. *What did Olav wonder [whether Astrid bought __]?

c. *What does Olav worry [if Astrid buys __]?

d. *What did Olav think it was Astrid [who bought __]?

Across languages, children reliably acquire grammars that permit unbounded dependencies and respect island constraints. Yet, the acquisition of islands presents a POS problem: The input to the learner is equally compatible with the hypothesis that island constraints exist and with the hypothesis that they do not. The mere absence of a particular set of FGDs does not necessarily entail that such dependencies are unacceptable. Consider example (3) above: Triply embedded dependencies are unlikely to appear in a learner's input, yet adults judge them as acceptable. Thus, children must generalize beyond the set of dependencies that they have seen. If children generalize beyond their input in the case of unboundedness, what prevents them from concluding that FGDs into islands are also possible?

One hypothesis, proposed in the generative tradition, is that innate language-specific constraints on syntactic operations block dependencies into islands. Examples include the *Subjacency Condition* of Chomsky (1973) and its successors (*Barriers*: Chomsky 1986; the *Phase Impenetrability Condition* of Chomsky 2001) and the *Condition on Extraction Domains* of Huang (1982).[2] These constraints limit

[1] We use the term 'filler-gap dependency,' which was coined by movement-based generative theories that assume that the extracted element is fronted, leaving a 'trace' in its base position in the sentence structure (Chomsky 1973). Trace-less theories (e.g., head-driven phase structure grammar: Pollard & Sag 1994; LFG: Kaplan & Bresnan 1995; or construction grammar: Goldberg 1995) assume a "gap-less" dependency between the filler and the head (e.g., the verb *bought* in 1.1). In this article, we do not take a strong stand on the existence of traces in the constituent structure and simply use the gap notation for ease of exposition.

[2] A class of *functionalist* theories represents a contrasting view according to which island effects arise because of a clash of information-related properties of the filler and/or the gap site (Erteschik-Shir 1973; Goldberg 2006; Abeillé et al. 2020; Cuneo & Goldberg 2023). Such accounts question the assumption that island effects are syntactic in nature and emphasize the importance of discourse-pragmatic factors in explaining why island-crossing dependencies are infelicitous (rather than ungrammatical). In this article, we set these accounts aside and explore whether island effects can be learned from syntactic distributions alone.

children's hypothesis space when it comes to the range of acceptable FGDs. For example, Chomsky's Subjacency Condition (1973) states that children come equipped with an innate grammar that prevents dependencies from crossing more than one bounding node. The precise definition of a bounding node may vary from language to language to account for patterns of cross-linguistic variation (Rizzi 1982). In English, NP (DP) and IP (S or TP) were considered bounding nodes.

Subjacency was introduced to explain several island effects simultaneously, including *wh*-, subject, and complex noun phrase islands. For example, the *wh*-dependency in (5), repeated from (4-b) has two intervening IP nodes, which are bounding nodes in English, making dependencies into *whether*-clauses ungrammatical.

(5) *What did [$_{\text{IP}}$ Olav wonder [$_{\text{CP}}$ whether [$_{\text{IP}}$ Astrid bought __]]]?

Constraints like Subjacency limit the hypothesis space that learners explore during the acquisition process, preventing them from considering such FGDs as possible. Additional constraints, required to rule out other island types, limit the hypothesis space further.

Committing to more constraints means assuming increasingly complex domain-specific biases to account for island acquisition. Such complexity prompts the question of whether it is possible to learn about islands using fewer domain-specific biases, less complex domain-specific biases, or perhaps no domain-specific biases at all. Recent work has turned to computational modeling to test this hypothesis. A growing body of research explores how algorithms that rely more heavily on domain-general statistical learning fare on the acquisition of islands from real language data.

A very recent line of research has used neural language models (NLMs) to test whether it is possible for statistical learners with weak domain-general biases to learn about islands in a given language when trained on raw, unannotated text (Chowdhury & Zamparelli 2018; Wilcox et al. 2018, 2021; Kobzeva, Arehalli et al. 2022; Kobzeva et al. 2023; Lan et al. 2024; Howitt et al. 2024).[3] Despite some similarities between the acquisition of syntactic rules by NLMs and children (Evanson et al. 2023), their relevance to human language acquisition (and cognition more broadly) is a topic of ongoing debate (Warstadt & Bowman 2022; Katzir 2023; Kodner et al. 2023; Piantadosi 2023; Cuskley et al. 2024). One line of criticism holds that NLMs are less informative for understanding and exploring possible human acquisition trajectories because they do not provide interpretive models of acquisition at Marr's (1982) *algorithmic* level (see Dunbar 2019 for discussion). Among other features that make NLMs less ideal models of human language acquisition include (i) backpropagation algorithms used to train NLMs, which make it challenging to identify parameters relevant to the learning process (McCloskey 1991); (ii) the nature of the representations learned by NLMs, which is not yet properly understood (Dunbar 2019); and (iii) the fact that NLMs are typically trained on large amounts of data, often from sources like Wikipedia, news, and web data, which differ qualitatively from child-directed speech (though see Warstadt & Bowman 2022; Warstadt et al. 2023). Though we do not deny that such models are still useful for proof-of-concept simulations and empirical explorations of learnability issues more broadly (Wilcox et al. 2021), and in the future it is possible that NLMs trained on developmentally plausible input enriched with data from other domains could give us deeper insights into human language development (Warstadt & Bowman 2022), for the current article we are interested in evaluating a proposal for a simple, interpretable, algorithmic-level procedure that operates over theoretically motivated linguistic representations—the distributional learner of Pearl & Sprouse (2013b).

Pearl & Sprouse (2013b) proposed a simple and explicit distributional learning algorithm for inducing island constraints from realistic amounts of *annotated* child input in tandem with fewer domain-specific learning biases than previously assumed (Pearl & Sprouse 2013a, 2013b; Pearl & Bates 2022). The guiding idea of their acquisition model is that a learner can acquire knowledge of the full range of acceptable dependencies in a language by tracking the probability of the structural "building blocks" that make up

[3] We note that although early work on NLMs learning FGDs in Norwegian suggested that NLMs could successfully mimic human knowledge (Kobzeva et al. 2023), more recent work probing the behavior of the models suggests that more caution is in order when drawing conclusions about model abilities (Kobzeva et al. submitted).

dependencies in the input. Unlike NLMs, the modeled learner is transparent for interpretation: It is easy to identify why and how the model succeeds or fails because all the components that go into the model, their parameters, and features are known to the experimenter. The model simply tracks co-occurrence frequencies over a limited window while operating over well-motivated phrase structure representations. Pearl & Sprouse argued that such a model can successfully approximate target knowledge of islands for English *wh*-dependencies (Pearl & Sprouse 2013a, 2013b). Here, we explore whether a modeled learner that implements the same general idea can learn the set of acceptable filler-gap dependencies in Norwegian—a language with a different set of island constraints than in English.

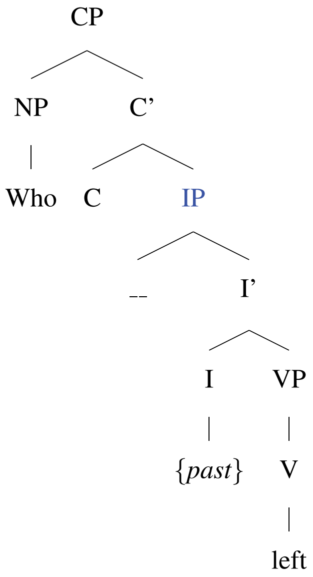
2. Pearl & Sprouse's computational learner

Pearl & Sprouse (2013b) proposed an explicit model of how "knowledge" of island constraints could be induced from child-directed input. The modeled learner works by tracking the co-occurrence frequency of syntactic features in attested *wh*-dependencies, which it can then use to estimate the probability of a novel dependency. Probability is used as a proxy for grammaticality: Learning is considered successful if grammatical dependencies are assigned significantly higher probabilities than ungrammatical dependencies, such as those that violate island constraints.

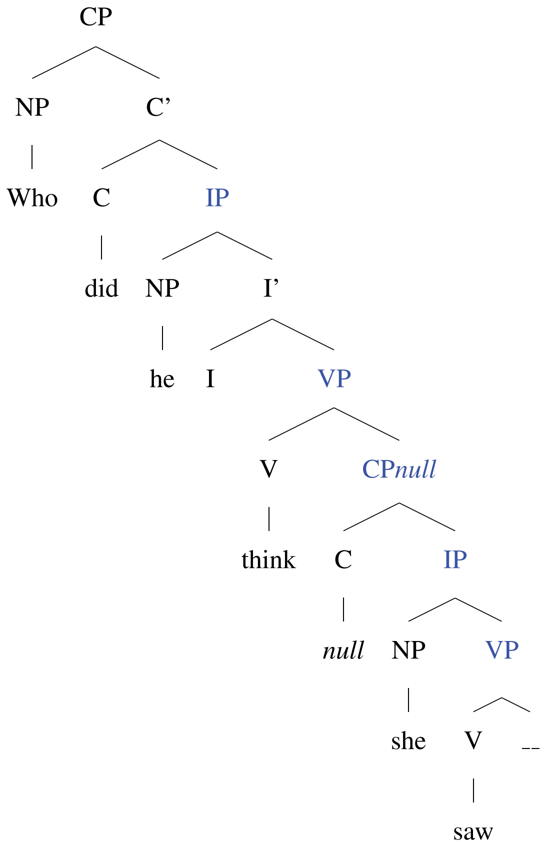
2.1. *The distributional learning algorithm*

The modeled learner starts by assuming that learners can parse input sentences into hierarchical syntactic representations. Pearl & Sprouse's modeled learner uses a phrase structure (PS) notation roughly equivalent to Penn Treebank notation (Marcus et al. 1993) supplemented with gaps. It also assumes that the representation of FGDs encodes a relation between the surface position of a filler and its corresponding gap, such that a *path* can be established between the two. The path between a filler and its gap is represented as the sequence of maximal projections that dominate the gap but not the filler. These projections are called *container nodes* and the path is referred to as a *container node sequence*. In the simple subject question (6-a), the only container node intervening between the filler *who* and its gap in the specifier of IP is the IP node, whereas in the long-distance object question (6-b) the container node sequence is IP-VP-CP$_{\text{null}}$-IP-VP, as also shown below with container nodes in blue.

(6) a. [$_{\text{CP}}$ Who [$_{\text{IP}}$ __ left]]?



b. [$_{\text{CP}}$ What did [$_{\text{IP}}$ he [$_{\text{VP}}$ think [$_{\text{CP}}$ [$_{\text{IP}}$ she [$_{\text{VP}}$ saw __]]]]]]?



According to this modeling approach, the probability of an FGD corresponds to the probability of its corresponding container node sequence. However, rather than tracking container node sequences as indivisible wholes, Pearl & Sprouse proposed that the learner break down the sequences into *container node trigrams*. To accommodate shorter dependencies into the trigram framework, and to track dependency boundaries, START and END nodes are appended to each sequence. Under this model, the overall probability P of any FGD can be calculated as the product of the probabilities of all container node trigrams along the path from filler to gap, as shown in (7):

(7) a. [$_{\text{CP}}$ What did [$_{\text{IP}}$ he [$_{\text{VP}}$ think [$_{\text{CP}_{null}}$ [$_{\text{IP}}$ she [$_{\text{VP}}$ saw __]]]]]]?

b. Sequence: START-IP-VP-VP$_{null}$-IP-VP-END

c. Trigrams:

START-IP-VP

IP-VP-CP$_{null}$

VP-CP$_{null}$-IP

CP$_{null}$-IP-VP

IP-VP-END

d. P(START-IP-VP-VP$_{null}$-IP-VP-END) = P(START-IP-VP) * P(IP-VP-CP$_{null}$) * ... * P(IP-VP-END)

Treating the probability of a given FGD as the product of trigram probabilities allows the modeled learner to generalize beyond the exact set of FGDs that it has seen during training. The modeled learner will assign a non-zero probability to any FGD whose container node sequence is entirely composed of previously seen trigrams. Grammatical dependencies that are made up of attested trigrams will be recognized as possible FGDs. Any sequence with one or more unattested trigrams, by contrast, should have a probability of zero because at least one of the trigrams being multiplied has zero probability. In practice, Pearl & Sprouse adopted a smoothing procedure that assigns a count of 0.5 to unattested trigrams, making their probability very low instead of zero.

How does the modeled learner distinguish island violations from acceptable dependencies? If island violations are unattested in the input, then their container node sequences should contain one or more unattested trigrams. Therefore, the probability assigned to the island violation should be much smaller than the probability of an acceptable comparison sentence. As an illustration, consider how the modeled learner could distinguish (8-a), which is a case of *wh*-island violation, from (8-b), which is an acceptable dependency.

(8) a. [$_{\text{CP}}$ What did [$_{\text{IP}}$ he [$_{\text{VP}}$ think [$_{\text{CP}_{that}}$ that [$_{\text{IP}}$ she [$_{\text{VP}}$ saw __]]]]]]?

b. *[$_{\text{CP}}$ What did [$_{\text{IP}}$ he [$_{\text{VP}}$ wonder [$_{\text{CP}_{whether}}$ whether [$_{\text{IP}}$ she [$_{\text{VP}}$ saw __]]]]]]?

Under the assumption that CP container nodes are annotated with lexical information about the complementizer head (CP$_{null}$, CP$_{that}$, CP$_{whether}$, CP$_{if}$), (8-b) will be ruled out because it contains the trigram VP-CP$_{whether}$-IP. This trigram should be unattested if the modeled learner has never seen a dependency into a *whether*-clause. (8-a) should be allowed because the trigram VP-CP$_{that}$-IP will be attested if the learner has ever seen a dependency into an embedded declarative clause.

2.2. *Input data and target state*

To train their learner on data representative of a child's input, Pearl & Sprouse assembled a set of 31,247 *wh*-dependencies drawn from the Brown (1973), Valian (1991), and Suppes (1974) corpora (together totaling 813,036 word tokens) of the CHILDES database (MacWhinney 2000). The corpora contained child-directed speech to 25 children from the age of 1;6 to 5 years old. Before training, the utterances were automatically parsed using Penn Treebank style annotation and then manually checked by human annotators who also inserted appropriate gaps into the tree representations.

To assess a computational acquisition model, one needs a representation of the target state that reflects the knowledge that children aim to achieve during language development. As this knowledge is not directly observable, researchers use behavioral data like acceptability judgments as a proxy. Ideally, child judgments collected at an age matched to the audience of the training corpus would be used. However, the relevant judgment studies have not been done (though see De Villiers et al. 2008 for 4-to-9-year-olds' judgments on English islands), so adult judgments are used as the "final" developmental target for the modeled learner.

As their target state, Pearl & Sprouse used acceptability judgment ratings of different *wh*-dependencies from acceptability judgment experiments collected in Sprouse et al. (2012). For each dependency tested, they compared the average human judgments to the probabilities that the modeled learner assigned to the same dependencies. Mapping the probabilities that the modeled learner assigns to dependency paths onto adult acceptability judgments is based on the assumption that grammaticality is a major (though not the only) factor underlying acceptability ratings (we return to this assumption in Section 5.3).

The judgment experiments followed the design for isolating and measuring island effects introduced by Sprouse (2007). Sentences containing FGDs are constructed according to a 2 × 2 design crossing the DISTANCE between a filler and its gap and STRUCTURE, which manipulates whether the test sentence contains an island structure or not. Applying the factorial design to test for *whether*-island effects yields an example experimental item set in (9).

(9) a. **Short, No Island**

Who$_i$ __$_i$ thinks that Olav stole the necklace?

b. **Short, Island**

Who$_i$ __$_i$ wonders whether Olav stole the necklace?

c. **Long, No Island**

What$_i$ does the detective think that Olav stole __$_i$?

d. **Long, Island**

What$_i$ does the detective wonder whether Olav stole __$_i$?

In (9), DISTANCE modulates whether the *wh*-filler *who* is associated with a gap in the matrix clause (*Short*) or the embedded clause (*Long*). STRUCTURE manipulates whether the embedded clause is a declarative complement clause (*no island*) or an embedded *whether*-question (*island*). The FGD in the *Long, island* condition corresponds to an 'island violation' as the filler is associated with a gap located inside an embedded question.

The design defines an island effect as a superadditive DISTANCE × STRUCTURE interaction that arises when (i) the *Long, island* condition is rated significantly lower than all other conditions and (ii) the unacceptability of the *Long, island* condition cannot be attributed to the simple effects of dependency length and structural complexity alone. The residual decrement in unacceptability is the island effect. The presence or absence of an island effect can be visually assessed on interaction plots as in Figure 1. If dependency length and structural complexity are sufficient to explain the relative unacceptability of the *Long, island* condition, there is no island effect, and plotting the ratings to the four conditions should yield two parallel lines. On the other hand, if there is an island effect, we should see two non-parallel lines, with the lowest ratings for *Long, island* condition.

Pearl & Sprouse (2013b) found good qualitative alignment between the human judgment patterns and the probabilities assigned to test sentences by the modeled learner. The modeled probabilities showed the same superadditive pattern as the human judgments for four different types of islands in English (Adjunct, Whether, Complex NP, and Subject phrases). *Long, island* FGDs received significantly lower probabilities than FGDs in the three non-island comparison conditions, because all of the *Long, island* container node sequences contained at least one unattested trigram. The authors thus concluded that the modeled learner successfully represents the adult target knowledge of island constraints in English.

Pearl & Sprouse's results demonstrate that a distributional learning algorithm can induce island constraints from realistic English input with the help of some supplemental biases: The model's need for

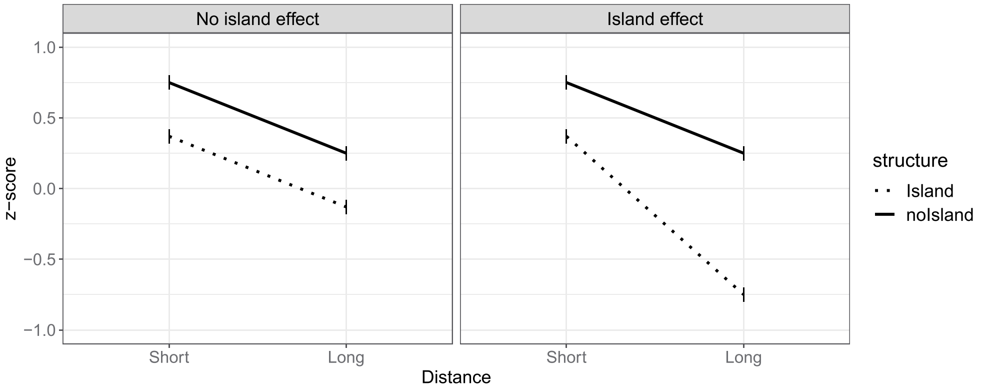


Figure 1. Visual definition of island effects: The left panel represents the absence of an effect, and the right panel represents a pattern when the island effect is present.

parsed input presupposes biases to assign hierarchical structure to strings and to apply labels to those structures from a given set. It also has biases that seem specific to FGDs: it tracks trigrams of container node sequences. Despite these assumptions, the model succeeds with fewer domain-specific biases than are traditionally assumed necessary because it is not endowed with specific constraints on FGD formation.

The success of the learner in this individual case invites follow-up questions about the general feasibility of the proposed algorithm. The first set of questions has to do with whether distributional learning is a viable strategy for acquiring islands given the possibility of cross-linguistic variation. It is known that some languages allow dependencies into constituents that are islands in English. In such languages, is there sufficient input for the distributional learner to acquire a different set of facts, or is there a poverty of the stimulus? The second set of questions has to do with how much this success depends on the idiosyncratic assumptions of Pearl & Sprouse's modeled learner. Are the results dependent on tracking trigrams of phrase structure nodes in a particular formalism, or is there an equivalence class of representational formats that would support learning these facts? If so, what types of information must the representation minimally encode to guarantee successful acquisition?

In this paper, we seek to answer the questions above by applying the idea behind the distributional learner of Pearl & Sprouse to Norwegian data annotated within LFG formalism.

3. Distributional learning and Norwegian islands

3.1. *The Norwegian target state*

Norwegian exhibits sensitivity to a subset of island constraints that English is sensitive to. Similar to other Mainland Scandinavian languages like Swedish and Danish, Norwegian also allows FGDs into some domains that are typically considered islands in English (Christensen 1982; Engdahl 1982, 1997; Christensen et al. 2013; Christensen & Nyvad 2014; Lindahl 2017; Kush et al. 2021; Müller & Eggers 2022). As a representation of the Norwegian target state, we use the results of a recent acceptability judgment study run by Kobzeva, Sant et al. (2022) that used the factorial design outlined above to test the acceptability of *wh*- and RC-dependencies from four different domains: subject phrases, embedded questions, relative clauses, and (conditional) adjunct clauses.[4] Subject island and adjunct island items were identical in structure to items used to test English subject and adjunct islands in Pearl & Sprouse (2013b) and similar to the items used in previous studies on Norwegian (Kush et al. 2018, 2019; Bondevik et al. 2021). Kobzeva, Sant et al. (2022) used different items to test the status of embedded questions in Norwegian from those used by Pearl & Sprouse. Where Pearl & Sprouse tested object *wh*-extraction from embedded *whether*-questions, Kobzeva, Sant et al. (2022) tested subject extraction from embedded constituent questions with RC-dependencies like (10) and *wh*-dependencies.

(10) Det var signalet$_i$ som sjømennene visste hva __$_i$ betydde.
That was signal.DEF RP fishermen.DEF knew what meant
'That was the signal$_i$ that the fishermen knew what __$_i$ meant.'

Kobzeva, Sant et al. (2022) also tested object extraction from existential relative clauses, an environment Pearl & Sprouse (2013) did not test. (11) is an example test sentence that contains a *wh*-dependency into a relative clause.[5]

(11) Hvilken bok$_i$ var det mange som likte __$_i$?
Which book were that many RP liked
'Which book$_i$ were there many people who liked __$_i$?'

[4] An example item set for each domain can be found in Appendix A.

[5] There is experimental evidence that such sentences are less degraded in English compared to other island types but some degree of cross-linguistic differences still persists (Kush et al. 2013; Christensen & Nyvad 2014; Vincent 2021; Müller & Eggers 2022).

Table 1. Summary of human judgment patterns from Kobzeva, Sant et al. (2022).

Constituent	*Wh*-dependencies	RC-dependencies
Subject phrases	Island effect	Island effect
Relative clauses	No effect	Island effect
Embedded questions	Island effect	No effect
Adjunct clauses	Island effect	No effect

Kobzeva, Sant et al. found clear island effects for *wh*- and RC-dependencies into subject islands, consistent with previous findings (Kush et al. 2018; Kush & Dahl 2020; Bondevik et al. 2021). They also found that island effects for the remaining environments differed by dependency type, as have others (Sprouse et al. 2016; Kush et al. 2018, 2019; Abeillé et al.2020; Bondevik et al. 2021; Bondevik & Lohndal 2023). First looking at relative clauses, the authors found island effects for RC-dependencies but not *wh*-dependencies. With embedded questions and conditional adjunct clauses, however, the authors found island effects with *wh*-dependencies but not with RC-dependencies. Interaction plots from Kobzeva, Sant et al. are presented alongside our model results for visual comparison in the Results section, but a summary of the human judgment patterns is given in Table 1.

In the absence of other assumptions, the modeled learner must come to distinct generalizations for each dependency type if it is to capture the cross-dependency variation in the Norwegian target state. To achieve this, we hard-coded our model to track frequencies for the two dependencies separately, which could be considered domain-specific prior knowledge. We return to this issue in the Discussion.

3.2. *The input*

3.2.1. *Input source*

To date, there are unfortunately no large-scale annotated corpora of child-directed speech in Norwegian that would be comparable to the corpus used by Pearl & Sprouse (2013b). As an approximation of child-directed input, we used the Norwegian children's fiction corpus (nob-child) from NorGramBank (Dyvik et al. 2016). The children's fiction corpus contains texts from 155 children's books, spanning various genres from picture books to young adult novels, which are aimed at children and youth between the ages of 3 to 18 (for a more detailed description of the corpus, including a breakdown of the number of different filler-gap dependencies by age group, see Kush et al. 2021).

It should be noted that there are important differences between our corpus and the corpus of Pearl & Sprouse, which stem from their different modalities. Child-directed texts tend to exhibit greater syntactic complexity compared to child-directed speech (Cameron-Faulkner & Noble 2013; Montag & MacDonald 2015; Mesmer 2016; Montag 2019), including more elaborate sentence structures with subordinate clauses, including relative clauses. Consequently, our analysis may yield inflated frequencies of such structures relative to their occurrence in child-directed speech. On the other hand, our corpus may underrepresent questions and embedded questions compared to child-directed speech that frequently employs questions to engage children and encourage interaction (Noble et al. 2018).

The corpus consists of 389,557 sentences and 4,111,213 tokens in total (roughly five times larger than the corpus of child-directed speech used by Pearl & Sprouse). The corpus was automatically parsed using the Norwegian Bokmål computational grammar NorGram within the LFG formalism described in the following section.[6] We queried the corpus to find all instances of *wh*- and RC-dependencies. For *wh*-dependencies, the query included all types of interrogative phrases followed by a verb (in order to exclude fragments like 'Then what?'). To find RC-dependencies, we ran a query specifying the clause type (clause-type *rel*). Because the corpus was automatically parsed, most sentences have multiple parses and only a small fraction have been disambiguated (less than 1% of the data). When downloading search results, only the first (most probable)

[6] The corpus can be queried with the help of the Infrastructure for the Exploration of Syntax and Semantics (INESS) (Rosén et al. 2009). Parsed sentences stored in the Tiger-XML treebank encoding format can be downloaded from the INESS project web page: https://clarino.uib.no/iness-prod/home.

parse was included in the data set. From the search results, 19,716 *wh*- and 41,220 RC-dependencies were automatically extracted. We then applied an algorithm to identify the path between the filler and the gap. For a portion of sentences, several possible paths were found due to difficulties stemming from automatic parsing. We manually corrected the paths for these sentences. The algorithm also identified false hits—sentences that did not contain an FGD—which were subsequently excluded from the data set. Finally, we identified remaining complex paths and manually corrected any errors. In total we manually checked 9,105 RC-dependency paths (22%) and 2,749 *wh*-dependency paths (14%).[7] After exclusions and corrections, 19,004 *wh*-dependencies and 40,702 RC-dependencies remained as input to the learner.

3.2.2. *Input representation: LFG*

Sentences in the input were parsed within the LFG formalism (Kaplan & Bresnan 1995; Dalrymple 2001), according to which a sentence is assigned distinct constituent structure (c-structure) and a functional structure (f-structure) representations.

A sentence's c-structure is a tree generated by context-free phrase structure rules where only whole lexical words serve as terminal nodes (Falk 2011; Bresnan et al. 2015; Dalrymple et al. 2019). A sentence's f-structure is a directed graph of attributes and values that represents the abstract functional organization of the sentence and serves as an interface between syntax and semantics. F-structure encodes predicate-argument structure, hierarchical functional relations (subject, object, complement, adjunct, etc.), as well as other features relevant to sentence interpretation.

Figure 2 illustrates the c-structure and f-structure for a simple transitive sentence in (12). The f-structure encodes that the predicate (PRED) of the clause *drink* has two arguments: *they* and *tea*, each of which bears an index. Each argument is assigned a functional role (SUBJ and OBJ, respectively) as a feature of the clause. Each argument's morphosyntactic features are stored in an attribute value matrix embedded under the functional role labels. Each matrix is assigned an index (5 and 2 for the subject and object matrices) that can be used as a pointer. Clause-level functional information such as

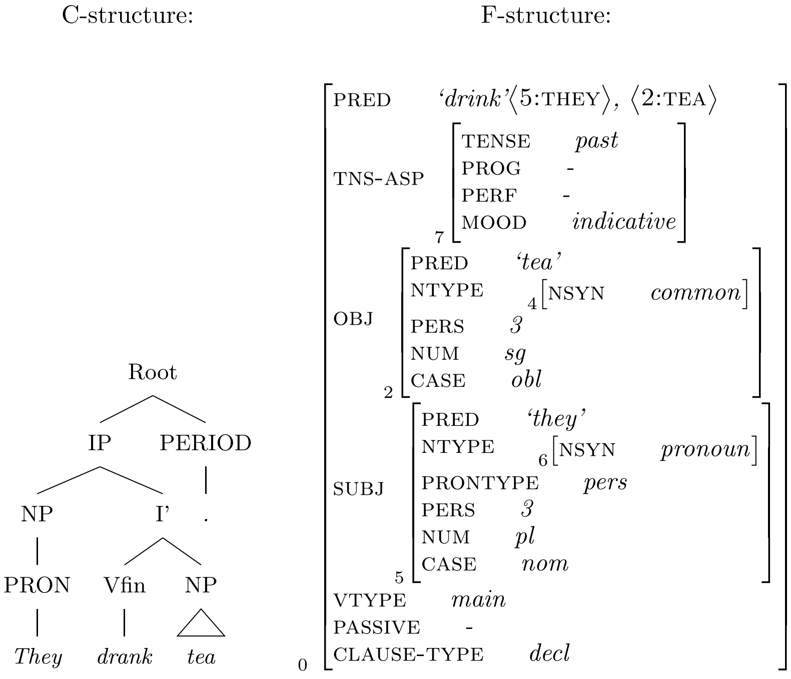


Figure 2. LFG representation for (12).

[7] We note that despite our efforts, a small degree of noise and some misparses may persist.

CLAUSE-TYPE (*declarative*), voice information, etc., are encoded as features at the same level of embedding as the grammatical role assignments in that clause.

(12) They drank tea.

FGDs are represented differently in LFG than in transformational frameworks like the one used by Pearl & Sprouse. In transformational theories, fillers are linked to a trace in the constituent structure. In LFG, it is generally assumed that empty categories are not represented in the c-structure.

(13) Hvem drakk te?
Who drink.PAST tea
'Who drank tea?'

Instead, FGDs are represented in the f-structure (Kaplan & Zaenen, 1989). To illustrate, consider the f-structure of a simple Norwegian sentence in (13) given in Figure 3.
In question (13), the *wh*-filler *hvem* 'who' bears interrogative focus, so it is labeled as FOCUS-INT of the matrix clause in the f-structure. To determine how it is interpreted, the filler must be assigned a grammatical function, in this case the SUBJ of the same clause. In the corpus, this linking relation is represented by using the *wh*-word's index, 2 as the value of SUBJ.[8] We defined the FGD path as the series of hierarchically nested labels between a boxed index value and the f-structure level bearing the filler's discourse functional feature. This path can be calculated iteratively. To illustrate: In Figure 3, beginning at 2, step up to SUBJ. Since SUBJ is at the same level as the FOCUS-INT bearing the same index, the path is complete.

The same iterative procedure can be used to calculate longer paths through (potentially unbounded) layers of embedding. In (14), the *wh*-filler *hva* 'what' is interpreted as the object of the embedded verb

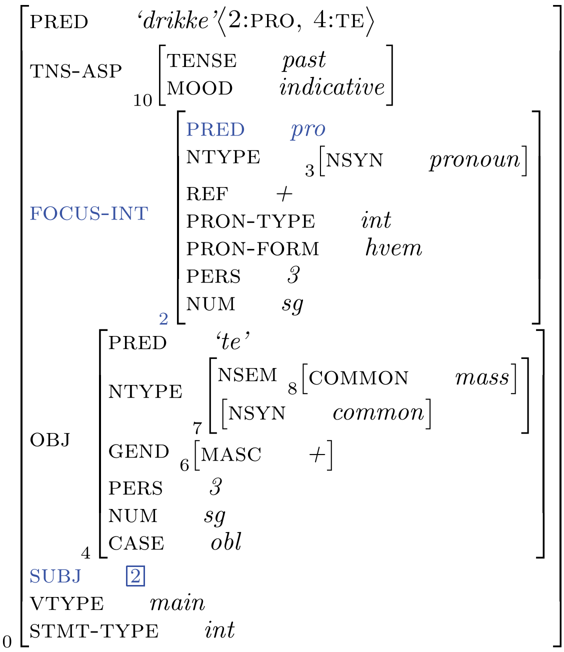


Figure 3. LFG f-structure for (13).

[8] The 'gap,' so to speak, in the matrix predicate's argument structure is represented as PRO, which is co-indexed with the filler and corresponds to the *pro* value for the filler's PRED attribute.

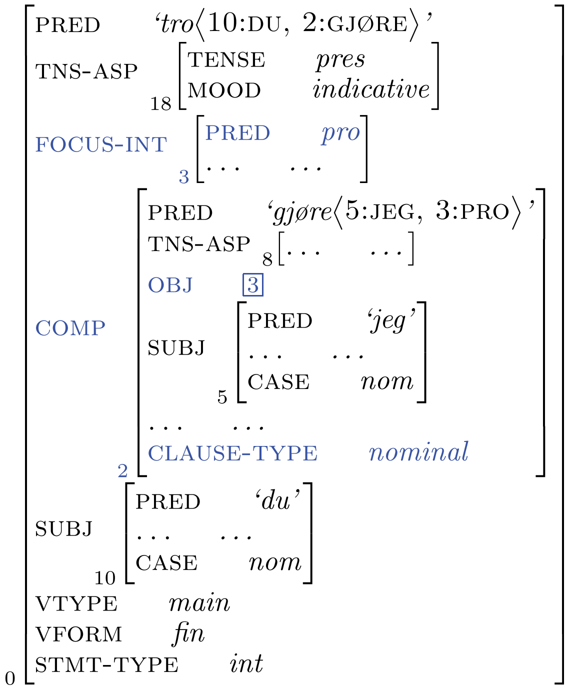

PRED ‘tro⟨10:DU, 2:GJØRE⟩’
TNS-ASP 18 [TENSE pres, MOOD indicative]
FOCUS-INT 3 [PRED pro, ...]
COMP
PRED ‘gjøre⟨5:JEG, 3:PRO⟩’
TNS-ASP 8 [... ...]
OBJ 3
SUBJ 5 [PRED ‘jeg’, ..., CASE nom]
... ...
CLAUSE-TYPE nominal
2
SUBJ 10 [PRED ‘du’, ..., CASE nom]
VTYPE main
VFORM fin
STMT-TYPE int
0

Figure 4. LFG f-structure for (14).

gjorde ‘did.’ To obtain the path, we start from 3, cross OBJ, then pop up via COMP (‘complement’) to the matrix clause, where we find the correct FOCUS-INT, as highlighted in blue in Figure 4.

(14) Hva tror du jeg gjorde?
What think you I did
‘What do you think I did?’

The paths mapped above correspond to the paths that would be represented in the right-hand side of traditional LFG equations (see Falk 2011; Bresnan et al. 2015; Dalrymple et al. 2019) for FGDs in (13) and (14) below:

(15) LFG equations for

a. example (13): (↑ FOCUS-INT) = (↑ SUBJ)

b. example (14): (↑ FOCUS-INT) = (↑ COMP OBJ)

We adopt the standard LFG representation of paths, with two modifications inspired by Pearl & Sprouse’s tagset. First, we append START and END labels to our paths to track dependency boundaries. Second, we augment our labels with information to track different types of finite clauses. Neither PS or LFG representations differentiate between types of finite complement clauses on the default label for such nodes (COMP in LFG, CP in PS), but being able to distinguish types of clauses is crucial for separating islands from non-islands. For this reason, Pearl & Sprouse chose to represent differences in clause type by annotating CP labels with lexical information about their head (e.g., CP$_{that}$). Instead of annotating COMP with the lexical complementizer, we used the CLAUSE-TYPE feature that each clause bears in an f-structure. For example, the embedded clause has the value *nominal* for its

CLAUSE-TYPE feature, so we modify COMP to COMP$_{nominal}$ (henceforth COMP$_{nom}$). Thus, the paths for (13) and (14) would be as follows:

(16) Paths for

a. example (13): START SUBJ END

b. example (14): START COMP$_{nom}$ OBJ END

3.2.3. *Path comparison*

In this section, we discuss differences between f-structure paths and paths calculated over phrase structure nodes and how these differences influence the modeled learner. Table 2 provides correspondences between PS sequence and f-structure sequences for different constructions (without START and END nodes at the sequence boundaries to preserve space).

F-structure paths encode more fine-grained information about clause structure and functional relations through a richer inventory of labels than paths stated over phrase structure nodes. For example, our tagset for *wh*-dependencies consists of 13 different f-structure labels compared to 9 phrase structure nodes reported in (Pearl & Sprouse 2013b); for relative clauses, our tagset consists of 18 different labels. This difference has a number of consequences. First, the greater number of f-structure labels leads to more unique n-grams. As a consequence, baseline n-gram probabilities may be lower on average than in a PS system because the probability mass is spread out over a larger set of n-grams.[9] Having more fine-grained n-grams may also exacerbate data sparsity issues. Second, f-structure paths are usually shorter than their phrase structure counterparts. Compare, for example, how f-structure paths for sentences with matrix subject, object, and oblique gaps are the same length but the length of the counterpart phrase structure paths varies with the depth of structural embedding. Within a clause, the functional structure is essentially flat, with no functional roles being more embedded than any others. Subsequently, the probabilities for FGDs spanning only the matrix clause directly reflect the raw corpus probabilities of the corresponding constructions. By contrast, probabilities calculated over phrase structure paths will exaggerate the actual corpus frequency differences between subject and non-subject FGDs because the path probabilities for non-subject FGDs will involve the multiplication of n-grams over clause-internal phrase structure nodes.

Nonfinite complement clauses are represented differently between PS and LFG formalisms. In PS representations, nonfinite complement clauses are treated as instances of IP (which may or may not fall under a CP), just like finite complement clauses. In LFG, nonfinite complements are analyzed as an instance of the generalized ‘open complement’ constituent XCOMP, distinct from finite complements. LFG treats a larger class of constructions as involving subordination of a nonfinite complement, which entails that XCOMP labels may show up in paths where an extra IP node would not be present in a PS

Table 2. Comparison between phrase structure paths and f-structure container node sequences.

Sentence	PS sequence	F-structure sequence
What happened?	IP	SUBJ
What is it?	IP	PREDLINK
What did they drink?	IP VP	OBJ
What are you talking about?	IP VP PP	OBL-TH
Where do you come from?	IP VP PP	ADJUNCT OBJ
Who do you think came?	IP VP CP$_{null}$ IP	COMP$_{nom}$ SUBJ
What do you think that she liked?	IP VP CP$_{that}$ IP VP	COMP$_{nom}$ OBJ
What do you think she wanted to buy?	IP VP CP$_{null}$ IP VP IP VP	COMP$_{nom}$ XCOMP OBJ

[9] However, the 7 most common f-structure labels account for 99.92% of all *wh*-dependencies and 99.41% of all RC-dependencies. These labels were SUBJ, OBJ, PREDLINK, OBL-TH, ADJUNCT, XCOMP, and COMP$_{nom}$.

path. For example, modals are analyzed as taking nonfinite clausal complements in LFG (as compared to VP complements in PS), so modals introduce XCOMP layers into an f-structure. We discuss one consequence of this analysis and a data sparsity issue it causes in Section 5.1.1.

As discussed above, finite complement clauses in LFG are labeled with COMP, compared to CP in PS representations. Pearl & Sprouse enriched the set of CP labels to include lexical complementizer information to distinguish between clause types. We annotated our COMP labels with their CLAUSE-TYPE feature, which ends up categorizing clause type more coarsely than using lexical complementizers would.

The implications of the different annotation scheme are most clearly seen with finite declarative complement clauses. A PS-based learner has at least two distinct CP nodes for such clauses: CP$_{null}$ when there was no complementizer and CP$_{that}$ when there was one. Our LFG-based learner only had one: COMP$_{nom}$, since declarative complement clauses bear the CLAUSE-TYPE: NOMINAL feature irrespective of whether they are headed by a lexical complementizer (*at* 'that'). This entails that the LFG-based system encodes the broader generalization that extraction is allowed out of declarative complement clauses using fewer n-grams than a PS-learner would. It also entails that the LFG-based learner has to encounter fewer distinct n-grams than the PS-learner, mitigating potential data sparsity issues for the acquisition of the broader generalization.

Embedded questions are also instances of COMP. For embedded questions, the choice between our f-structure labeling scheme and a PS-based system does not seem to make an appreciable difference. In our f-structures, there are two clause-type features for embedded questions: CLAUSE-TYPE: POLAR-INT for polar interrogatives and CLAUSE-TYPE: WH-INT for all others. A PS-based learner would similarly only need one label for polar questions in Norwegian, as there is only one complementizer, *om* 'whether,' that introduces polar interrogatives (compared to *if* and *whether* in English). Under the assumption that the same complementizer (C$_{+wh}$) introduces all constituent questions regardless of the identity of the *wh*-phrase in the specifier of CP, there would only be a single label for the questions in a PS-based system (CP$_{wh}$).[10]

The final relevant difference between the tagsets relates to how adjuncts are annotated in LFG. The basic set of labels in LFG contains ADJUNCT, which is applied to a diverse set of constituents including non-obligatory prepositional phrases as in (17), adverbial *wh*-words as *hvordan* 'how' in (18), and adverbial clauses as in (19). Relative clauses are also treated as adjuncts embedded under the role assigned to their head noun, as in (20), which is illustrated in Figure 5.

(17) Hvor kommer du fra?
Where come you from
'Where do you come from?'

(18) Hvordan går det?
How goes it?
'How is it going?'

(19) Jeg sov da du kom.
I slept when you came
'I slept when you came.'

(20) Du så lampa jeg kjøpte.
You saw lamp.DEF I bought
'You saw the lamp I bought.'

[10] We note that Pearl & Sprouse only tested polar interrogatives with CP$_{whether}$ and did not discuss how other embedded questions would be represented. We have assumed CP$_{wh}$ for the sake of simplicity, but alternative container node schema are possible. For example, CP nodes could be annotated with the *wh*-phrase in their specifier, leading to different container nodes for different embedded question types (e.g., CP$_{who}$, CP$_{what}$, etc.). Under such a scheme, learning the broad generalization that extraction from embedded questions is allowed would require learning separate generalizations about different question types.

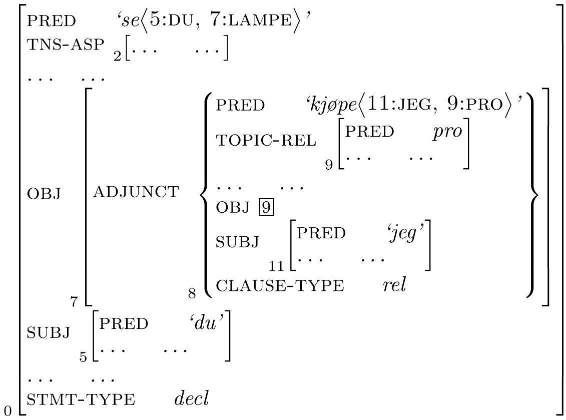


Figure 5. LFG f-structure for (20).

Clausal adjuncts are distinguished from nonclausal adjuncts, in that they are marked with a CLAUSE-TYPE feature: CLAUSE-TYPE: REL for relative clauses and CLAUSE-TYPE: ADV for adverbial adjunct clauses. As we did with COMP, we added clause-type information to the ADJUNCT label for any clausal adjunct, resulting in the tags ADJUNCT$_{rel}$ and ADJUNCT$_{adv}$. Consequently, extraction from these domains should include ADJUNCT$_{rel}$ or ADJUNCT$_{adv}$ in the container node sequence.

A side-by-side comparison of the PS and f-structure container node sequences for all our test sentences (with START and END nodes omitted) is presented in Table 3. As can be seen, both formalisms can distinguish between different sentence types and have unique sequences for all potential island structures.

To demonstrate how the paths can distinguish FGDs into potential islands from non-islands, consider subject extraction from an embedded question (EQ), for which an experimental *wh*-dependency item is in (21). Both *Short* conditions feature matrix subject extraction. The sentences

Table 3. Test conditions with container node sequences.

Constituent	Condition	PS sequence	F-structure sequence
Subject	Short, no island	IP	SUBJ
Subject	Long, no island	IP-VP-CP$_{that}$-IP	COMP$_{nom}$ SUBJ
Subject	Short, island	IP	SUBJ
Subject	Long, island	IP-VP-CP$_{that}$-IP-NP-PP	COMP$_{nom}$ SUBJ ADJUNCT OBJ
RC	Short, no island	IP	SUBJ
RC	Long, no island	IP-VP-CP$_{that}$-IP-VP	COMP$_{nom}$ OBJ
RC	Short, island	IP	SUBJ
RC	Long, island	IP-VP-CP$_{rel}$-IP-VP	ADJUNCT$_{rel}$ OBJ
EQ	Short, no island	IP	SUBJ
EQ	Long, no island	IP-VP-CP$_{that}$-IP	COMP$_{nom}$ SUBJ
EQ	Short, island	IP	SUBJ
EQ	Long, island	IP-VP-CP$_{wh}$-IP	COMP$_{wh-int}$ SUBJ
Adjunct	Short, no island	IP	SUBJ
Adjunct	Long, no island	IP-VP-CP$_{that}$-IP-VP	COMP$_{nom}$ OBJ
Adjunct	Short, island	IP	SUBJ
Adjunct	Long, island	IP-VP-CP$_{if}$-IP-VP	ADJUNCT$_{adv}$ OBJ

correspond to the dependency IP or SUBJ. Given that the filler-gap paths are short and direct evidence for them is likely abundant, we would expect a modeled leaner to assign high probabilities to these conditions. The *Long, no island* condition is an example of subject extraction from an embedded declarative clause, so the filler-gap path is longer: IP-VP-CP$_{that}$-IP or COMP$_{nom}$ SUBJ. Because the dependency is longer and there are likely fewer examples of the dependency in the corpus, we would expect the dependency paths in this condition to have lower probability compared to the *Short* ones. Finally, the *Long, island* condition includes subject extraction from an embedded question, which is reflected in the complementizer specification in both path representations (IP-VP-CP$_{wh}$-IP VS. COMP$_{wh-int}$ SUBJ). The probability of this condition relative to *Long, no island* condition will depend on whether the learner has seen dependencies into embedded questions during the acquisition period. If the training corpus does not contain direct evidence for such structures, we expect much lower probability, as n-grams containing COMP$_{wh-int}$/CP$_{wh}$ would be unattested. If the corpus does contain examples of the structure, then the probability of the *Long, island* dependency compared to the *Long, no island* dependency will depend on the relative frequency of sentences containing subject extraction from embedded declaratives and embedded questions in the corpus.

(21) a. **Short, no island**

Hvilken snekker sa at hylla skulle monteres i stuen?
which carpenter said that shelf.DEF should install.PASS in living.room.DEF
'Which carpenter __ said that the shelf should be installed in the living room?'

b. **Long, no island**

Hvilken hylle sa snekkeren at skulle monteres i stuen?
which shelf said carpenter.DEF that should install.PASS in living.room.DEF
'Which shelf did the carpenter say (that) __ should be installed in the living room?'

c. **Short, island**

Hvilken snekker sa hvor hylla skulle monteres?
which carpenter said where shelf.DEF should install.PASS
'Which carpenter __ said where the shelf should be installed?'

d. **Long, island**

Hvilken hylle sa snekkeren hvor skulle monteres?
which shelf said carpenter.DEF where should install.PASS
'Which shelf did the carpenter say where __ should be installed?'

3.3. *Model training*

To derive probability estimates for container sequences corresponding to our test dependencies, we trained our modeled learner on attested f-structure paths according to the procedure introduced by Pearl & Sprouse, with minor modifications. First, the modeled learner was trained on *wh*- and RC-dependencies separately. Second, in addition to tracking container node trigrams, as Pearl & Sprouse did, we ran a separate version of the algorithm that tracked the frequencies of container node bigrams. We reasoned that f-structure bigrams might be sufficient for capturing the relevant co-occurrence restrictions given that f-structure paths tend to be shorter and encode structural features more compactly than their PS counterparts (compare paths in Table 3).

Pearl & Sprouse reported a single probability value for each dependency path, calculated directly from the empirical corpus probabilities. Such single point estimates do not provide information about the uncertainty around the predicted probabilities or how data sparsity and variation in the input could affect the modeled learner's predictions. To quantify uncertainty, we used bootstrapping techniques to construct confidence intervals over the probability estimates for test dependency paths (DiCiccio & Efron 1996). To do so, we generated 1,000 training sets for *wh*- and RC-dependencies apiece by sampling dependencies from the corpus with replacement. Each training set was equal in size to the original set (19,004 *wh*-dependencies and 40,702 RC-dependencies). We then ran the learning algorithm on each training set to obtain bigram and trigram probabilities, applying a smoothing procedure that assigned a count of 0.5 (instead of 0) to any unattested n-grams. Finally, we calculated the probabilities for test dependency paths in each iteration. After training we took the middle 95% of the distribution over predicted probabilities as our confidence interval for each n-gram and dependency combination.

4. Results

In this section, we present our results and evaluate the performance of the learning algorithm.

4.1. *Overview of direct evidence: Corpus counts*

Before investigating how the modeled learner's probabilities align with human judgments, we present an overview of how often FGDs that exemplify our different test conditions are found in the corpus.

Raw counts and the log probabilities that the modeled learner associates to the corresponding container node sequences can be found in Table 4 split by dependency type and n-gram window. We use both bigrams and trigrams to derive the probability. The counts represent the frequency of relevant direct evidence available to the learner.

Looking first at the baseline grammatical dependencies, we see that the input contains ample direct evidence for most structures of interest. Short-distance dependencies are overwhelmingly more frequent than long-distance dependencies. Though infrequent, long-distance dependencies into embedded declarative clauses are observed for all but one cell of the paradigm: Among the nearly 20,000 *wh*-dependencies analyzed, there were no instances of long-distance subject questions like *Who do you think __ came?* that mapped to the container node sequence START-COMP$_{nom}$-SUBJ-END. The input data included one instance embedded subject extraction, shown in (22).

(22) Hva ville du aller helst [__ skulle skjedd]?
What wanted you of.all preferably should happened
lit. 'What did you want should have happened?'
≈ 'What would you have preferred to have happen?'

Even though (22) is an example of subject extraction from a tensed complement clause, it is not included in our counts because SUBJ is not adjacent to COMP$_{nom}$ in the FGD's container node sequence, which is START-COMP$_{nom}$-XCOMP-SUBJ-END. According to convention, the finite auxiliary *skulle* 'should' introduces an XCOMP layer between the nominal complement and the subject gap, essentially treating the embedded clause as having two layers of clausal embedding. To our knowledge, this is the only instance where the LFG formalism creates a data sparsity issue that has any appreciable consequences for the model learner (or where there would be any divergence from a PS-based learner). As we discuss below, the absence of these dependencies has consequences for the

Table 4. Corpus counts and model probabilities for baseline grammatical filler-gap dependencies and dependencies that correspond to *Long, island* test conditions. START and END nodes in the 'Filler-gap path' column are omitted to preserve space. The columns labeled 'Island?' report whether humans exhibited an island effect for the structure in Kobzeva, Sant et al. (2022).

	Island?			Raw count		Log probability			
Structure	RC	Wh	Filler-gap path	RC	Wh	RC-bigr	RC-trigr	Wh-bigr	Wh-trigr
			Short conditions (grammatical structures)						
Matrix subj.	No	No	SUBJ	16,418	1,493	−3.29	−1.23	−6.36	−2.79
Matrix obj.	No	No	OBJ	7,934	2,623	−4.33	−1.95	−4.97	−2.23
Copula pred.	No	No	PREDLINK	1,010	3,930	−8.58	−4.01	−4.69	−1.82
			Long, no island condition (grammatical structures)						
Emb. subj.	No	No	COMP_{nom} SUBJ	11	0	−16.15	−16.68	−20.23	−21.58
Emb. obj.	No	No	COMP_{nom} OBJ	81	32	−14.66	−12.93	−15.20	−13.21
			Long, island condition (potential island structures)						
Subject	Yes	Yes	COMP_{nom} SUBJ ADJ. OBJ	0	0	−32.30	−34.86	−35.26	−36.28
RC	Yes	No	ADJUNCT_{rel} OBJ	0	0	−26.18	−23.25	−24.91	−21.58
EQ	No	Yes	COMP_{wh-int} SUBJ	20	0	−18.38	−15.88	−25.73	−21.58
Adjunct	No	Yes	ADJUNCT_{adv} OBJ	0	0	−26.18	−23.25	−24.91	−21.58

modeled learner's ability to adequately approximate the target state and arguably poses a potential Poverty of the Stimulus challenge.

Turning to potential island structures, we see that all but one of the test dependencies are absent. There is no direct evidence of *wh*-dependencies into subjects, relative clauses, embedded questions, or adjuncts. Similarly, the training corpus lacks examples of RC-dependencies into subjects, relative clauses, or adjuncts. Because they are unattested, the modeled learner should treat these dependencies as island violations. The corpus does contain, however, examples of RC-dependencies into embedded questions. Interestingly, RC-dependencies into embedded questions are observed slightly more often than their counterparts into declarative complements (START-COMP$_{wh-int}$-SUBJ-END VS. START-COMP$_{nom}$-SUBJ-END). Thus, the modeled learner receives direct evidence that embedded questions are not islands in Norwegian—but for RC-dependencies only.

4.2. *Success metrics*

We evaluate the success of the modeled learner by comparing probabilities that it assigns to dependency paths discussed in Section 3.2 to human judgments from (Kobzeva, Sant et al. 2022). When assessing whether the modeled learner exhibits an island effect, we ask whether its predicted probabilities resemble the superadditive pattern. That is, whether (i) the *Long, island* condition is less probable than all other conditions and (ii) the difference between the *Long, island* and *Long, no island* conditions is greater than the difference between the two *Short* conditions. Following Pearl & Sprouse, we consider two different outcomes as successes: (i) any instance where Kobzeva, Sant et al. (2022) observed a significant superadditive interaction and the modeled learner predicts a superadditive pattern and (ii) any instance where Kobzeva, Sant et al. (2022) failed to find an interaction effect and the model does not predict one. In either case, we say that the modeled learner's predictions "align" with the human judgments. Importantly, all comparisons are qualitative rather than quantitative. We refrain from directly comparing effect sizes between the modeled learner and human data, nor do we attempt to establish a mapping between model probabilities and "absolute" human ratings. This is because our learner's probabilities only track syntactic probability of the FGD-path, abstracting away from the interplay of additional lexical, semantic, syntactic, discourse-pragmatic, and processing factors that impact and potentially skew human acceptability judgments.

To facilitate comparison, we plot modeled probabilities (derived from both bigrams and trigrams) and human judgments side by side for each dependency-island combination. In all figures, acceptability ratings are plotted on the left, with the y-axis representing mean condition ratings z-scored by participants and error bars representing 95% confidence intervals. The modeling results are plotted in the middle for bigrams and on the right for trigrams with the y-axis reflecting log probability and error bars representing bootstrapped 95% confidence intervals. Following Pearl & Sprouse, we report the modeling results on a log probability scale because n-gram probabilities (and their products) are very small numbers. All log probabilities are negative numbers. The closer a log probability is to 0, the more probable it is. Adopting the (simplifying) assumption that probability maps directly to acceptability, we interpret probabilities further away from 0 as "less acceptable" sequences.

4.3. *Subject island*

Example subject island sentences from Kobzeva, Sant et al. (2022) for RC- and *wh*-dependencies are presented in (23). In (23-c), the label ADJUNCT links the gap-containing prepositional phrase introduced by *med* 'with' to the subject *avtalen* 'the agreement.'

(23) a. Det var bankene$_i$ som rådgiveren trodde avtalen med __$_i$ vil
That was banks.DEF RP advisor.DEF thought agreement.DEF with will
forsterke det politiske samspillet.
strengthen the political interaction.DEF
'Those were the banks$_i$ that the advisor thought the agreement with __$_i$ will strengthen the political interaction'.

b. Hvilke banker$_i$ trodde rådgiveren at avtalen med __$_i$ vil forsterke
Which banks thought advisor.DEF that agreement.DEF with will strengthen
det politiske samspillet?
the political interaction.DEF
'Which banks$_i$ did the advisor think that the agreement with __$_i$ will strengthen the political interaction?'

c. F-structure path: START COMP$_{nom}$ SUBJ ADJUNCT OBJ END
Bigrams: (START COMP$_{nom}$), (COMP$_{nom}$ SUBJ), (SUBJ ADJUNCT), (ADJUNCT OBJ), (OBJ END)
Trigrams: (START COMP$_{nom}$ SUBJ), (COMP$_{nom}$ SUBJ ADJUNCT), (ADJUNCT OBJ END)

Figure 6 plots average human acceptability judgments on the left and modeled probabilities for bigrams and trigrams on the right. Human participants in (Kobzeva, Sant et al. 2022) showed clear island effects for both RC- and *wh*-dependencies into NP subjects. The qualitative patterns of the modeled learner's bigram and trigram log probabilities exhibit similar island effects, as the modeled learner assigns a much lower log probability to container node sequences corresponding to subject island violations.

Because FGDs into subjects are unattested in the corpus, the container node sequence in (23-c) has parts that are unattested, summarized in Table 5. In the case of bigram-based probabilities with RC-dependencies, only one bigram is unattested in the *Long, island* condition, reflecting the fact that extraction from subjects is absent in the corpus. With *wh*-dependencies, this number doubles due to the absence of embedded subject extraction in the grammatical *Long, no island* condition.

Despite the apparent success of the modeled learner, the interaction in the *wh*-dependency subset obscures a disconnect between the modeled and human judgments. Human participants rate embedded subject extraction—the *Long, no island* condition—as only moderately less acceptable than either *Short* condition and within the range typically assigned to other grammatical sentences in the experiment. The modeled learner, on the other hand, treats embedded subject extraction on par with ungrammatical island violations: It assigns the same absolute log probability to the grammatical *Long, no island* condition in the subject island experiment as it does to other RC island and adjunct island violations because the corpus contains no evidence of *wh*-dependencies with embedded subjects. Thus, the numerical difference between the *Long, no*

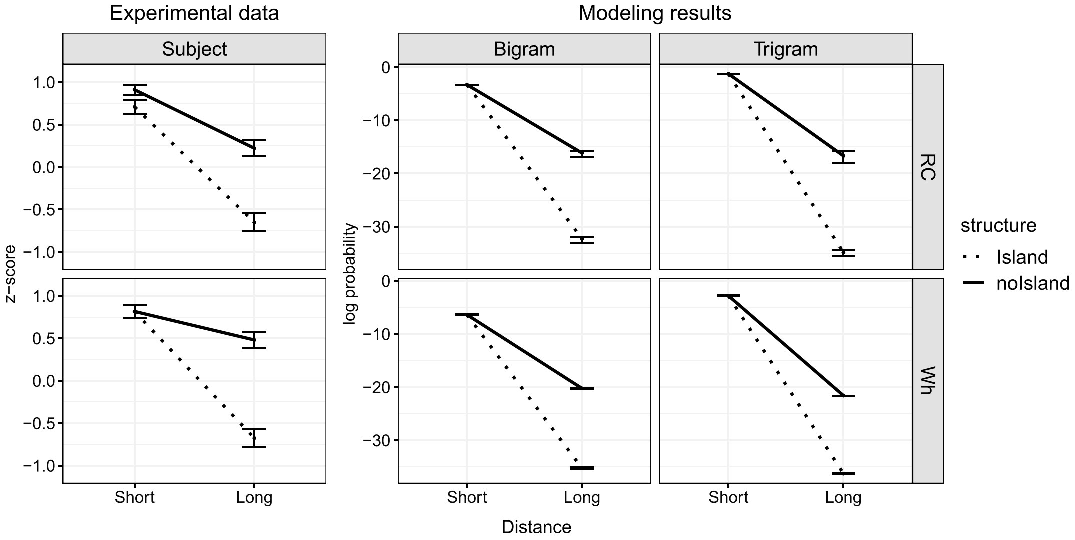


Figure 6. Interaction plots for experimental data and modeling results for Subject island. Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

Table 5. Unattested n-grams, Subject island.

Dependency	Condition	Unattested n-grams
		Bigram
RC	*Long, island*	SUBJ ADJUNCT
Wh	*Long, no island*	COMP$_{nom}$ SUBJ
	Long, island	COMP$_{nom}$ SUBJ, SUBJ ADJUNCT
		Trigram
RC	*Long, island*	COMP$_{nom}$ SUBJ ADJUNCT, SUBJ ADJUNCT OBJ
Wh	*Long, no island*	START COMP$_{nom}$ SUBJ, COMP$_{nom}$ SUBJ END
	Long, island	START COMP$_{nom}$ SUBJ, COMP$_{nom}$ SUBJ ADJUNCT, SUBJ ADJUNCT OBJ

island condition and the ungrammatical *Long, island* conditions arises because the path in the former condition is both longer and contains more unattested n-grams than the path in the latter. For bigrams, calculating the probability of *Long, no island* involves one unattested bigram, whereas the calculation for the Subject island violation involves multiplying two unattested bigrams (specified in Table 5). Similarly for trigrams, calculating the probabilities of *Long, no island* and *Long, island* conditions involves the multiplication of two and three unattested trigrams respectively.

Overall, the modeled learner clearly approximates the target state for RC-dependencies. Focusing solely on whether there is a superadditive pattern, it also appears that the learner also approximates the target state for *wh*-dependencies, as the *Long, island* condition is judged to be much less probable than the *Long, no island* condition.

4.4. *Relative clauses*

Example test sentences from Kobzeva, Sant et al. (2022) for both dependency types are in (24). Human judgments and modeled probabilities are plotted in Figure 7.

Human judgments and modeled probabilities align for RC-dependencies: The modeled learner predicts island effects found in acceptability ratings. The human data and modeled probabilities do

(24) a. Det var boka$_i$ som det var mange som likte $\_\_{}_i$.
That was book.DEF RP that was many RP liked
'It was the book$_i$ that there were many (people) who liked $\_\_{}_i$.'

b. Hvilken bok$_i$ var det mange som likte $\_\_{}_i$?
Which book were that many RP liked
'Which book$_i$ were there many people who liked $\_\_{}_i$?'

c. F-structure path: START ADJUNCT$_{rel}$ OBJ END

Bigrams: (START ADJUNCT$_{rel}$), (ADJUNCT$_{rel}$ OBJ), (OBJ END)

Trigrams: (START ADJUNCT$_{rel}$ OBJ), (ADJUNCT$_{rel}$ OBJ END)

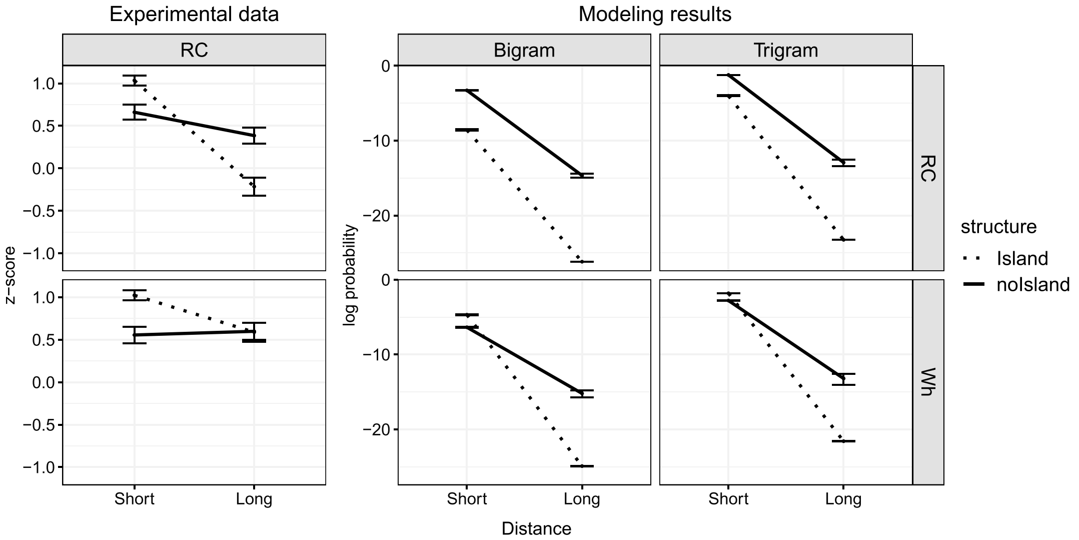


Figure 7. Interaction plots for experimental data and modeling results for RC island. Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

not align with *wh*-dependencies. Human judgments show no island effects for *wh*-dependencies into RCs, but the modeled learner exhibits an island effect with both bigrams and trigrams due to no instances of extraction from RCs in the corpus (see Table 6).[11]

[11] Differences between the *Short, island* and *Short, no island* conditions can be attributed to the fact that the test items in the behavioral experiment deviated from a strict factorial design for the *Short, island* condition; see Kobzeva, Sant et al. (2022) for more details.

Table 6. Unattested n-grams, relative clauses.

Dependency	Condition	Unattested n-grams
	Bigram	
RC	*Long, island*	START ADJUNCT$_{rel}$, ADJUNCT$_{rel}$ OBJ
Wh	*Long, island*	START ADJUNCT$_{rel}$, ADJUNCT$_{rel}$ OBJ
	Trigram	
RC	*Long, island*	START ADJUNCT$_{rel}$ OBJ, ADJUNCT$_{rel}$ OBJ END
Wh	*Long, island*	START ADJUNCT$_{rel}$ OBJ, ADJUNCT$_{rel}$ OBJ END

4.5. *Embedded questions*

Example sentences from Kobzeva, Sant et al. (2022) for testing RC and *wh*-dependencies into EQs and the corresponding container node sequences are in (25).

(25) a. Det var signalet$_i$ som sjømennene visste hva __$_i$ betydde.
That was signal.DEF RP fishermen.DEF knew what meant
'That was the signal$_i$ that the fishermen knew what __$_i$ meant'.

b. Hvilket signal$_i$ visste sjømennene hva __$_i$ betydde ?
Which signal knew sailors.DEF what meant
'Which signal$_i$ did the sailors know what __$_i$ meant?'

c. F-structure path: START COMP$_{wh-int}$ SUBJ END

Bigrams: (START COMP$_{wh-int}$), (COMP$_{wh-int}$ SUBJ), (SUBJ END)

Trigrams: (START COMP$_{wh-int}$ SUBJ), (COMP$_{wh-int}$ SUBJ END)

As can be seen in Figure 8, no island effects are observed in the human judgment patterns or the modeled probabilities for RC-dependencies. For *wh*-dependencies, the human judgments show an island effect, and so does the modeling result based on bigrams, but the result calculated using trigrams does not.

The absence of an island effect for RC-dependencies is expected because relativization of a subject from an EQ is just as (if not slightly more) frequent than relativization of a subject from an embedded declarative clause.

The probabilities are low for both dependencies, but they fall within the "attested" range of probabilities assigned to "attested" structures.[12] For *wh*-dependencies, there is a difference between bigram- and trigram-based modeled learners.

With bigrams, there is only one unattested bigram in *Long, no island* condition and two in *Long, island*, which explains the island effect. With trigrams, the modeled learner treats both sentences as equally *improbable* because *Long, no island* and *Long, island* conditions are both unattested in the corpus, as summarized in Table 7. If the input had contained any evidence that *wh*-extraction of an

[12] At an anonymous reviewer's request, we present density distributions of modeled probabilities for attested and unattested structures in Appendix C, alongside density plots of z-scored human ratings.

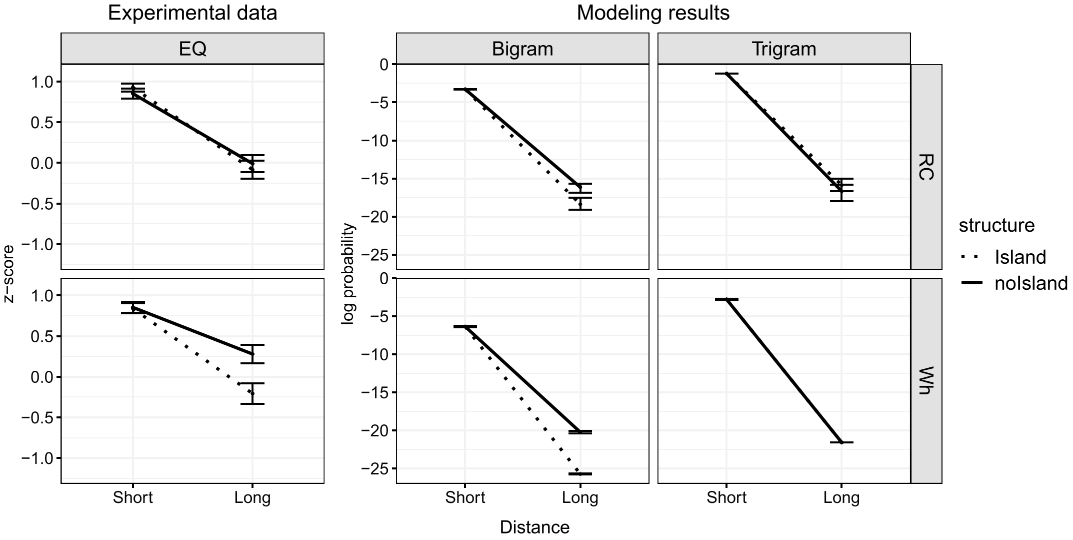


Figure 8. Interaction plots for experimental data and modeling results for EQ island (subject position). Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

Table 7. Unattested n-grams, EQs.

Dependency	Condition	Unattested n-grams
	Bigram	
Wh	*Long, no island*	COMP$_{nom}$ SUBJ
	Long, island	START COMP$_{wh-int}$, COMP$_{wh-int}$ SUBJ
	Trigram	
Wh	*Long, no island*	START COMP$_{nom}$ SUBJ, COMP$_{nom}$ SUBJ END
	Long, island	START COMP$_{wh-int}$ SUBJ, COMP$_{wh-int}$ SUBJ END

embedded subject was possible, then the modeled learner would have predicted an interaction effect with trigrams as well (and a larger interaction effect with bigrams).[13]

4.6. *Adjunct island*

Examples of FGDs into conditional adjunct clauses from (Kobzeva, Sant et al. 2022) are presented in (26) with the corresponding container node sequences in (26-c).
In Figure 9, we see a misalignment between human and modeled learner judgments for RC-dependencies. Human ratings show no adjunct island effect, but an interaction is apparent in the modeled probabilities. Relative alignment is observed between human judgments of *wh*-dependencies and the modeled learner's predicted probabilities, all of which exhibit adjunct island effects that are smaller in magnitude than island effects observed with other constituents.

The modeled learner predicts island effects for both dependency types because there are no examples of RC- or *wh*-dependencies into conditional/clausal adjuncts in the corpus; the unattested n-grams are presented in Table 8. Once again, we can evaluate the learner's performance as mixed, given that it captures the observed pattern of average judgments for *wh*-dependencies but not RC-dependencies.

[13] This is indeed the result when testing the modeled learner's predictions for *wh*-dependencies with embedded object gaps, which are present in the input (see Table 4 for counts and Appendix B for demonstration).

(26) a. Det er maleriet$_i$ som kvinnen blir glad om han kjøper __$_i$.
That is painting.DEF RP woman.DEF becomes happy if he buys
'That's the painting$_i$ that the woman will be happy if he buys __$_i$.'

b. Hvilket maleri$_i$ blir kvinnen glad om han kjøper __$_i$?
Which painting becomes woman.DEF happy if he buys
'Which painting$_i$ will the woman be happy if he buys __$_i$?'

c. F-structure path: START ADJUNCT$_{adv}$ OBJ END

Bigrams: (START ADJUNCT$_{adv}$), (ADJUNCT$_{adv}$ OBJ), (OBJ END)

Trigrams: (START ADJUNCT$_{adv}$ OBJ), (ADJUNCT$_{adv}$ OBJ END)

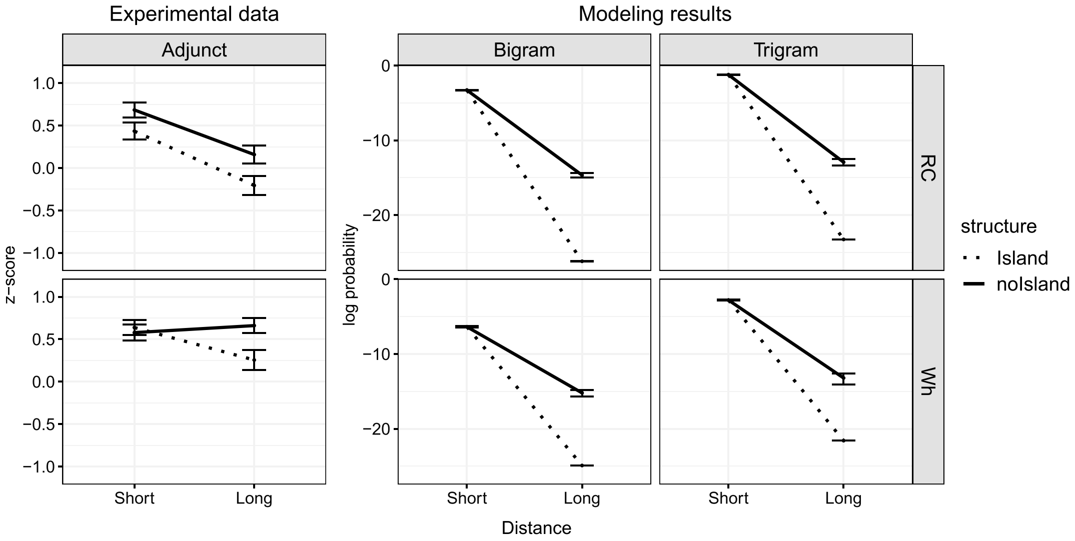


Figure 9. Interaction plots for experimental data and modeling results for Adjunct island. Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

5. Discussion

Prior work (Pearl & Sprouse 2013a,b; Pearl & Bates 2022) suggests that distributional learning models that track container node n-grams can learn about islands for *wh*-dependencies in English. We investigated whether a similar model could learn about island effects in Norwegian to assess whether the method could be a viable acquisition model for islands cross-linguistically. Norwegian is a relevant comparison case because the language allows FGDs into some constituents that are islands in English.

We trained our modeled learner on *wh*- and RC-dependencies taken from an LFG-parsed corpus of child-directed text. Dependencies were represented as sequences of container nodes along the path between the filler and the gap in the hierarchical f-structure representation of a sentence. The learner tracked the probability of container node n-grams along the dependency path so that it could later estimate the probability of observed or novel dependencies as the product of their constituent n-grams. We used two

Table 8. Unattested n-grams, Adjunct island.

Dependency	Condition	Unattested n-grams
	Bigram	
RC	*Long, island*	START ADJUNCT$_{adv}$, ADJUNCT$_{adv}$ OBJ
Wh	*Long, island*	START ADJUNCT$_{adv}$, ADJUNCT$_{adv}$ OBJ
	Trigram	
RC	*Long, island*	START ADJUNCT$_{adv}$ OBJ, ADJUNCT$_{adv}$ OBJ END
Wh	*Long, island*	START ADJUNCT$_{adv}$ OBJ, ADJUNCT$_{adv}$ OBJ END

different n-gram windows for probability estimation: bigrams and trigrams. These probabilities were then compared against native speakers' acceptability judgments of similar dependencies.

Our results were mixed. The model did not succeed across all islands and dependency types, though it performed worse with *wh*-dependencies than with RC-dependencies. By and large, the bigram-based modeled learner performed as well as or even better than the trigram-based modeled learner. Our results are summarized in Table 9.

5.1. *Direct evidence, data sparsity, and the poverty of the stimulus*

We asked whether the modeled learner was able to recover a number of different generalizations about FGD formation in Norwegian from the input it received. Overall, the modeled learner succeeds (perhaps not surprisingly) when the input contains direct evidence of the dependencies in question. One notable success is that the modeled learner recovers the generalization that EQs are not islands for RC-dependencies in Norwegian, because the input corpus contains 31 relevant examples like (27). Because sentences like (27) are essentially as frequent as RC-dependencies into embedded declarative clauses, the modeled learner treats relativization from an embedded question just as acceptable as from an embedded declarative clause.

(27) Vi begge var redde for noe$_i$ vi ikke visste [hva __$_i$ var].
We both were afraid of something we NEG knew what was.
'We both were afraid of something we did not know what __ was.'

Table 9. Summary of the modeling results.

	RC-dependencies		*Wh*-dependencies	
(Potential) island type	Bigram	Trigram	Bigram	Trigram
Subject phrases	✓	✓	✓	✓
Relative clauses	✓	✓	✗	✗
Embedded questions	✓	✓	✓	✗
Adjunct clauses	✗	✗	✓	✓
Total successes:	3	3	3	2

✓ represents cases where the learning algorithm succeeded at representing the target state, ✗ where it did not.

We saw that the modeled learner was not uniformly successful at recovering human-like patterns of (in) sensitivity to islands: In one case, it treats long-distance subject questions, a grammatical "control" FGD, as equally improbable as an island violation. In two other cases, the modeled learner predicts a superadditive island effect where humans do not exhibit one: It assigns very low probability to *wh*-dependencies into relative clauses and to RC-dependencies into adjunct clauses. In all three cases, the modeled learner fails because it has not seen direct evidence of the three dependencies in question. As we discuss below, in one case, the absence of relevant direct evidence can be attributed to how the model represents chunks and distinguishes between input dependencies. For this case, we discuss how modifications to the model might overcome some of the problems. In other cases, the lack of evidence may, in fact, constitute a true POS problem.

5.1.1. *Long-distance subject questions*

The modeled learner assigns a very low probability to acceptable baseline sentences featuring *wh*-extraction of a subject from an embedded declarative clause, as in (28), which corresponds to the path START-COMP$_{nom}$-SUBJ-END. The modeled learner thus could not distinguish the acceptable path from unacceptable island-violating paths in some comparisons, so it could not learn the correct generalization.

(28) Hvem sa du [__ kom]?
Who said you __ left?
'Who did you say left?'

F-structure path: START-COMP$_{nom}$-SUBJ-END

The failure of the modeled learner in this instance highlights how choices of input representation or model parameters can exacerbate data sparsity problems. The low probability assigned to (28) reflects the fact that n-grams containing the sequence COMP$_{nom}$-SUBJ are unattested in the corpus.[14] Within the LFG-based system that we used, the only sentences that could provide evidence for those n-grams are sentences like (28) themselves or the Norwegian equivalent of more deeply embedded subject *wh*-extraction like (29), where the labels COMP$_{nom}$ and SUBJ are directly adjacent. However, such sentences are also absent from the corpus.

(29) Who did you say that Ole believed that Marit thought __ left?

F-structure path: START COMP$_{nom}$ COMP$_{nom}$ SUBJ END

Even though examples of *wh*-extraction of embedded objects (COMP$_{nom}$ OBJ) and oblique arguments (COMP$_{nom}$ OBL) are in the input, such sentences do not count as indirect evidence toward learning the generalization that embedded subjects can be *wh*-extracted because the attested paths are different from the crucial COMP$_{nom}$ SUBJ sequence. A learner trained on PS-trigrams from the same input corpus would encounter the same generalization problem, since the path for embedded subject extraction contains a unique trigram, CP-IP-END, that could not be learned indirectly from other sentence types. It is worth mentioning that in a hypothetical case if non-subject (object or oblique) EQs were missing from the corpus, the PS-based learner would have an advantage over the LFG-based one, because non-subject EQs share a trigram CP-IP-VP in their container node sequence, and one type of non-subject EQ would provide evidence for the other. In the case of the LFG-based learner, the

[14] For the bigram-based learner only (COMP$_{nom}$-SUBJ) is unattested; for the trigram-based learner (START-COMP$_{nom}$-SUBJ) and (COMP$_{nom}$-SUBJ-END) are unattested.

learner has to learn separate generalizations for embedded object and embedded oblique questions as they include unique bigrams (COMP$_{nom}$ OBJ AND COMP$_{nom}$ OBL).

The choice of f-structure labels resulted in one data sparsity issue that specifically impacted the modeled learner's ability to learn dependencies like (28). Recall that we saw that the input contained an example sentence that would commonly be described as *wh*-extraction of an embedded subject: (22), reprinted below:

(22) Hva ville du aller helst [_ skulle skjedd]?
What wanted you of.all preferably should happened

lit. 'What did you want that should (have) happened?'

≈ 'What would you have preferred to have happen?'

F-structure path: START COMP$_{nom}$ XCOMP SUBJ END

As discussed in Section 4, even though (22) contains extraction of an embedded subject, our modeled learner cannot not count it as evidence in favor of (28) because of the embedded modal *skulle* ('should'). Since modals are assumed to add an extra layer of nonfinite embedding to the f-structure, *skulle* inserts an XCOMP label between COMP$_{nom}$ and SUBJ in the dependency path. Consequently, the modeled learner does not see evidence of the relevant sequence. We see that for an LFG-based learner, acquiring the broader generalization that embedded subjects can be *wh*-extracted requires learning two effectively independent generalizations encoded as n-gram sequences: that embedded subjects can be extracted when there is no modal verb and when there is one. A PS-based learner would not fall victim to the same problem, because under a PS representation, either of the two questions would constitute sufficient evidence to learn the broader generalization.

The discussion above highlights one of the challenges our modeled learner faced in learning a crucial generalization when the input corpus does not contain (direct) positive evidence of a specific sentence type. It also reveals ways in which the learner fails to generalize from 'neighbor' examples in the absence of direct evidence. Though the issue clearly highlights a conceptual concern, we do not think that learning the correct generalization represents an insurmountable practical problem, because we know that sentences of the relevant kinds occur at least occasionally in regular speech and are found in (adult-oriented) texts (Kush et al. 2021). The absence of examples like (28) in our training set is most likely an "accidental" gap, at least partly reflecting our choice of a relatively small corpus drawn from written texts, which are likely to underrepresent the distribution of questions and embedded questions compared to child-directed speech (Noble et al. 2018). In other words, if the number of sentences with *wh*-dependencies in our corpus was comparable to the number of sentences with RC-dependencies, the learner would be more likely to encounter the relevant examples.

5.1.2. *RC island*

Norwegian speakers judge *wh*-dependencies into (existential) relative clauses as acceptable, but our modeled learner treats them as islands. The learner's behavior reflects that there is no direct evidence for the relevant dependency path in our training corpus, and necessary n-grams for the path in question cannot be acquired indirectly from other sentence types because RCs bear a unique label (ADJUNCT$_{rel}$) not found in other constructions.

We consider it highly unlikely that the absence of these dependencies from our corpus is just an accidental gap. We base our speculation on recent corpus findings from Müller & Eggers (2022). Danish reportedly allows the same set of FGDs into RCs as Norwegian; however, Müller & Eggers found no examples of *wh*-dependencies into relative clauses in a large collection of Danish corpora (*KorpusDK*: 56 million words, *BySoc* corpus: 1.3 million words, and *SamtaleBank*: 6h, 20m of naturalistic conversation (MacWhinney & Wagner 2010)) or in targeted searches on Google. The complete absence of relevant examples from such a large sample of texts suggests that they do not exist.

We see no reason to suspect that the Norwegian distribution differs from the Danish in this regard. It seems that Norwegians must learn the generalization through some other means. The modeled learner in its current form is not capable of doing so.

One modification to the distributional learner could possibly overcome this problem. The distributional learner currently tracks separate n-gram probabilities for different dependency types, which is, on the one hand, motivated by the need to capture the cross-dependency variation observed in human judgments (Kobzeva, Sant et al. 2022). On the other hand, tracking distinct n-grams goes against common practice in generative linguistics to treat both dependency types as created using the same basic operation (i.e., A-bar movement; Chomsky 1973, 1986) and subject to the same constraints.

There is reason to believe that collapsing across dependency types to track n-grams for A-bar movement generally could mitigate the problem of unattested *wh*-dependencies into relative clauses. Even though *wh*-dependencies are unattested, there is evidence that other A-bar dependencies can cross into relative clauses. For example, although we did not include them in our study, *topicalization* dependencies into relative clauses are quite common in Mainland Scandinavian languages (Engdahl 1997; Christensen 1982; Lindahl 2017; Müller & Eggers 2022), and work has verified that dependencies like (30) are in the input to children in Norwegian (Kush et al. 2021).

(30) Det$_i$ er det ingen [som vet __$_i$]
That is it no.one REL knows
'That, there's no one who knows __.'

It is also the case that cross-dependency n-gram tracking could have helped with the data sparsity issue related to long-distance subject questions: Even though the right type of *wh*-extraction of an embedded subject was missing from our corpus, there were many examples of long-distance relativization of a subject. So, a modeled learner that ignored dependency type could use examples of long-distance relativization to infer that long-distance subject questions were also possible. Such *indirect* positive evidence has been used as part of successful strategies for learning other linguistic phenomena (Perfors et al. 2011; Pearl & Mis 2016).

Of course, collapsing across dependencies would mean that the modeled learner would lose the ability to model apparent fine-grained cross-dependency variation in island effects. For example, it would predict no island effects for RC-dependencies into relative clauses and *wh*-dependencies into embedded questions, contrary to findings in Kobzeva, Sant et al. (2022). If such a model were adopted, cross-dependency differences would need to be explained with supplemental constraints (as discussed by Kush et al. 2021). Alternatively, one could imagine a learner that assigned less weight to paths observed with a different dependency type in the probability calculation than direct evidence of paths observed with the dependency in question.

5.1.3. *Adjunct island*

Conditional adjuncts appear not to be islands for RC-dependencies in Norwegian (Kobzeva, Sant et al. 2022; Bondevik & Lohndal 2023), but our corpus contains no direct evidence for such dependencies. In the absence of direct evidence, the learner is unlikely to acquire the component n-grams required to represent the corresponding dependency path indirectly because, as with RCs, conditional adjuncts bear the label ADJUNCT$_{adv}$. The label is not unique to *conditional* adjuncts, but it is unique to clausal adjunct configurations on the whole. Therefore, the only evidence that would count toward learning the generalization is directly observing extraction from some type of clausal adjunct. No such examples exist in our corpus.

Though we do not find any dependencies in our children's corpus, Müller & Eggers (2022) reported 18 examples of RC-dependencies into conditional adjuncts in their broader sample of Danish mentioned above. It is difficult to estimate the exact frequency of relevant evidence from Müller & Eggers's results because we do not know the size of the Danish web, but extrapolating from the quantifiable subset of their data suggests that evidence is rare: Only four examples were found in the 57 million word set made up by KorpusDK and BySoc. The very low frequency of relevant examples in adult-directed language once again suggests that direct evidence likely does not occur frequently or reliably enough in the input to children to guarantee acquisition. If Norwegian children know that conditional adjuncts are not islands at an early age, there would seem to be a serious POS problem.

5.2. *Domain-specific information and formalism dependence*

A second debate that our study contributes to is what domain-specific information or biases a distributional learner might need to induce island effects from the input. Our learning model, like Pearl & Sprouse's, presupposes that the learner can parse the input into labeled, hierarchical representations *before* the acquisition of filler-gap constraints can begin. The learner is biased to attend to a specific subset of information contained within a larger representation (i.e., labels/nodes along a path between filler and gap) and to "chunk" that information into units of a particular size (trigrams or bigrams). We briefly consider the extent to which the assumptions or biases above implicate domain specificity.

The ability to categorize input and label hierarchically arranged objects is required for the acquisition of almost all syntactic phenomena. When it comes to phrase structure representations, there is a long-standing debate around whether such categories are innate or learned through distributional analysis (Cartwright & Brent 1997; Mintz 2003; Liang et al. 2022; Zhu & Clark 2022). Within LFG, it is commonly assumed that the set of f-structure features is universal and therefore innately specified (Dalrymple 2001; however, see Perfors et al. (2011) for a perspective that these structures do not require domain-specific category information). In keeping with the assumptions of the formalism, a learner that uses LFG-like representations would seem to rely on domain-specific category information.

Turning to the question of what kind of information the learner attends to and how that information is encoded, there is a clear need for domain-specific bias. Generally speaking, the success of distributional learning depends on the ability to represent certain crucial distinctions. Chiefly, the input must somehow distinguish between embedded declaratives, embedded questions, relative clauses, and adjunct clauses. Even if we remain agnostic as to whether clause-type information is innately specified or learned distributionally from cues in the input (lexical, syntactic, semantic, and pragmatic; see Geffen & Mintz 2015), the need for domain-specific bias relates to how clause-type information is encoded. Neither the standard Penn treebank annotation nor the LFG formalism encodes clause-type information in a default format that would be easily usable by a model tracking container node sequences. Both formalisms separate the label assigned to a complement clause from the information encoding the clause type. To import clause-type information into container node sequences, Pearl & Sprouse modified CP nodes with lexical information about the complementizer (e.g., *null* vs. *that* vs. *whether*). We annotated COMP labels with CLAUSE-TYPE values. Both choices represent a bias toward privileging a subset of information as relevant for the task, which does not follow straightforwardly from domain-general considerations alone. Similarly, though it is possible to learn that *wh*-dependencies and RC-dependencies are separate dependency types from distributional cues, our specification that the model track frequencies for the two dependency types independently represents a built-in bias.

The last built-in bias for our modeled learner is the n-gram window size. Pearl & Sprouse argued that the decision to use phrase structure trigams in their modeled learner arose out of empirical necessity. Windows larger than trigrams were inappropriate for modeling short dependencies, while smaller windows failed to capture some island effects. A unigram model could learn *wh*- and adjunct islands but not subject islands (or complex NP islands), and a bigram model was insufficient because there were no phrase structure bigrams that uniquely identified subject island violations to the exclusion of other grammatical FGDs. There was, however, a unique trigram for excluding subject islands (IP NP PP). As the authors rightfully noted, there is a trade-off between the complexity of the representational formalism and the size of the n-gram window needed to capture island effects: the simpler the formalism, the larger the window and vice versa. We showed that f-structure bigrams worked equally well as, if not better than, trigrams. This is because LFG's richer tagset afforded unique bigrams for all island violations, including those for subject islands (SUBJ ADJUNCT). Thus, a modeled learner based on LFG f-structure labels could use either trigrams or bigrams. Bigrams might arguably be easier to motivate as domain-general chunk size, as they constitute a representational minimum of what can be considered a "collocation."

Despite differences between LFG and PS formalisms, we suspect that most of the conclusions about the success and failures of learning Norwegian island constraints using container node n-grams would hold irrespective of the choice between the two formalisms. As noted above, any system that has nodes or n-grams that distinguish different clause types (including relative and adjunct clauses) and modification of subjects has the basic representational capacity to support the different generalizations that must be encoded. In many cases, both formalisms can make the relevant distinctions using unigrams, and where they cannot, there are unique bigram or trigram sequences for all remaining island types. The only substantive difference that we identified was related to the fact that LFG sometimes made finer-grained structural distinctions than PS representations would. Recall that the f-structures for long-distance subject questions differed depending on whether or not the embedded clause contained a tensed modal. In general, finer-grained structural distinctions entail that broad generalizations will be encoded in a larger set of possible container node sequences and that a narrower set of sentences will be direct evidence for each sequence. We saw one case where this difference impeded extracting the broader generalization from sparse input. Future research can investigate the consequences of such fine-grained differences in greater detail.

5.3. *Further limitations of the modeled learner*

We briefly address some additional limitations of the modeled learner presented here. As noted by Pearl & Sprouse, the learner is unable to induce human-like knowledge of filler-gap constructions where gaps inside islands become licensed by other gaps in the same sentence, such as parasitic gaps and across-the-board (ATB) extraction. In parasitic gap constructions like (31-a), there is a gap in an island environment (e.g., after *buying*) that is only acceptable when coupled with a licit gap in a non-island environment (e.g., after *try on*). In ATB constructions like (31-b), there are two (or more) gaps, each of which occurs in a conjunct of a coordinate structure. Each gap inside the coordinate structure would be illicit in isolation due to the coordinate structure constraint (Ross 1967), but the two gaps become licit in conjunction.

(31) a. This is the dress that Astrid did not try on __ [before buying __].

b. What dress did Astrid like __ and then buy __?

As it stands, the the modeled learner is unable to represent the contingent acceptability of the gaps inside islands. If the algorithm was exposed to examples of either (31-a) or (31-b), it would likely take the examples as evidence that single filler-gap dependencies into nonfinite adjuncts and coordinate structures were possible.

A major conceptual limitation of the algorithm is that it is clear that probability cannot be a proxy for intuitions of grammaticality. When making comparisons between lexically and length-matched sentences, the algorithm can reliably distinguish between ungrammatical island violations and grammatical structures. However, if we move away from closely matched comparison sentences, there are entire equivalence classes of very long grammatical sentences that the modeled learner would assign the same low probability to as it would assign to island violations. In this sense, the modeled learner fails to represent a core intuition that there is a fundamental distinction between extremely long filler-gap dependencies that are unattested because they are unlikely and dependencies that are outright ungrammatical (see Phillips 2013 for an extensive discussion).

It is also clear that probability is an imperfect proxy for acceptability, as well (see also Lau et al. 2017). We saw that there was, at times, reasonable qualitative alignment between the modeled learner and human judgments on whether there was a superadditive effect or not. However, quantitative alignment between empirical probabilities and acceptability is not possible on the absolute scale. Some of the quantitative misalignment can be attributed to the influence on the acceptability of lexical, semantic, processing, and pragmatic factors that the modeled learner, which only takes into account the dependency path, abstracts away from. A recent study modifying the current modeled learner (Dickson et al. 2022) showed that including more lexical properties (e.g., lexicalizing not only CP but also other phrasal categories such as VP) and allowing for different-sized fragments to represent FGDs can lead to more informative and efficient dependency representations. This modification could also capture factive and manner-of-speaking island effects (Liu et al. 2022) because through VP lexicalization, verb-frame frequency information is included in probability calculation. Nevertheless, there are other misalignments that potentially persist even when some of these other nuisance factors are taken into account. For example, it is unclear that the modeled learner reliably predicts differences in effect sizes between different island violations. This is most apparent in the case of so-called subliminal island effects (Almeida 2014), where a superadditive interaction is observed but the participants' average acceptability rating falls in the "acceptable" range (e.g., above $z = 0$). Some of the island effects in Kobzeva, Sant et al. (2022), such as the effect of *wh*-extraction from a conditional adjunct, can be categorized as subliminal effects, but the modeled learner essentially treats those violations as on par with other much larger effects. In other words, the current version of the modeled learner could benefit from an improved linking mechanism between the induced probabilities and human acceptability judgments. One potential way of enhancing our linking hypotheses involves incorporating lexical and/or semantic factors into the probability calculation, as proposed by Dickson et al. (2022). Other approaches could include integrating processing constraints (e.g., memory cost) to derive gradient acceptability judgments from categorical grammars (De Santo 2020) or introducing coercion of less acceptable elements into more acceptable ones to establish gradient grammaticality ranges (Villata & Tabor 2022). These modifications to the learner can be implemented through either one- or two-stage mechanisms, potentially shedding light on the interplay between categorical grammars and processing theories in modulating acceptability. We consider this line of inquiry a promising direction for future research.

6. Conclusion

We tested whether a distributional learner inspired by the learner of Pearl & Sprouse (2013b) could learn islands facts in Norwegian by tracking n-grams of container nodes along dependency paths. We used LFG f-structure labels as our container nodes. On the one hand, we found that the proposed learning strategy can capture some patterns of island-insensitivity in Norwegian when examples of "island-violating" dependencies are in the input. On the other hand, it failed to learn other important patterns because the relevant dependencies are unattested in children's input. Our findings suggest that given limited input data, a simple n-gram-based distributional learning over structured LFG representations may not be sufficient to fully recover human-like knowledge of filler-gap dependency relations and island constraints cross-linguistically.

Acknowledgments

We are grateful to the two reviewers whose feedback and comments allowed us to significantly improve the manuscript. We thank Paul Meurer and Helge Dyvik for their assistance with querying and downloading the corpus data. We also thank members of EyeLands Lab (Øyelab) and the former AcqVA group at NTNU for their comments on earlier versions of this work. All materials and analysis scripts are available in the following OSF repository: https://osf.io/24zhg/.

Disclosure statement

No potential conflicts of interest are reported by the author(s).

ORCID

Anastasia Kobzeva http://orcid.org/0000-0001-7810-4959

Declaration of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article.

References

Abeillé, Anne, Barbara Hemforth, Elodie Winckel & Edward Gibson. 2020. Extraction from subjects: Differences in acceptability depend on the discourse function of the construction. *Cognition* 204. 104293. https://doi.org/10.1016/j.cognition.2020.104293.

Almeida, Diogo. 2014. Subliminal wh-islands in Brazilian Portuguese and the consequences for syntactic theory. *Revista da ABRALIN* 13(2). 55–93.

Perfors, Amy, Joshua B. Tenenbaum & Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition* 118 (3). 306–338. https://doi.org/10.1016/j.cognition.2010.11.001, http://dx.doi.org/10.1016/j.cognition.2010.11.001.

Bondevik, Ingrid, Dave Kush & Terje Lohndal. 2021. Variation in adjunct islands: The case of Norwegian. *Nordic Journal of Linguistics* 44 (3). 223–254. https://doi.org/10.1017/S0332586520000207.

Bondevik, Ingrid & Terje Lohndal. 2023. Extraction from finite adjunct clauses: An investigation of relative clause dependencies in Norwegian. *Glossa: A Journal of General Linguistics* 8(1). 1–41. https://doi.org/10.16995/glossa.9033.

Bresnan, Joan, Ash Asudeh, Ida Toivonen & Stephen Wechsler. 2015. *Lexical-functional syntax*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119105664.

Brown, Roger. 1973. *A first language: The early stages*. Cambridge, MA: Harvard University Press. https://doi.org/10.4159/harvard.9780674732469.

Cameron-Faulkner, Thea & Claire Noble. 2013. A comparison of book text and child directed speech. *First Language* 33. 268–279. https://doi.org/10.1177/0142723713487613.

Cartwright, Timothy A. & Michael R. Brent. 1997. Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition* 63(2). 121–170.

Chater, Nick, Alexander Clark, John A. Goldsmith & Amy Perfors. 2015. *Empiricism and language learnability*, 1st edn. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198734260.001.0001.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge: the MIT press.

Chomsky, Noam. 1973. Conditions on transformations. In Morris Halle, Stephen R. Anderson & Paul Kiparsky (eds.), *A Festschrift for Morris Halle*, 232–286. New York: Holt, Rinehart and Winston.

Chomsky, Noam. 1986. *Knowledge of language: Its nature, origin, and use*. Westport, CT: Praeger.

Chomsky, Noam. 2001. Derivation by phase. In Michael Kenstowicz (ed.), *Ken Hale: A life in language*, 1–52. Cambridge, MA: The MIT Press.

Chowdhury, Shammur Absar & Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In E. M. Bender, L. Derczynski, and P. Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, 133–144. Santa Fe, NM: Association for Computational Linguistics. https://aclanthology.org/C18-1012.

Christensen, Ken Ramshøj, Johannes Kizach & Anne Mette Nyvad. 2013. Escape from the island: Grammaticality and (reduced) acceptability of wh-island violations in Danish. *Journal of Psycholinguistic Research* 42. 51–70. https://doi.org/10.1007/s10936-012-9210-x.

Christensen, Ken Ramshøj & Anne Mette Nyvad. 2014. On the nature of escapable relative islands. *Nordic Journal of Linguistics* 37(1). 29–45. https://doi.org/10.1017/S0332586514000055.

Christensen, Kirsti Koch. 1982. On multiple filler-gap constructions in Norwegian. In Elisabet Engdahl & Eva Ejerhed, *Readings on unbounded dependencies in Scandinavian languages*, 77–98. Stockholm: Almquist & Wiksell.

Clark, Alexander & Shalom Lappin. 2010. *Linguistic nativism and the poverty of the stimulus*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781444390568.

Crain, Stephen & Paul Pietroski. 2001. Nature, nurture and universal grammar. *Linguistics and Philosophy* 24(2). 139–186.

Cuneo, Nicole & Adele E. Goldberg. 2023. The discourse functions of grammatical constructions explain an enduring syntactic puzzle. *Cognition* 240. 105563. https://doi.org/10.1016/j.cognition.2023.105563, https://www.sciencedirect.com/science/article/pii/S001002772300197X.

Cuskley, Christine,Rebecca Woods &d Molly Flaherty. 2024. The limitations of large language models for understanding human language and cognition. *Open Mind* 8. 1058–1083. https://doi.org/10.1162/opmi_a_00160.

Dalrymple, Mary. 2001. *Lexical functional grammar*. San Diego, CA: Academic Press.

Dalrymple, Mary, John J. Lowe & Louise Mycock. 2019. *The Oxford reference guide to lexical functional grammar*. Oxford, UK: Oxford University Press.

De Santo, Aniello. 2020. MG parsing as a model of gradient acceptability in syntactic islands. In A. Ettinger , G. Jarosz, and M. Nelson (eds.), *Proceedings of the Society for Computation in Linguistics 2020*, 53–63. Amherst, MA: University of Massachusetts Amherst. https://doi.org/10.7275/srck-2j50.

De Villiers, Jill, Thomas Roeper, Linda Bland-Stewart & Barbara Pearson. 2008. Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics* 29(1). 67–103. https://doi.org/10.1017/S0142716408080041.

DiCiccio, Thomas J. & Bradley Efron. 1996. Bootstrap confidence intervals. *Statistical Science* 11(3). 189–228. https://doi.org/10.1214/ss/1032280214.

Dickson, Niels, Lisa Pearl & Richard Futrell. 2022. Learning constraints on wh-dependencies by learning how to efficiently represent wh-dependencies: A developmental modeling investigation with fragment grammars. *Proceedings of the Society for Computation in Linguistics* 5(1). 220–224. https://doi.org/10.7275/7fd4-fw49.

Dunbar, Ewan. 2019. Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language* 95(1). e87–e98.

Dyvik, Helge, Paul Meurer, Victoria Rosén, Koenraad De Smedt, Petter Haugereid, Gyri Smørdal Losnegaard, Gunn Inger Lyse & Martha Thunes. 2016. NorGramBank: A ‘deep’ treebank for Norwegian. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, 3555–3562. Portorož, Slovenia: European Language Resources Association (ELRA). https://aclanthology.org/L16-1565 .

Engdahl, Elisabet. 1982. Restrictions on unbounded dependencies in Swedish. In Elisabet Engdahl & Eva Ejerhed (eds.), *Readings on unbounded dependencies in Scandinavian languages*, 151–174. Stockholm: Almquist & Wiksell.

Engdahl, Elisabet. 1997. Relative clause extractions in context. *Working Papers in Scandinavian Syntax* 60. 51–79.

Erteschik-Shir, Nomi. 1973. *On the nature of island constraints*. Cambridge, MA: MIT dissertation.

Evanson, Linnea, Yair Lakretz & Jean-Rémi King. 2023. Language acquisition: Do children and language models follow similar learning stages? *arXiv preprint arXiv:2306.03586*.

Falk, Yehuda. 2011. *Lexical-functional grammar*. Oxford, UK: Oxford University Press.

Geffen, Susan & Toben H. Mintz. 2015. Can you believe it? 12-month-olds use word order to distinguish between declaratives and polar interrogatives. *Language Learning and Development* 11(3). 270–284.

Goldberg, Adele. 1995. *Constructions: a construction grammar approach to argument structure*. Chicago, IL: The University of Chicago Press.

Goldberg, Adele. 2006. *Constructions at work: The nature of generalization in language*. Oxford, UK: Oxford University Press.

Howitt, Katherine, Sathvik Nair, Allison Dods & Robert Melvin Hopkins. 2024. Generalizations across filler-gap dependencies in neural language models. In Libby Barak & Malihe Alikhani (eds.), *Proceedings of the 28th Conference on Computational Natural Language Learning*, 269–279. Miami, FL: Association for Computational Linguistics.https://aclanthology.org/2024.conll-1.21.

Huang, Cheng-Teh James. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: MIT dissertation.

Kaplan, Ronald M. & Joan Bresnan. 1995. Formal system for grammatical representation. In M. Dalrymple, R. M. Kaplan, J. T. Maxwell III, and A. Zaenen (eds.), *Formal issues in lexical-functional grammar*, 29–130. Stanford, CA: CSLI Publications.

Kaplan, Ronald M. & Annie Zaenen. 1989. Long-distance dependencies, constituent structure, and functional uncertainty. In M. R. Baltin and A. S. Kroch (eds.), *Alternative conceptions of phrase structure*, 17–42. Chicago, IL: The University of Chicago Press.

Katzir, Roni. 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. *Biolinguistics* 17. 1–12. https://doi.org/10.5964/bioling.13153.

Kobzeva, Anastasia, Suhas Arehalli, Tal Linzen & Dave Kush. Manuscript submitted for publication. Learning filler-gap dependencies with Neural language models: Testing island sensitivity in Norwegian and English. Trondheim, Norway: Norwegian University of Science and Technology.

Kobzeva, Anastasia, Suhas Arehalli, Tal Linzen & Dave Kush. 2022. LSTMs can learn basic Wh-and relative clause dependencies in Norwegian. In J. Culbertson, A. Perfors, H. Rabagliati, and V. Ramenzoni (eds.), *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2974–2980. Oakland, CA: escholarship, California Digital Library. https://escholarship.org/uc/item/012683gb.

Kobzeva, Anastasia, Suhas Arehalli, Tal Linzen & Dave Kush. 2023. Neural networks can learn patterns of island-insensitivity in Norwegian. In T. Hunter and B. Prickett (eds.), *Proceedings of the Society for Computation in Linguistics* 6(1). 175–185. Amherst, MA: University of Massachusetts Amherst. https://doi.org/10.7275/qb8z-qc91 .

Kobzeva, Anastasia, Charlotte Sant, Parker T. Robbins, Myrte Vos, Terje Lohndal & Dave Kush. 2022. Comparing island effects for different dependency types in Norwegian. *Languages* 7(3). 195–220. https://doi.org/10.3390/languages7030195.

Kodner, Jordan, Sarah Payne & Jeffrey Heinz. 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). *arXiv preprint arXiv:2308.03228*.

Kush, Dave & Anne Dahl. 2020. L2 Transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research* 38(2). 1–32. https://doi.org/10.1177/0267658320956704.

Kush, Dave, Terje Lohndal & Jon Sprouse. 2018. Investigating variation in island effects: A case study of Norwegian Wh-extraction. *Natural Language & Linguistic Theory* 36(3). 743–779. https://doi.org/10.1007/s11049-017-9390-z.

Kush, Dave, Terje Lohndal & Jon Sprouse. 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language* 95(3). 393–420. https://doi.org/10.1353/lan.2019.0051.

Kush, Dave, Akira Omaki & Norbert Hornstein. 2013. Microvariation in islands? In Jon Sprouse & Norbert Hornstein (eds.), *Experimental syntax and island effects*, 239–264. Cambridge: Cambridge University Press.

Kush, Dave, Charlotte Sant & Sunniva Briså Strætkvern. 2021. Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. *Glossa: A Journal of General Linguistics* 6(1). 1–50. https://doi.org/10.16995/glossa.5774.

Lan, Nur, Emmanuel Chemla & Roni Katzir. 2024. Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry* 1–56. https://doi.org/10.1162/ling_a_00533.

Lau, Jey Han, Alexander Clark & Shalom Lappin. 2017. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive science* 41(5). 1202–1241.

Laurence, Stephen & Eric Margolis. 2001. The poverty of the stimulus argument. *British Journal for the Philosophy of Science* 52(2). 217–276.

Liang, Kevin, Diana Marsala & Charles Yang. 2022. Distributional learning of syntactic categories. In Ying Gong & Felix Kpogo (eds.), *Proceedings of the 46th annual Boston University Conference on Language Development*, 442–455. Somerville, MA: Cascadilla Press.

Lindahl, Filippa. 2017. *Extraction from relative clauses in Swedish*. Gothenburg, Sweden: University of Gothenburg dissertation.

Liu, Yingtong, Elodie Winckel, Anne Abeillé, Barbara Hemforth & Edward Gibson. 2022. Structural, functional, and processing perspectives on linguistic island effects. *Annual Review of Linguistics* 8. 495–525.

MacWhinney, Brian. 2000. *The CHILDES project: The database*. Vol. 2. Hove, UK: Psychology Press.

MacWhinney, Brian & Johannes Wagner. 2010. Transcribing, searching and data sharing: The CLAN software and the TalkBank data repository. *Gesprachsforschung: Online-Zeitschrift zur verbalen Interaktion* 11. 154.

Marcus, Mitchell, Beatrice Santorini& Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330. https://aclanthology.org/J93-2004.

Marr, David. 1982. *Vision*. Cambridge, MA: MIT Press.

McCloskey, Michael. 1991. Networks and theories: The place of connectionism in cognitive science. *Psychological Science* 2(6). 387–395. https://www.jstor.org/stable/pdf/40062715.pdf.

Mesmer, Heidi Anne E. 2016. Text matters: exploring the lexical reservoirs of books in preschool rooms. *Early Childhood Research Quarterly* 34. 67–77. https://doi.org/10.1016/j.ecresq.2015.09.001.

Mintz, Toben H. 2003. Frequent frames as a cue for grammatical categories in child-directed speech. *Cognition* 90(1). 91–117.

Montag, Jessica L. 2019. Differences in sentence complexity in the text of children's picture books and child-directed speech. *First Language* 39. 527–546. https://doi.org/10.1177/0142723719849996.

Montag, Jessica L. & Maryellen C. MacDonald. 2015. Text exposure predicts spoken production of complex sentences in eight and twelve year old children and adults. *Journal of Experimental Psychology: General* 144. 447–468. https://doi.org/10.1037/xge0000054.

Müller, Christiane & Clara Ulrich Eggers. 2022. Island extractions in the wild: A corpus study of adjunct and relative clause islands in Danish and English. *Languages* 7(2). 125. https://doi.org/10.3390/languages7020125.

Noble, Claire H., Thea Cameron-Faulkner & Elena Lieven. 2018. Keeping it simple: The grammatical properties of shared book reading. *Journal of Child Language* 45(3). 753–766.

Pearl, Lisa. 2022. Poverty of the stimulus without tears. *Language Learning and Development* 18(4). 415–454.

Pearl, Lisa & Alandi Bates. 2022. A new way to identify if variation in children's input could be developmentally meaningful: Using computational cognitive modeling to assess input across socio-economic status for syntactic islands. *Journal of Child Language* 51(4). 800–833. https://doi.org/10.1017/S0305000922000514.

Pearl, Lisa & Benjamin Mis. 2016. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric 'one'. *Language* 92(1). 1–30.

Pearl, Lisa & Jon Sprouse. 2013a. Computational models of acquisition for islands. In Jon Sprouse & Norbert Hornstein (eds.), *Experimental syntax and island effects*, 109–131. Cambridge, UK: Cambridge University Press.

Pearl, Lisa & Jon Sprouse. 2013b. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition* 20(1). 23–68.

Phillips, Colin. 2013. On the nature of island constraints II: Language learning and innateness. In Jon Sprouse and Norbert Hornstein (eds.), *Experimental syntax and island effects*, 132–158. Cambridge, UK: Cambridge University Press.

Piantadosi, Steven T. 2023. Modern language models refute Chomsky's approach to language. In Edward Gibson and Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett*, 353–414. Berlin, Germany: Language Science Press.

Pollard, Carl & Ivan A. Sag. 1994. *Head-driven Phrase Structure Grammar*. University of Chicago Press.

Pullum, Geoffrey K. & Barbara C Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19(1–2). 9–50.

Rizzi, Luigi. 1982. *Issues in Italian syntax*, 49–76. Dordrecht: Foris.

Rosén, Victoria, Paul Meurer & Koenraad De Smedt. 2009. LFG Parsebanker: A toolkit for building and searching a treebank as a parsed corpus. In Frank, Van Eynde, Anette Frank, Koenraad de Smedt & Gertjan van Noord (eds.), *Proceedings of the Seventh International Workshop on Treebanks and Linguistic Theories, Groningen, The Netherlands, January 23-24, 2009*, 127–133. Amsterdam: LOT.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation. https://dspace.mit.edu/handle/1721.1/15166.

Sprouse, Jon. 2007. *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge*. College Park, MD: University of Maryland dissertation. https://drum.lib.umd.edu/handle/1903/7283.

Sprouse, Jon, Ivano Caponigro, Ciro Greco & Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory* 34(1). 307–344. https://doi.org/10.1007/s11049-015-9286-8, https://doi.org/10.1007/s11049-015-9286-8.

Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1). 82–123. https://muse.jhu.edu/article/469088 .

Suppes, Patrick. 1974. The semantics of children's language. *American Psychologist* 29(2). 103. https://doi.org/10.1037/h0036026.

Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition* 40(1–2). 21–81. https://doi.org/10.1016/0010-0277(91)90046-7.

Villata, Sandra & Whitney Tabor. 2022. A self-organized sentence processing theory of gradience: The case of islands. *Cognition* 222. 104943.

Vincent, Jake Wayne. 2021. *Extraction from relative clauses: An experimental investigation into variable island effects in English - or - this is a dissertation that we really needed to find someone who'd write*. University of California Santa Cruz dissertation.

Warstadt, Alex & Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In Shalom Lappin and Jean-Philippe Bernardy (eds.), *Algebraic Structures in Natural Language*, 17–60. Boca Raton, FL: CRC Press.

Warstadt, Alex, Leshem Choshen, Aaron Mueller, Adina Williams, Ethan Wilcox & Chengxu Zhuang. 2023. Call for papers—The BabyLM challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2301.11796*.

Wilcox, Ethan, Richard Futrell & Roger Levy. 2021. Using computational models to test syntactic learnability. *Linguistic Inquiry* 55(4). 805–848. https://doi.org/10.1162/ling_a_00491.

Wilcox, Ethan, Roger Levy, Takashi Morita & Richard Futrell. 2018. What do RNN language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.

Zhu, Haiting & Alexander Clark. 2022. Distributional lattices as a model for discovering syntactic categories in child-directed speech. *Journal of Psycholinguistic Research* 51(4). 917–931.

Appendices

Appendix A. Item examples from Kobzeva, Sant et al. (2022)

A.1. *Subject island*

(32) a. **Short x No Island x *Wh*-dependency**

Hvilke aktivister er redde for at fabrikken skader miljøet?
which activists are worried C factory.DEF harms environment.DEF
'Which activists are worried that the factory is harming the environment?'

b. **Short x No Island x RC-dependency**

Det er aktivistene som er redde for at fabrikken skader miljøet.
those are activists.DEF REL are worried C factory.DEF harms environment.DEF
'Those are the activists that are worried that the factory is harming the environment.'

c. **Long x No Island x *Wh*-dependency**

Hvilken fabrikk er aktivistene redde for at skader miljøet?
which factory are activists.DEF worried C harms environment.DEF
'Which factory are the activists worried __ is harming the environment?'

d. **Long x No Island x RC-dependency**

Det er fabrikken som aktivistene er redde for at skader miljøet.
that is factory.DEF REL activists.DEF are worried C harms environment.DEF
'That's the factory that the activists worry __ is harming the environment.'

e. **Short x Island x *Wh*-dependency**

Hvilke aktivister er redde for at avfall fra fabrikken skader miljøet?
which activists are worried C waste from factory.DEF harms environment.DEF
'Which activists are worried that waste from the factory is harming the environment?'

f. **Short x Island x RC-dependency**

Det er aktivistene som er redde for at avfall fra fabrikken skader
those are activists.DEF that are worried C waste from factory.DEF harms
miljøet.
environment.DEF
'Those are the activists that are worried that waste from the factory is harming the environment.'

g. **Long x Island x *Wh*-dependency**

Hvilken fabrikk er aktivistene redde for at avfall fra skader miljøet?
which factory are activists.DEF worried C waste from harms environment.DEF
'Which factory are the activists worried that waste from __ harms the environment?'

h. **Long x Island x RC-dependency**

Det er fabrikken som aktivistene er redde for at avfall fra skader
that is factory.DEF that activists.DEF are worried C waste from harms
miljøet.
environment.DEF
'That's the factory that the activists are worried that waste from __ is harming the environment.'

A.2. *Relative clauses*

(33) a. **Short x No Island x *Wh*-dependency**

Hvilken servitør sa at mange bestilte ølet?
which waiter said C many ordered beer.DEF
'Which waiter said that many people ordered the beer?'

b. **Short x No Island x RC-dependency**

Det var servitøren som sa at mange bestilte ølet.
that was waiter.DEF REL said that many ordered beer.DEF
'That was the waiter that said that many people ordered the beer.'

c. **Long x No Island x *Wh*-dependency**

Hvilket øl sa servitøren at mange bestilte?
which beer said waiter.DEF C many ordered
'Which beer did the waiter say many people ordered __?'

d. **Long x No Island x RC-dependency**

Det var ølet som servitøren sa at mange bestilte.
that was beer.DEF REL waiter.DEF said C many ordered
'That was the beer that the waiter said many people ordered __.'

e. **Short x Island x *Wh*-dependency**

Hvor mange var det som bestilte ølet?
how many was it REL ordered beer.DEF
'How many were there that ordered the beer?'

f. **Short x Island x RC-dependency**

Det var mange som bestilte ølet.
it was many REL ordered beer.DEF
'There were many people that ordered the beer.'

g. **Long x Island x *Wh*-dependency**

Hvilket øl var det mange som bestilte?
which beer was it many REL ordered
'Which beer were there many people that ordered __?'

h. **Long x Island x RC-dependency**

Det var ølet som det var mange som bestilte.
that was beer.DEF REL it was many REL ordered
'That was the beer that there were many people that ordered __.'

A.3. *Embedded questions*

(34) a. **Short x No Island x *Wh*-dependency**

Hvilken snekker sa at hylla skulle monteres i stuen?
which carpenter said C shelf.DEF should install.pass in living.room.DEF
'Which carpenter said that the shelf should be installed in the living room?'

b. **Short x No Island x RC-dependency**

Det var snekkeren som sa at hylla skulle monteres i stuen.
that was carpenter.DEF REL said C shelf.DEF should install.pass in living.room.DEF
'That was the carpenter that said that the shelf should be installed in the living room.'

c. **Long x No Island x *Wh*-dependency**

Hvilken hylle sa snekkeren at skulle monteres i stuen?
which shelf said carpenter.DEF C should install.pass in living.room.DEF
'Which shelf did the carpenter say __ should be installed in the living room?'

d. **Long x No Island x RC-dependency**

Det var hylla som snekkeren sa at skulle monteres i stuen.
that was shelf.DEF REL carpenter.DEF said C should install.pass in living.room.DEF
'That was the shelf that the carpenter said __ should be installed in the living room.'

e. **Short x Island x *Wh*-dependency**

Hvilken snekker sa hvor hylla skulle monteres?
which carpenter said where shelf.DEF should install.pass
'Which carpenter said where the shelf should be installed?'

f. **Short x Island x RC-dependency**

Det var snekkeren som sa hvor hylla skulle monteres.
that was carpenter.DEF REL said where shelf.DEF should install.pass
'That was the carpenter that said where the shelf should be installed.'

g. **Long x Island x *Wh*-dependency**

Hvilken hylle sa snekkeren hvor skulle monteres?
which shelf said carpenter.DEF where should install.pass
'Which shelf did the carpenter say where __ should be installed?'

h. **Long x Island x RC-dependency**

Det var hylla som snekkeren sa hvor skulle monteres.
that was shelf.DEF that carpenter.DEF said where should install.pass
'That was the shelf that the carpenter said where __ should be installed.'

A.4. *Adjunct island*

(35) a. **Short x No Island x *Wh*-dependency**

Hvilken kokk misliker at hun bruker den skarpe kniven?
which chef dislikes C she uses the sharp knife.DEF
‘Which chef dislikes that she uses the sharp knife?’

b. **Short x No Island x RC-dependency**

Det er kokken som misliker at hun bruker den skarpe kniven.
that is chef.DEF REL dislikes C she uses the sharp knife.DEF
‘That’s the chef that dislikes that she uses the sharp knife.’

c. **Long x No Island x *Wh*-dependency**

Hvilken kniv misliker kokken at hun bruker?
which knife dislikes chef.DEF C she uses
‘Which knife does the chef dislike that she uses __?’

d. **Long x No Island x RC-dependency**

Det er kniven som kokken misliker at hun bruker.
that is knife.DEF REL chef.DEF dislikes C she uses
‘That’s the knife that the chef dislikes that she uses __.’

e. **Short x Island x *Wh*-dependency**

Hvilken kokk blir sur om hun bruker den skarpe kniven?
which chef gets angry if she uses the sharp knife.DEF
‘Which chef gets angry if she uses the sharp knife?’

f. **Short x Island x RC-dependency**

Det er kokken som blir sur om hun bruker den skarpe kniven.
that is chef.DEF REL gets angry if she uses the sharp knife.DEF
‘That’s the chef that gets angry if she uses the sharp knife.’

g. **Long x Island x *Wh*-dependency**

Hvilken kniv blir kokken sur om hun bruker?
which knife gets chef.DEF angry if she uses
‘Which knife does the chef get angry if she uses __?’

h. **Long x Island x RC-dependency**

Det er kniven som kokken blir sur om hun bruker.
that is knife.DEF REL chef.DEF gets angry if she uses
‘That’s the knife that the chef gets angry if she uses __.’

Appendix B. Other comparisons

Here we provide modeling results for different potential island constituents compared to human data from other acceptability judgment studies (when available).

B.1. *CNP island*

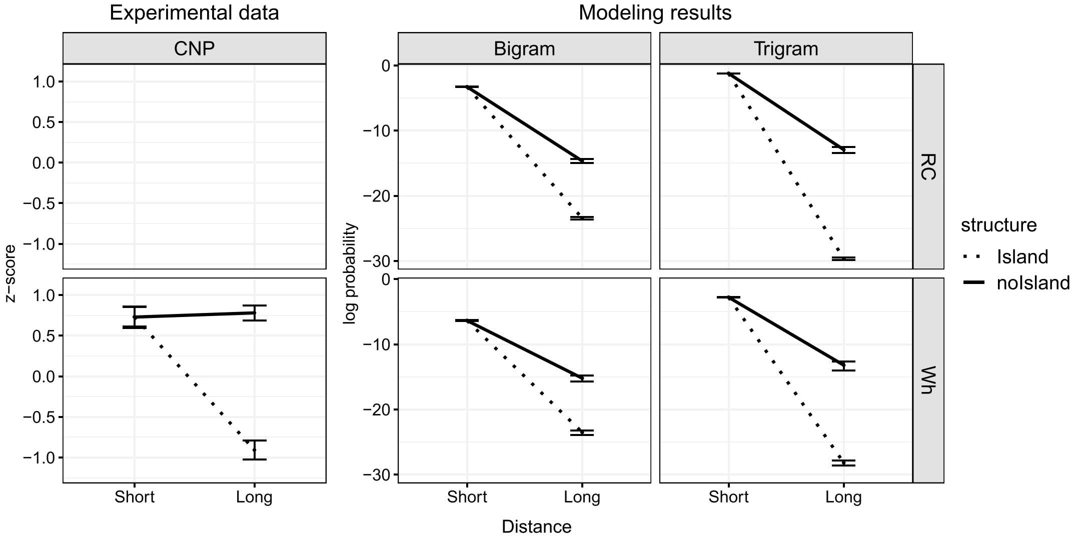


Figure B1. Interaction plots for modeling results and judgment data from (Kush et al. 2018, Experiment 1) for CNP island. Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

B.2. *Embedded questions, object extraction*

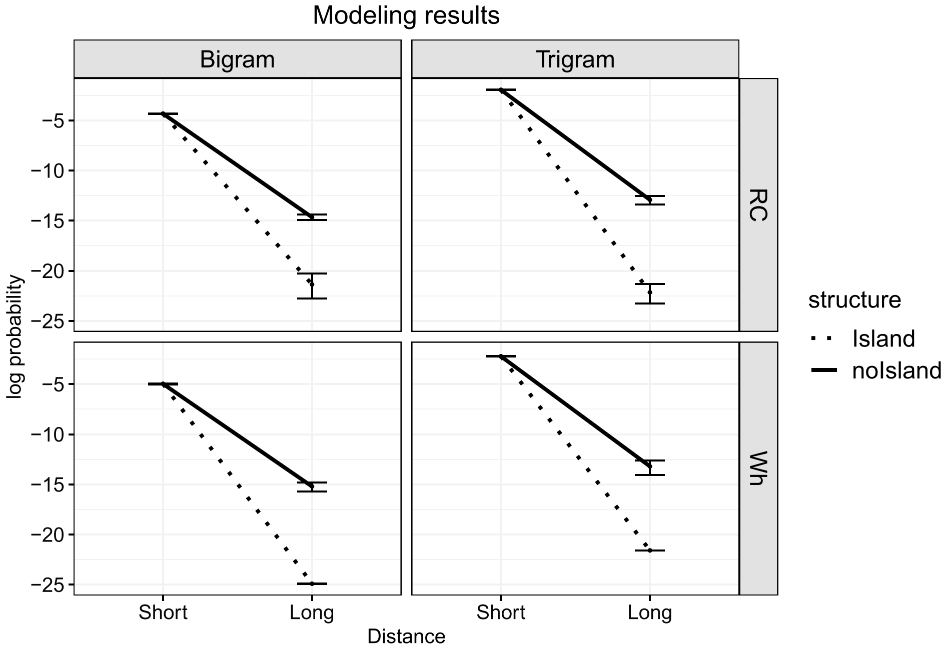


Figure B2. Interaction plots for modeling results for EQs (object position). Error bars indicate bootstrapped 95% confidence intervals.

B.3. *Whether-clauses, subject extraction*

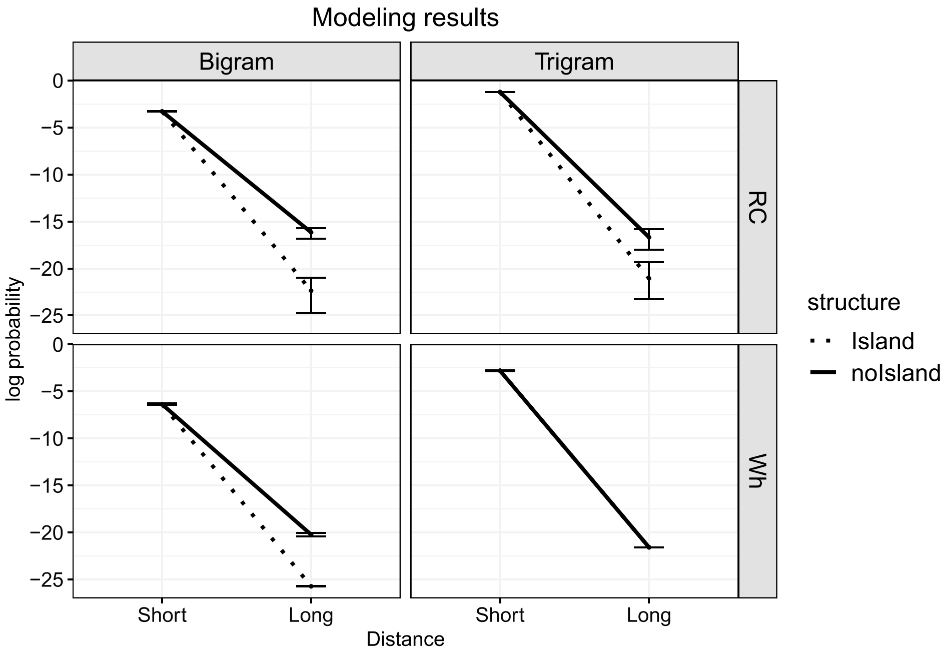


Figure B3. Interaction plots for modeling results for *Whether*-clauses (subject position). Error bars indicate bootstrapped 95% confidence intervals.

B.4. *Whether-clauses, object extraction*

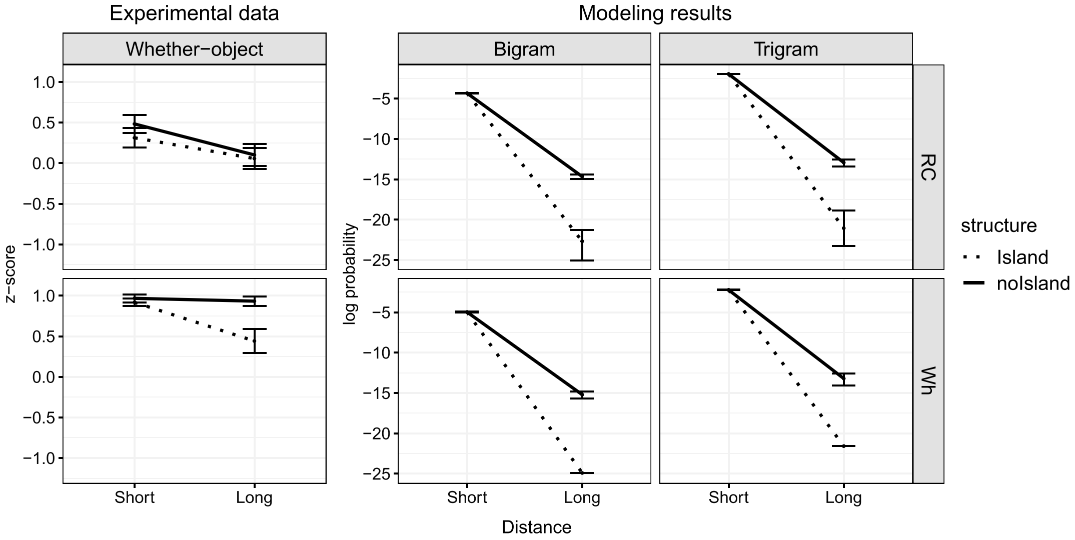


Figure B4. Interaction plots for modeling results and judgment data from (Bondevik & Lohndal 2023, RC-dependencies) and (Kush et al. 2018, *wh*-dependencies) for *Whether*-clauses (object position). Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

B.5. *Relative clauses, extraction from a PP-complement*

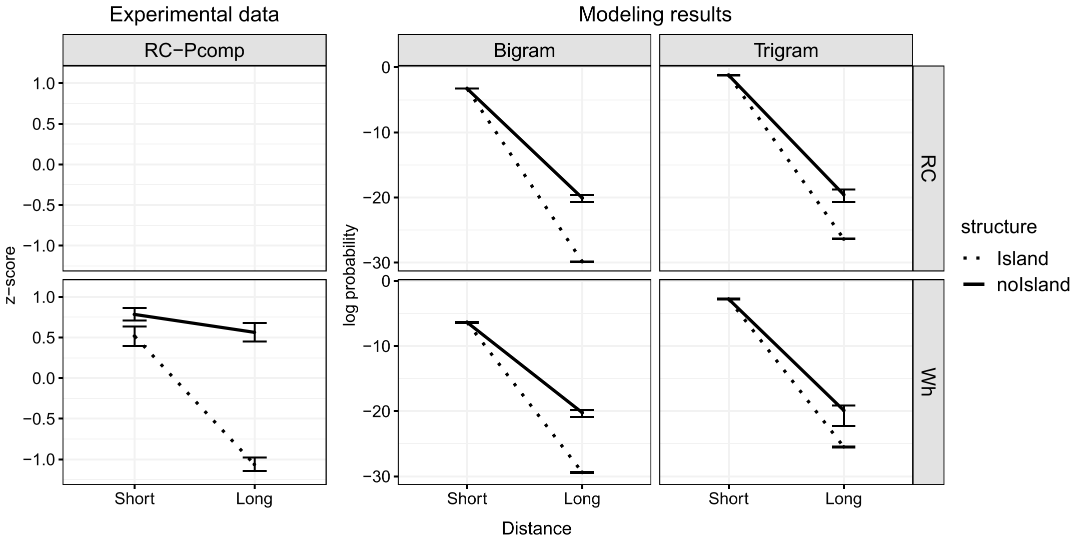


Figure B5. Interaction plots for modeling results and judgment data from (Kush et al. 2018, Experiment 2) for RCs (PP-complement position). Error bars indicate 95% confidence intervals (observed for the experimental data and bootstrapped for the modeling data).

Appendix C. Density plots

In Figure 15, we plot the density distributions of z-scored human ratings and modeling results, split by experimental factors distance (with levels shown on different panels) and structure (different line types). Additionally, the modeling results are split by bigram- and trigram-based learners (middle and lower rows, respectively). The shaded areas represent the range of probabilities assigned to unattested structures (all structures in *Short* conditions are attested).

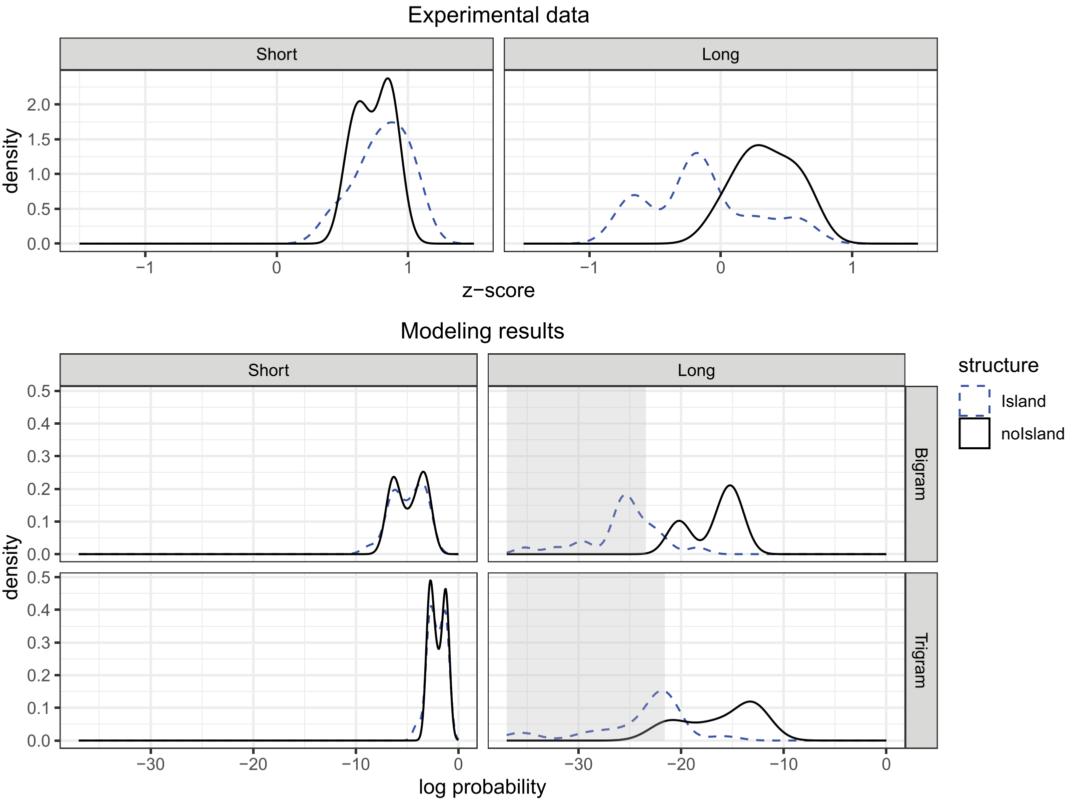


Figure C1. Density plots of z-scored human ratings and modeled probabilities.

Article J3

Kobzeva, A., Arehalli, S., Linzen, T. & Kush, D. (2025). Learning Filler-Gap Dependencies with Neural Language Models: Testing Island Sensitivity in Norwegian and English. *Accepted to the Journal of Memory and Language.*

Learning Filler-Gap Dependencies with Neural Language Models: Testing Island Sensitivity in Norwegian and English

Anastasia Kobzeva[a], Suhas Arehalli[b], Tal Linzen[c], Dave Kush[d]

[a] *Department of Language and Literature, Norwegian University of Science and Technology, Trondheim, Norway*
[b] *Macalester College, Saint Paul, USA*
[c] *New York University, New York City, USA*
[d] *University of Toronto, Toronto, Canada*

Abstract

Human linguistic input is often claimed to be impoverished with respect to linguistic evidence for complex structural generalizations that children induce. The field of language acquisition is currently debating the ability of various learning algorithms to accurately derive target generalizations from the input. A growing body of research explores whether Neural Language Models (NLMs) can induce human-like generalizations about filler-gap dependencies (FGDs) in English, including island constraints on their distribution. Based on positive results for select test cases, some authors have argued that the relevant generalizations can be learned without domain-specific learning biases (Wilcox et al., 2023), though other researchers dispute this conclusion (Lan et al., 2024b; Howitt et al., 2024). Previous work focuses solely on English, but broader claims about filler-gap dependency learnability can only be made based on multiple languages and dependency types. To address this gap, we compare the ability of NLMs to learn restrictions on FGDs in English and Norwegian. Our results are mixed: they show that although these models acquire some sophisticated generalizations about filler-gap dependencies in the two languages, their generalizations still diverge from those of humans. When tested on structurally complex environments, the models sometimes adopt narrower generalizations than humans do or overgeneralize beyond their input in non-human-like ways. We conclude that current evidence does not support the claim that FGDs and island constraints on them can be learned without domain-specific biases.

Keywords: Filler-gap dependencies, Island constraints, Learnability, Neural language models, Norwegian

1. Introduction

Children acquire language rapidly and relatively effortlessly despite the fact that linguistic competence requires complex and abstract generalizations. The field of language acquisition is currently debating the ability of various learning algorithms to

Preprint accepted to Journal of Memory and Language

accurately derive target generalizations from the input. One central issue is the relative contribution that language-specific and domain-general mechanisms and biases make to the learning process. The nativist tradition has assumed that domain-general learning procedures and biases alone are insufficient to guarantee the acquisition of the full range of generalizations that humans come to master from an impoverished input. In order to overcome the Poverty of the Stimulus (henceforth POS), domain-general procedures must be supplemented by innate, language-specific biases (Chomsky, 1965; Phillips, 2013a; Lasnik and Lidz, 2016; Crain and Pietroski, 2001). An alternative, empiricist view holds that acquisition need only rely on domain-general biases and learning mechanisms, while relevant domain-specific information can be derived from linguistic exposure (Clark and Lappin, 2010; Reali and Christiansen, 2005; Perfors et al., 2011; Clark and Lappin, 2012; Landauer and Dumais, 1997; Christiansen and Chater, 2016). A recent series of studies has sought to contribute to this debate by exploring whether Neural Language Models (NLMs) without substantial linguistic biases can induce complex linguistic generalizations from the input they receive.

NLMs produce probability distributions over word sequences based on a corpus. In recent years, researchers have started using these systems to explore the types of generalizations that can be induced based on the statistical regularities of the input. Since the nature of representations learned by NLMs is not yet properly understood, the models are typically evaluated through behavioral experiments that examine whether the probabilities assigned by the models to minimal pairs of sentences, one grammatical and one ungrammatical, align with sentence acceptability. In this way, NLMs serve as proxies for learners with minimal linguistic bias. Proponents of this approach hold that NLM simulations provide a proof of concept for what can *in principle* be acquired by domain-general learning procedures alone (Wilcox et al., 2023).[1]

Since the early explorations of NLMs' linguistic abilities (Linzen et al., 2016; Bernardy and Lappin, 2017; Gulordava et al., 2018), many studies have uncovered impressive performance on certain structure-dependent linguistic phenomena (Hu et al. 2020; Linzen and Baroni 2021; Lake and Baroni 2023; Ahuja et al. 2024, a.o.). *Filler-Gap Dependencies* (FGDs), the focus of the present paper, are one such phenomenon. A growing body of work explores the potential of NLMs to induce complex rules about FGDs, including certain restrictions called *island constraints*, which we discuss in more detail shortly (Chowdhury and Zamparelli, 2018; Wilcox et al., 2019a,b, 2023; Chaves, 2020; Bhattacharya and van Schijndel, 2020; Ozaki et al., 2022; Suijkerbuijk et al., 2023; Lan et al., 2024b; Howitt et al., 2024). We extend this line of research by exploring whether NLMs can learn complex properties of FGDs and patterns of cross-linguistic variation in island facts from exposure to Norwegian and English text.

[1] Others hold that NLMs can even implement genuine theories of language (Piantadosi, 2023) — a view that has recently received much critique (Katzir 2023; Kodner et al. 2023; Cuskley et al. 2024, a.o.). Here we follow Wilcox et al. and use NLMs to study the kinds of generalizations that are in principle recoverable from the input via domain-general procedures, without making commitments about how human-like those learned representations are.

FGDs are contingencies between a filler, for example, a *wh*-word ‘what’ in (1-a) and a gap position (denoted with __ throughout the paper) later in the sentence where the filler is ultimately interpreted. The *wh*-question in (1-a) is an example of a filler-gap dependency where *what* is related to the gap that is a complement to the preposition *on*. Relative Clauses (RCs) like (1-b) are another example, where the head of the RC *the topic* is linked to a gap in the same position.

(1) a. What$_i$ did you write your first paper on __$_i$?
b. That’s the topic$_i$ that you wrote your first paper on __$_i$.

Acquiring the grammar of filler-gap dependency formation requires mastering a number of complex, abstract generalizations about the distribution of fillers and gaps. The most basic generalization is the *bidirectional* relationship between fillers and gaps. If a filler is not linked to a later gap, the sentence is ill-formed (2-a). Similarly, if a gap is not linked to a filler, the sentence is also ungrammatical (2-b).

(2) a.*What did you write your first paper on the topic?
b.*Did you write your first paper on __?

Learning the bidirectional contingency between fillers and gaps is not sufficient. There are additional generalizations that govern the configurations in which filler-gap dependencies are licensed, some of which vary by language. We review three such generalizations that are relevant to our paper.

First, FGDs are potentially *unbounded*: setting aside limitations imposed by working memory capacity, there is no limit on the linear or hierarchical distance between a filler and its corresponding gap. As the *wh*-FGD in (3) illustrates, one can interpolate multiple successively embedded clauses between the filler and the gap in both English (3-a) and Norwegian (3-b).

(3) a. Which topic$_i$ [did you say that [Marit thought [that Odd knew [that . . . you wrote your article about __$_i$?]]]]
b. Hvilket tema$_i$ [sa du at [Marit trodde [at Odd visste [at . . . du skrev
Which topic said you that Marit thought that Odd knew that . . . you wrote
artikkelen din om __$_i$?]]]]
article.DEF your about

Second, though (potentially) unbounded, FGDs are nevertheless constrained. Certain environments, referred to as *islands* (Ross, 1967), appear to block the association between fillers and gaps. Various structures have been identified as islands cross-linguistically. For example, subject phrases have been identified as islands in English and Norwegian alike. Therefore, attempting to link a gap inside a subject phrase to a filler outside the subject phrase leads to unacceptability of examples like (4), as confirmed by many formal judgment studies (Sprouse et al., 2012, 2016; Kush et al., 2018, 2019; Kobzeva et al., 2022b).

(4) a.*What$_i$ did [the letter about __$_i$] create problems?

b. *Hva$_i$ har [brevet om __$_i$] skapt problemer?
What has letter.DEF about created problems

Finally, though some environments appear to be islands across many languages, there is cross-linguistic variation when it comes to the islandhood of other environments. For example, embedded polar and adjunct questions are so-called *wh*-islands in English, as examples in (5) illustrate, but Norwegian appears to allow FGD-formation into these domains, as in (6) (Christensen, 1982; Kush et al., 2021; Kush and Dahl, 2020; Kush et al., 2023; Kobzeva et al., 2022b).

(5) a. ENGLISH EMBEDDED POLAR QUESTION (*whether*-island)
*That was the book$_i$ that I wondered [whether he had read __$_i$].

b. ENGLISH EMBEDDED ADJUNCT QUESTION (*wh*-island)
*Those are the students$_i$ that I don't know [where __$_i$ come from].

(6) a. NORWEGIAN EMBEDDED POLAR QUESTION
Det var boka$_i$ som jeg lurte på [om han hadde lest __$_i$].
That was book.DEF REL I wondered on whether he had read

b. NORWEGIAN EMBEDDED ADJUNCT QUESTION
Det er studentene$_i$ som jeg ikke vet [hvor __$_i$ kommer fra].
It is students.DEF REL I NEG know where come from

Many researchers acknowledge that learning the generalizations above presents a POS problem (Chomsky 1971; Phillips 2013b; Pearl 2022) because learners' input data are, in principle, compatible with multiple distinct hypotheses about the adult target state. To illustrate the problem: children may observe sentences in which one or two clauses — but never more — intervene between a filler and its gap (Hollebrandse and Roeper, 2014; Pearl and Sprouse, 2013b), which is consistent with unboundedness, but also with the more restrictive generalization that FGD-formation is bounded above two clauses. To arrive at the target generalization, children must generalize beyond their input to a class of unseen sentences. At the same time, they must also avoid overgeneralizing the possibility of FGD-formation to other unseen structural configurations if they are to capture island constraints. Human learners of the same language (and often across different languages) effectively strike this balance and converge on the same constrained generalizations. How?

Researchers in the generative tradition have assumed that innate language-specific biases guide filler-gap dependency acquisition. The POS problem that islands arguably pose, taken together with their abstract nature and (near) cross-linguistic uniformity in island facts, led to a search of possible unifying principles behind island acquisition (Phillips, 2013b). In particular, it has been proposed that knowledge of islands follows from innate constraints on what is a possible dependency, such as the Subjacency Condition (Chomsky, 1973) or Phases (Chomsky, 2001). For example, Chomsky's Subjacency condition postulated that a dependency cannot cross more than one bounding node (a certain phrase type intervening between the filler and the gap) in one application of a movement rule. For English, NP (DP) and IP (S or TP) were proposed to be bounding nodes, preventing movement out of embedded questions, which in turn renders examples like (5) above ungrammatical (Chomsky, 1973). To allow for some cross-linguistic variation, the set of bounding nodes may vary from language to language (Rizzi, 1982). In such traditional generative frameworks, island acquisition involves setting language-specific parameters in place (e.g., bounding nodes), while the set of parameters, their possible values, and abstract constraints on operations like movement are innately specified by Universal Grammar.

More recent attempts to model FGD acquisition while eschewing complex language-specific constraints have not eliminated domain-specific biases completely. Pearl and Sprouse (2013b) proposed a distributional learning algorithm that could successfully recover island constraints on English *wh*-dependencies from parsed child-directed speech, but only if it was biased to attend to select linguistic features of the input representations[2] (see also Pearl and Bates 2022; Dickson et al. 2022; Gulrajani and Lidz 2024).

Empiricist accounts predict that domain-general knowledge and learning mechanisms, such as pattern recognition and statistical learning, should be sufficient to induce the full set of generalizations on FGD formation from the input. Over the past few years, researchers have begun using NLM simulations to test these claims and have argued that NLMs can successfully recover abstract generalizations and distributional constraints, including that FGDs are potentially unbounded and subject to island constraints (Wilcox et al. 2018, 2019a,b, 2023). According to this line of reasoning, positive learnability results with NLMs provide empirical evidence against POS arguments in the domain of *wh*-movement.

Complicating the empirical picture, however, several recent studies revisiting Wilcox et al.'s work present empirical evidence that NLMs struggle when tested on more complex environments and might not in fact approximate the linguistic generalizations underlying filler-gap dependencies (Chaves, 2020; Da Costa and Chaves, 2020; Lan et al., 2024b; Howitt et al., 2024; Bhattacharya and van Schijndel, 2020). Moreover, NLMs' performance has been shown to vary depending on the type of FGD tested (Ozaki et al., 2022; Howitt et al., 2024), and what little cross-linguistic work has been done also suggests that success may vary across languages (Suijkerbuijk et al., 2023). As a general argument against domain-specific biases can only be made if the models are equally successful on a broad range of languages and dependencies, controlled cross-linguistic and cross-construction comparisons are especially informative.

To this end, this paper presents a controlled cross-linguistic comparison of FGD learnability in Norwegian and English. Our research questions pertain to the properties of filler-gap dependencies outlined above. We ask: (1) *Do the models learn that FGDs are structurally unbounded?* (2) *Do they induce island constraints on FGDs that*

[2] In particular, their modeled learner was trained on syntactically annotated child input and hard-wired to track the probability of trigrams of structural 'building blocks' that make up FGDs — phrase structure nodes such as IP, VP, and lexically annotated CPs.

Norwegian and English have in common? (3) *Do they learn patterns of cross-linguistic variation in island facts?* After conducting experiments that address these questions, we also conduct a restricted corpus analysis to better understand the type of input that the model uses to extract its generalizations in Norwegian.

To preview our results, we present mixed evidence regarding whether NLMs can learn the properties of filler-gap dependencies in both English and Norwegian. While the models successfully generalize in some cases (Experiment 2), they sometimes *undergeneralize*, adopting narrower generalizations than humans (Experiment 1). Additionally, while in some instances the models seem to capture the patterns of cross-linguistic variation correctly (Experiment 3), they also fail to do so in other instances, appearing to *overgeneralize* and predict that a subset of island violations are possible in English (Experiment 4). We conclude that although such models may acquire some sophisticated generalizations about filler-gap dependencies in the two languages, they do not successfully approximate the target human generalizations.

2. Method

2.1. Language Models

Language models take a sequence of words as input and compute a probability distribution over the model's vocabulary to predict the next word. In this paper, we evaluate the performance of two types of models, Long Short-Term Memory Recurrent Neural Networks (LSTM RNNs, Hochreiter and Schmidhuber (1997)) and a model based on the Transformer architecture (Vaswani et al., 2017)—specifically the GPT-2 variant (Radford et al., 2019)—on both Norwegian and English. The Norwegian LSTM and GPT-2 models were trained on Norwegian Wikipedia text (113 million tokens), while the English counterparts were trained on a subset of English Wikipedia (90 million tokens). The English LSTM was the most successful model reported in (Gulordava et al., 2018)[3], while the Norwegian LSTM was taken from Kobzeva et al. (2022a). The LSTM models were trained using the same procedure and architecture: both contained 2 layers with 650 hidden units in each layer and had a vocabulary consisting of the 50000 most frequent words in their respective corpora. Both LSTM models were trained for 40 epochs with a batch size of 128, a dropout rate of 0.2, and a learning rate of 20.0. The Norwegian LSTM achieved a perplexity of 30 on the validation set, whereas the English LSTM's perplexity was 52.[4] The GPT-2 models were based on GPT-2 small (117 M parameters) and were trained on the same data and had the same vocabulary size as the LSTM models. Here we report two models that achieved the lowest validation perplexities: for English, that model occurred during

[3] Downloaded from `https://github.com/facebookresearch/colorlessgreenRNNs/tree/main/data`

[4] Perplexities cannot be directly compared across languages and corpora due to the different corpora sizes, potential language-specific variation, corpus representativeness, differences in morphological complexity, etc.

epoch 9 (out of 54) and achieved a perplexity of 42, while for Norwegian, that model occurred during epoch 12 (out of 40) and achieved a perplexity of 27.

One important concern about the utility of language models for POS debates is that they are oftentimes trained on data amounts exceeding child input multiple times (Frank, 2023; Warstadt and Bowman, 2022). In our case, this concern is alleviated by the fact the input corpora sizes are relatively small. For English, 90 million words roughly correspond to the linguistic experience of a child between 8 (Hart and Risley, 1992) and 13 years of age (Gilkerson et al., 2017). For Norwegian, no such statistics exist, but given the typological proximity of the two languages, it is reasonable to assume that the estimates will be similar. To that end, the models do not have a considerable unfair advantage over humans in terms of data size (Warstadt and Bowman, 2022).

2.2. Dependent Measure

We assess how the models fare as incremental processors on sentences with filler-gap dependencies by looking at *surprisal*, an information-theoretic measure of how (un)predictable a word is given its context (Hale, 2001; Levy, 2008). Surprisal is defined as the negative logarithm of the conditional probability of a word given the previous context. In cognitive modeling, surprisal has been shown to be a strong predictor of processing difficulty as manifested by both behavioral measures like reading times and neural responses such as the amplitude of event-related brain potentials (Smith and Levy, 2013; Shain et al., 2024; Michaelov et al., 2024). In our models, surprisal values were calculated over the models' respective vocabularies calculated from their softmax layer.

2.3. Definition of Effects

To probe the models' generalizations about the distribution of filler-gap dependencies, we adopted the evaluation framework introduced by Wilcox et al. (2018). This evaluation involves a comparison between the surprisal values that models assign to target words in test sentences created according to a 2×2 factorial design which manipulates the presence of a filler and the presence of a gap as in (7).

(7) a. He knows that the student used AI on the exam. −FILLER, −GAP
b. *He knows what the student used AI on the exam. +FILLER, −GAP
c. *He knows that the student used __ on the exam. −FILLER, +GAP
d. He knows what the student used __ on the exam. +FILLER, +GAP

The design allows us to test the models' sensitivity to both parts of the *bidirectional* relationship between fillers and gaps by comparing minimal sentence pairs. We look for two different effects as a measure of a model's ability to construct a filler-gap dependency in a particular position: *unlicensed gap effects* and *filled-gap effects*.

Unlicensed gap effects quantify how the presence of an earlier filler influences the processing of a later gap. The unlicensed gap effect is intended to index if a models are sensitive to the fact that a gap depends on a previously-seen filler. Unlicensed gap effects are defined as the difference in surprisal at the region immediately following the gap in +GAP conditions (i.e., at *on the exam* in (7-c) v. (7-d)). If the model 'knows' that the gap in (7-d) is licensed, *on the exam* should be less surprising in that condition than in (7-c). Subtracting the latter from the former (i.e., subtracting −FILLER from +FILLER) should yield a *negative* difference. We consider unlicensed gap effects to be a direct window into model generalizations about possible gap positions.

Filled-gap effects quantify how the presence of an earlier filler influences the processing of a noun phrase in any *potential* gap position. The comparison is intended to measure whether the models' representations reflect the fact that a filler requires a later gap. The logic of the comparison rests on the assumption that having seen a filler should create an expectation for a gap in an upcoming position. Filled-gap effects are defined as the difference in surprisal at the potential gap site between −GAP conditions (i.e., at *AI* in (7-b) v. (7-a)). If the model 'knows' that seeing the filler *what* in (7-a) increases the likelihood of a gap after *used* compared to (7-b), then we should expect an increased surprisal value at *AI* in (7-a). Filled-gap effects have been observed in behavioral experiments investigating how humans resolve filler-gap dependencies during incremental processing (Crain and Fodor, 1985; Stowe, 1986). For example, Stowe (1986) found that participants took longer to read the direct object 'us' following a filler, *who*, in sentences like (8-b) compared to control sentences without a filler-gap dependency (8-a):

(8) a. My brother wanted to know if Ruth will bring us home to Mom at Christmas. −FGD
b. My brother wanted to know who Ruth will bring us home to __ at Christmas. +FGD

Stowe interpreted the slowdown as evidence that comprehenders actively predicted a gap in object position and experienced difficulty when the true direct object 'us' disconfirmed that prediction, potentially triggering reanalysis.

2.4. Diagnosing Sensitivity

We measure unlicensed gap and filled-gap effects across positions and environments to determine whether our models (i) can establish a relationship between an earlier filler and a gap in a given position and (ii) actively consider a gap as an option in that position. We note, however, an important asymmetry in the inferences that we are licensed to draw from the presence or absence of the two effect types: An unlicensed gap effect indicates that the model can establish an FGD with that position and the absence of an unlicensed gap effect entails that the model cannot establish an FGD. The same bidirectional reasoning does not apply to filled-gap effects. The implication holds only in one direction: A filled-gap effect signals an expectation for a gap, which entails that the model can establish an FGD, but the reverse does not hold: We cannot directly infer from the absence of a filled-gap effect in position X that a model cannot establish an FGD in position X. This inference would only be licensed if position X were the only grammatical gap site in the sentence, but it is often the case that other gap sites are possible later in the sentence (as seen in (8-b)). Given this, we consider unlicensed gap effects to be more reliable measures of the models' generalizations about FGDs.

Following previous work (Wilcox et al., 2018, 2023; Kobzeva et al., 2022a, 2023), we test the models' basic ability to establish grammatical filler-gap dependencies by looking for both effects in positions where gaps are licensed. We test for island sensitivity by asking whether unlicensed and filled-gap effects are suspended in island environments. Unlicensed gap effects should be suspended inside islands because the models should avoid associating gaps inside islands with fillers outside of the island domain. Filled-gap effects should be extinguished inside islands because the models should not expect to see gaps in unlicensed positions.

Human behavioral studies have shown that filled-gap effects are suspended inside islands. Stowe (1986) found that in contrast to (8), the participants did not actively pursue gaps inside subject phrases (subject islands) after encountering an upstream filler. In (9), there is a possible gap site inside the prepositional phrase attached to the subject *the story* which is 'filled' with a noun phrase *Greg's brother*. If humans were considering this slot as a potential gap site, then one would expect a filled-gap effect at *Greg's brother* in (9-b), where the filler *what* is present, as compared to (9-a) where it's not. Stowe found that there were no differences in reading times between the two conditions (9-b) and (9-a), suggesting that the language processor respects island constraints by suppressing active expectation for gaps inside island environments.

(9) a. The teacher asked if the story [about Greg's brother] was supposed to mean anything.
b. The teacher asked what the story [about Greg's brother] was supposed to mean.

Extending these diagnostics to potential island environments, if a learner has acquired the relevant restrictions on FGDs, we expect to find near-zero unlicensed gap effects and filled-gap effects inside embedded questions in English, but not in Norwegian. Consistent with this prediction, a recent behavioral study found filled-gap effects inside embedded *whether*-clauses in Norwegian, confirming the non-island status of this domain from a processing perspective (Kobzeva and Kush, 2024).

2.5. Statistical Analysis

We use two separate metrics to assess model performance. First, we assess whether there are significant differences across conditions in the *relative* size of filled-gap and unlicensed gap effects. Second, we ask whether the *absolute* size of any individual effect is different from zero. Statistical analysis of relative differences was performed using linear mixed-effects models with filler effects as the dependent variable. Filler effects were defined as the difference in surprisal values assigned to the critical region between +FILLER sentences and their −FILLER counterparts. We ran separate models for the two filler effects: one for filled-gap effects in the filled NP region (e.g., *AI* in (7-a) and (7-b)), and one for unlicensed gap effects in the region following the gap (e.g., *on the exam* in (7-c) and (7-d)). All statistical models had a fixed effect of CONDITION, which manipulated the location of the gap and varied across experiments. Because the number of conditions and contrasts varied across experiments, the contrast coding scheme for CONDITION differed between experiments and is therefore described in each experiment's subsection. Statistical models were fit in R (R Core Team, 2021) using the lme4 package (Bates et al., 2015). The models had the maximal random effect structure justified by the design Barr et al. (2013), which included by-item random intercepts and slopes for CONDITION. In cases where statistical models did not converge, only by-item random intercepts were included.

The relative comparison allows us to ask the following question: *Does the model assign lower probabilities to gaps inside islands than non-islands?*. However, even if the answer to that question is yes, we cannot necessarily conclude that the model cannot establish dependencies inside islands. Establishing that the model cannot represent gaps inside islands requires a more stringent criterion: filled-gap effects and unlicensed gap effects should be around zero. To assess whether the absolute size of any filler effect is different from zero, we checked whether the 95% confidence interval for that effect included zero, following (Wilcox et al., 2023).

3. Experiments

In this section, we present the results of four experiments investigating whether NLMs can learn FGDs and constraints on them in Norwegian and English. Alongside *wh*-dependencies tested by Wilcox et al. (2018, 2023), we included RC-dependencies into the test set to see whether the models make similar generalizations about different dependency types and how they are reflected in the input corpus data.

To create our Norwegian test items, the basic 2×2 design for *wh*-dependencies illustrated in (7) was translated into Norwegian, resulting in (10).

(10) a. −FILLER, −GAP, WH
Han vet at studenten brukte KI på prøven.
He knows that student.DEF used AI on exam.DEF

b. +FILLER, −GAP, WH
*Han vet *hva* studenten brukte KI på prøven.
He knows what student.DEF used AI on exam.DEF

c. −FILLER, +GAP, WH
*Han vet at studenten brukte __ på prøven.
He knows that student.DEF used on exam.DEF

d. +FILLER, +GAP, WH
Han vet *hva* studenten brukte __ på prøven.
He knows that student.DEF used on exam.DEF

In all of the experiments below, we created closely matched test sentences with RC-dependencies by modifying the corresponding *wh*-dependency sentences. (11) illustrates an adapted item set.

(11) a. −FILLER, −GAP, RC

Han fikk vite fra noen at studenten brukte KI på prøven.
He got know.INF from someone that student.DEF used AI on exam.DEF

b. +FILLER, −GAP, RC
*Han fikk vite om noe som studenten brukte KI på prøven.
He got know.INF about something that student.DEF used AI on exam.DEF

c. −FILLER, +GAP, RC
*Han fikk vite fra noen at studenten brukte __ på prøven.
He got know.INF from someone that student.DEF used on exam.DEF
‘He found out from someone that the student used __ on the exam.’ ‘He found out about something that the student used AI on the exam.’

d. +FILLER, +GAP, RC
Han fikk vite om noe som studenten brukte __ på prøven.
He got know.INF about something that student.DEF used on exam.DEF
‘He found out about something that the student used __ on the exam.’

To create the RC-dependency test items, we changed embedding verbs like *vet* ‘knows’ in (10) to verbs or predicates like the idiomatic *fikk vite* ‘got to know’ that had flexible subcategorization frames. The structure of the sentences after the main predicate differed depending on the levels of the FILLER factor. In −FILLER conditions, the predicate was followed by a prepositional phrase that introduced a source/goal argument (*fra noen*, ‘from someone’) and then a complement declarative clause (*at studenten . . .* , ‘that the student . . . ’) as in (11-a). In +FILLER conditions, the predicate was followed by a prepositional phrase headed by *om* ‘about’ that contained either the indefinite pronoun *noen* ‘someone’ or *noe* ‘something’. Relative clauses, headed by the relative pronoun *som* ‘that’, modified the indefinite NP, as in (11-b) and (11-d). This way, the main clause provided a licit filler for the upcoming gap in the relative clause, analogous to *wh*-words in *wh*-FGDs.

Translation equivalent items were created for *wh*-dependencies in English. Unfortunately, it was not possible to create comparable RC-dependency test sentences in English because the relative pronoun ‘that’ and the declarative complementizer ‘that’ are homonyms in the language. Compare (12-a) and (12-b) below, which are the translations of (11-b) and (11-d) above. In the two sentences, ‘that’ has different meanings: it either introduces an embedded declarative clause where there is no filler-gap dependency (12-a), or it serves as a relative pronoun inside a relative clause (+FGD case, (12-b)). Therefore, sentences without filler-gap dependencies but with an overt declarative complementizer could often be interpreted as containing relative clauses.

(12) a. He got to know from someone that the student used AI on the exam. −FGD
b. He got to know about something that the student used __ on the exam. +FGD

This makes it impossible to reliably distinguish between +FILLER, −FILLER conditions in English — a distinction on which this factorial design crucially relies. Therefore, we test *wh*-dependencies in both Norwegian and English, while RC-dependencies are only evaluated in Norwegian.

All four experiments used the factorial logic outlined above, with additional experiment-specific modifications described in the subsequent Materials subsections.

3.1. Experiment 1: Unboundedness

Experiment 1 tested whether the models learned the basic bidirectional relation between a filler and its gap and, if so, whether they could learn the generalization that the dependency between a filler and its gap can span an arbitrary *hierarchical* distance. To do so, we tested if the filled-gap effects and the unlicensed gap effects were observed when the filler and the gap were contained in the same clause, and if the effects persisted as the number of embedded clauses separating the filler and its gap increased.

3.1.1. Experiment 1: Materials

To create our items we crossed the basic design 2×2 in (10) with an additional factor, NUMBER OF LAYERS, that had five levels, yielding a 2×2×5 design. NUMBER OF LAYERS systematically manipulated the structural distance between the clause where the filler was introduced and the clause that could contain a gap. In the 1 LAYER condition, no clause intervened between the filler and the gap, and this condition tested whether the models could learn the simplest case of a filler-gap dependency with an object gap. In the 5 LAYERS condition, four nested clauses intervened. For reasons of space, we only illustrate the 1 and 5 LAYER conditions below. Layers of embedding are numbered in the examples:

(13) a. 1 LAYER (+FILLER, +GAP)
Han vet $[_1$hva studenten brukte __ på prøven.]
He knows what student.DEF used on exam.DEF
'He knows what the student used __ on the exam'.

b. 5 LAYERS (+FILLER, +GAP, WITH 'THAT')
Han vet $[_1$hva hun trodde $[_2$at foreldrene fant ut $[_3$at skolen
He knows what she thought that parents.DEF found out that school.DEF
mistenkte $[_4$at læreren visste $[_5$at studenten brukte __ på prøven.]]]]]
suspected that teacher.DEF knew that student.DEF used on exam.DEF
'He knows what she thought that the parents found out that the school suspected that the teacher knew that the student used __ on the exam'.

We also manipulated whether the intervening clauses were introduced by the declarative complementizer (*at* in Norwegian, *that* in English) or a zero complementizer. Wilcox et al. (2023) demonstrated that filler effects persisted across multiple clauses in English when the complementizer *that* was not present. Under the assumption that the presence or absence of the complementizer is orthogonal to the structure of the clause, the unboundedness generalization should not depend on such a low-level lexical factor. Thus, if the models have learned the correct generalization, their predictions should not be strongly influenced by the presence/absence of the complementizer. If, on the other hand, the models' behavior is significantly impacted by the presence of the complementizer, then that would suggest that the model is following a more restrictive generalization.

(13-b) provides an example item with overt complementizers, and (14) illustrates the corresponding condition without complementizers.[5]

(14) 5 LAYERS (+FILLER, +GAP, WITHOUT ‘THAT’)

Han vet [$_1$hva hun trodde [$_2$foreldrene fant ut [$_3$skolen mistenkte
He knows what she thought parents.DEF found out school.DEF suspected
[$_4$læreren visste [$_5$studenten brukte __ på prøven.]]]]]
teacher.DEF knew student.DEF used on exam.DEF

‘He knows what she thought the parents found out the school suspected the teacher knew the student used __ on the exam’.

50 lexically distinct, matched test items were created for *wh-* and RC-dependencies in Norwegian and *wh*-dependencies in English, yielding 1000 test sentences per dependency-language combination.

3.1.2. Experiment 1: Results and Discussion

The results of the Unboundedness experiment are presented in Figure 1. Here and in all remaining plots filler effects plotted on the y-axis represent the average difference between +FILLER, −FILLER conditions, which correspond to filled-gap effects (pink bars) and unlicensed gap effects (blue bars).

Filler effects are robust at 1 layer of embedding for all language-dependency pairs tested across both types of models, which establishes that the models can represent a local bidirectional relationship between fillers and gaps in object position. Thus, we replicate the English finding from Wilcox et al. (2018, 2023) and extend it to Norwegian *wh-* and RC-dependencies.

[5] In the 1 LAYER condition, there is no difference between with and without ‘that’ cases as no clause intervenes between the filler and the gap.

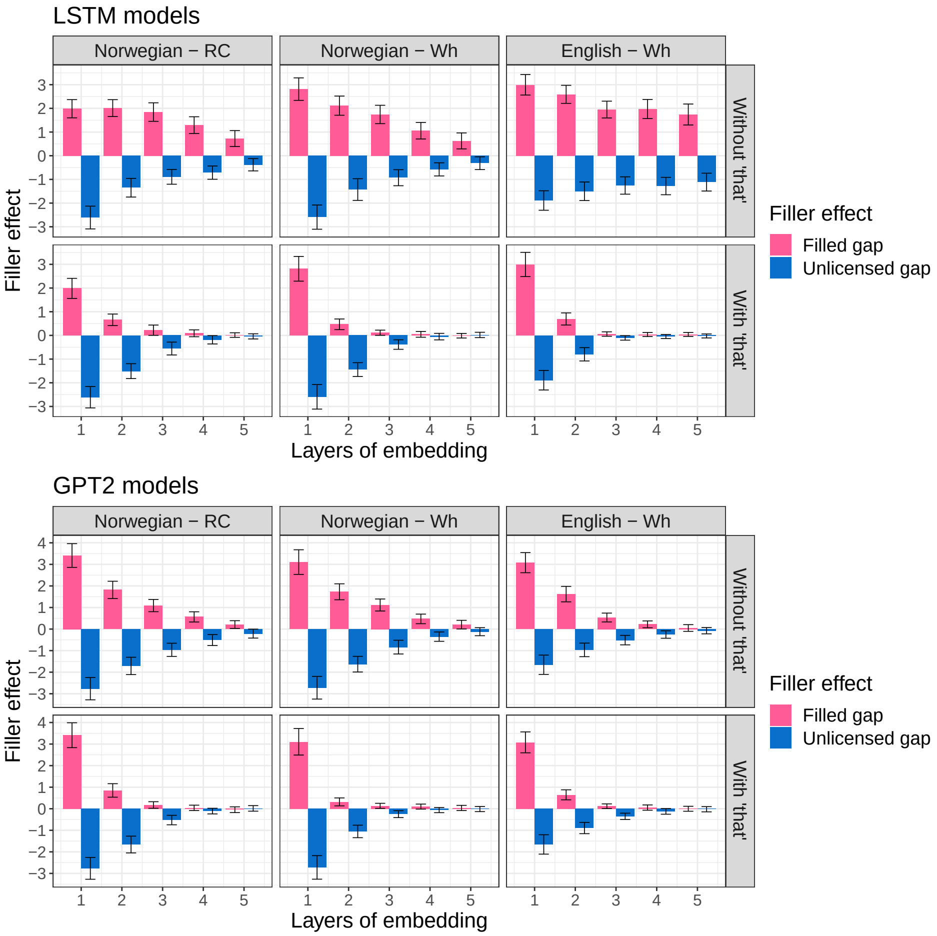


Figure 1: Results of Experiment 1 testing unboundedness of filler-gap dependencies. Error bars represent 95% confidence intervals.

In test sentences that *do not* contain an overt complementizer, filled and unlicensed gap effects are observed at deeper layers of embedding (upper rows of Figure 1). For both model types, effect sizes steadily diminish as layers of embedding increase, but the GPT-2 models exhibit a sharper reduction than the LSTM models with the effects trending towards zero at 4 and 5 layers of embedding. A general reduction in effect size as a function of embedding depth is in line with previous findings (Wilcox et al., 2023; Da Costa and Chaves, 2020).

To test how filler effects changed as a function of embedding, we fit linear mixed-effects regression models with filler effects as our response variable and NUMBER OF

LAYERS as our predictor variable. The predictor variable was backwards difference coded so as to compare the mean effect at one layer to the mean of the previous layer (2 v. 1, 3 v. 2 and so on). The output of all models can be found in Appendix A, Table A.6, but we summarize the main takeaways here. For sentences without a complementizer, filler effects remain comparable with the previous layer for both model types and languages up to 4 levels of embedding in the majority of statistical comparisons. Significant differences between 4 and 5 levels are observed in many statistical analyses. For sentences with complementizers, filler effects significantly decrease between 2 and 1 layers and continue to decrease significantly with every additional layer in nearly all comparisons (see Table A.6 for more details).

Insofar as the models exhibit non-zero filler effects across multiple layers of embedding when complementizers are absent, it appears that they can generalize that FGDs are unbounded in certain circumstances. However, the sharp decrease in filler effects with overt complementizers suggests that the models have come to a different, more restrictive generalization for FGDs in these sentences. How does this align with human generalizations? Under the assumption that complementizers are *optional* in the kinds of long-distance object questions that we tested, their presence should not have a marked effect on the ability to establish an FGD. Consistent with this assumption, Ritchart et al. (2016) found that an overt complementizer did not negatively impact humans' acceptability judgments of FGDs with two levels of embedding.[6] Assuming that humans exhibit the same insensitivity to an overt complementizer at greater depths of embedding, it would seem that the model fails to arrive at a human-like generalization (whether judgments or incremental behavior is the object of modeling).

Anecdotally, it seems that complementizers tend to be dropped in long-distance dependency production, which could mean that our training corpora lacked (sufficient) evidence of long-distance extraction across overt complementizers to generalize broadly. If that is the case, the models' behavior suggests that the models have extracted a generalization that hews more closely to observed distributions in the corpora.

One might ask how the models' poorer performance on deeply embedded FGDs with complementizers bears on their utility for testing island sensitivity, as islands are often nested clauses. We think that the results here prompt some caution, but we believe that the models' abilities are sufficient to proceed with island experiments. All of the test conditions in the coming experiments require FGDs across 2 layers of embedding at most, which the models are capable of representing.

[6] A recent eye-tracking study Chow and Zhou (2019) suggests that plausibility mismatch effects used to identify active gap-filling may be reduced in size when an extra layer of embedding is interpolated, potentially aligning with model predictions. We note that the length-dependent reduction in effect size was observed in the post-critical region, but plausibility mismatch effects in the critical region were not appreciably different across different lengths. As such, we do not think that there is strong evidence that the initial prediction of a gap dwindles with distance (though later interpretive processes associated with reanalysis may be affected by dependency length).

3.2. Experiment 2: Subject islands

Having shown that the models are sensitive to grammatical FGDs (with up to two layers of sentential embedding), we now proceed to test if they can limit this sensitivity in island environments where FGDs are ungrammatical. The second question we sought to answer was whether the models could recover the generalization that subject phrases are islands for filler-gap dependency formation in both Norwegian and English.

3.2.1. Experiment 2: Materials

Subject island effects arise when part of a subject phrase is extracted. To test for sensitivity to subject islands we created test items following the 2×2 design exemplified (15). Unlike in (10), the gap site in Experiment 2 was located inside a prepositional phrase, *i brevet* 'in the letter', attached to a subject NP *opplysningene* 'the information'.

(15) a. SUBJECT ISLAND (−FILLER, −GAP)
Hun oppdaget at [opplysningene i brevet] vil bekrefte mistanken
She discovered that information.DEF in letter.DEF will confirm suspicion.DEF
under rettssaken.
during trial.DEF
'She discovered that the information in the letter will confirm the suspicion during the trial.'

b. SUBJECT ISLAND (−FILLER, +GAP)
*Hun oppdaget at [opplysningene i __] vil bekrefte mistanken under
She discovered that information.DEF in __ will confirm suspicion.DEF during
rettssaken.
trial.DEF
'*She discovered that the information in __ will confirm the suspicion during the trial.'

c. SUBJECT ISLAND (+FILLER, -GAP)
*Hun oppdaget hva [opplysningene i brevet] vil bekrefte mistanken
She discovered what information.DEF in letter.DEF will confirm suspicion.DEF
under rettssaken.
during trial.DEF
'*She discovered what the information in the letter will confirm the suspicion during the trial.'

d. SUBJECT ISLAND (+FILLER, +GAP)
*Hun oppdaget hva [opplysningene i __] vil bekrefte mistanken under
She discovered what information.DEF in __ will confirm suspicion.DEF during
rettssaken.
trial.DEF
'*She discovered what the information in __ will confirm the suspicion during the trial.'

When the full phrase is extracted, embedded subject gaps are usually grammatical in Norwegian and English. To assess whether the models could link fillers to acceptable gaps in embedded subject positions, we included two control comparisons alongside the subject island condition. The first control condition (16-a) tested an embedded subject gap in the same clause as the filler. The second control condition (16-b) interpolated an embedded clause between the filler and the subject gap. Each control comparison followed the full 4-condition FILLER × GAP design, though we only present the +FILLER, +GAP condition to illustrate.[7]

(16) a. SUBJECT CONTROL (+FILLER, +GAP)
Hun oppdaget hva som __ vil bekrefte mistanken under rettssaken.
She discovered what C will confirm suspicion.DEF during trial.DEF
'She discovered what __ will confirm the suspicion during the trial.'

b. EMBEDDED CONTROL (+FILLER, +GAP)
Hun oppdaget hva han trodde __ vil bekrefte mistanken under rettssaken.
She discovered what he believed will confirm suspicion.DEF during trial.DEF
'She discovered what he believed __ will confirm the suspicion during the trial.'

We expect to see both filled-gap and unlicensed gap effects in the SUBJECT CONTROL and EMBEDDED CONTROL comparisons. If the models have learned that subjects are islands, though, we should expect no filled-gap effects at *brevet* 'the letter' in (15-a) v. (15-c) and no unlicensed gap effect at *vil* 'will' in (15-b) v. (15-d).

3.2.2. Experiment 2: Results and Discussion

A breakdown of filler effects by condition and dependency is presented in Figure 2. We see a similar pattern of results across the models we tested: both filled-gap and unlicensed gap effects were at or near zero in the subject island condition across all language and dependency pairs, suggesting that the models do not represent illicit FGDs into subject phrases. The absence of filler effects in the subject island condition contrasts with the non-zero filler effects in the two control comparisons. The large filled-gap and unlicensed gap effects in the subject control condition indicate that the models are capable of extracting subjects across short distances. Similar effects in the embedded control condition show that the models can still extract more deeply embedded subjects, suggesting that the absence of the effects in the subject island condition does not simply reflect difficulty with embedding alone.

[7] In (16-a) the local subject is extracted. When the local subject is extracted in a Norwegian embedded question, the complementizer *som* must follow the *wh*-filler. The complementizer is not observed in embedded questions where the *wh*-filler is linked to a gap in any other position. The presence of *som* in (16-a) therefore serves as a diagnostic cue for a local subject gap. *Som* is also used as a relative pronoun in RC-dependencies, but its presence does not entail a local subject gap. In this regard, therefore, we expect stronger expectations for a subject gap — and therefore stronger effects — with *wh*-dependencies in the SUBJECT CONTROL conditions than with other dependencies and other conditions.

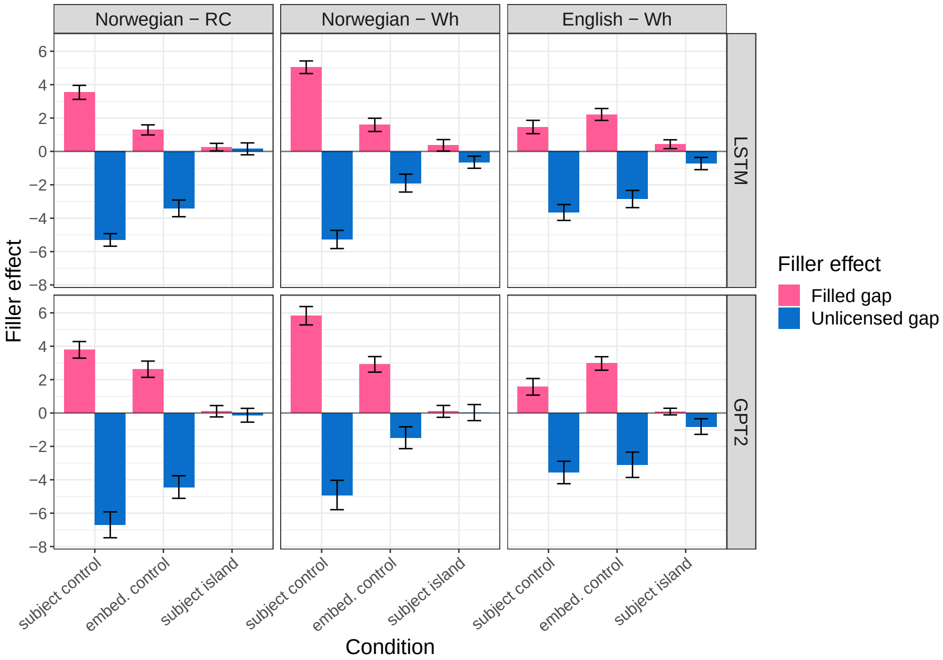


Figure 2: Results of Experiment 2 testing subject islands. Error bars represent 95% confidence intervals.

We used forward difference coding to define contrasts for statistical analysis, which compares the mean filler effect at one level of CONDITION to the mean filler effect of the next adjacent level. With three levels of CONDITION, this resulted in two contrasts: the *control contrast* compared the mean of filler effects between the two control conditions (SUBJECT CONTROL V. EMBEDDED CONTROL). The *island contrast* compared the filler effects between the EMBEDDED SUBJECT and the SUBJECT ISLAND conditions. We chose this control condition as the baseline for this contrast because it is more comparable to the island condition in terms of structural depth.

Statistical analysis revealed that for both the LSTM and GPT-2 models, control contrasts were significant in Norwegian suggesting that the additional level of embedding has a non-negligible impact, in line with the Unboundedness experiment (see Table 1). In English, control contrasts were only significant with UGEs but not FGEs. All island contrasts were significant reflecting reduced effects in subject islands compared to the embedded subject control across languages and models. However, non-zero filler effects are still present in half of the comparisons: 95% confidence intervals do not cross zero with either unlicensed gap effect in English, and the Norwegian LSTM shows non-zero effects in all cases except the unlicensed gap effect for RC dependencies.

	Norwegian - RC		Norwegian - Wh		English - Wh	
	FGE	UGE	FGE	UGE	FGE	UGE
			LSTM			
control contrast	3.7***	-4.9***	5.4***	-5.3***	0.2	-2.5***
island contrast	2.9***	-6.0***	3.9***	-3.9***	1.9***	-3.4***
			GPT-2			
control contrast	3.2***	-5.9***	5.8***	-5.6***	0.06	-2.1***
island contrast	4.1***	-7.2***	5.7***	-4.3***	2.9***	-3.4***

Table 1: Output of the linear mixed-effects models for Experiment 2 that tested subject islands. Control contrast compared the two control conditions to one another, while island contrast compared filler effects in the EMBEDDED CONTROL condition to the SUBJECT ISLAND condition. Reported values are model coefficients and diacritics represent significance levels (***p <.001).

3.3. Experiment 3: Embedded Polar Questions/'Whether-islands'

Having investigated sensitivity to an island constraint that is shared between Norwegian and English, we investigated a point of divergence between them: embedded polar questions or '*whether*-islands'. As discussed above, prior studies have found that native speakers of Norwegian produce FGDs into embedded polar questions and often judge them as acceptable (Kush et al., 2018, 2019; Kobzeva et al., 2022b; Kush et al., 2021), whereas English speakers consistently exhibit island effects when judging such constructions (Sprouse et al., 2012, 2016; Pañeda et al., 2024). Recent work has suggested that NLMs trained on English corpora exhibit *whether*-island sensitivity (Wilcox et al., 2018, 2023). We tested whether NLMs trained on Norwegian data would arrive at a different conclusion.

3.3.1. Experiment 3: Materials

We created the experimental stimuli according to a 2 × 2 design that crossed the factors FILLER and GAP, as before. The test gap was located in the object position in an embedded clause. We crossed the basic design with a third factor, CLAUSE, which manipulated whether the embedded clause was introduced by a declarative complementizer (DECL-COMP, control condition) or by a complementizer *whether* (WHETHER-COMP, potential *whether*-island). Examples of the +FILLER, +GAP condition with both declarative control and *whether*-embedded clauses are given below.

(17) a. DECL-COMP (+FILLER, +GAP)

Han vet hva professoren kunne fortelle at studenten brukte __ på
He knows what professor.DEF could tell that student.DEF used __ on
prøven.
exam.DEF

'He knows what the professor could tell that the student used __ on the exam'.

b. WHETHER-COMP (+FILLER, +GAP)

Han vet hva professoren kunne fortelle om studenten brukte __ på
He knows what professor.DEF could tell whether student.DEF used __ on
prøven.
exam.DEF
‘He knows what the professor could tell whether the student used __ on the exam’.

We created 50 items following the 2 × 2 × 2 design for *wh*- and RC-dependencies in Norwegian and translation equivalent *wh*-dependencies in English, resulting in 400 sentences per language-dependency combination.

3.3.2. Experiment 3: Results and Discussion

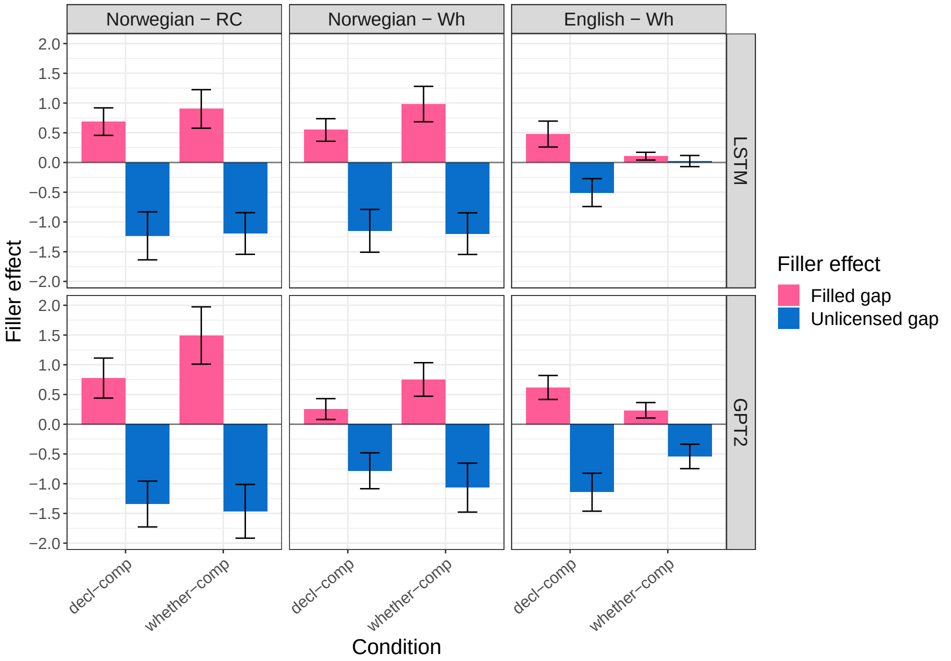


Figure 3: Results from Experiment 3 testing object extraction from embedded polar questions/‘whether-islands’. Error bars represent 95% confidence intervals.

Average filler effects across comparisons in Experiment 3 are presented in Figure 3. There are clear differences between the models’ predictions between the two languages: the Norwegian models exhibit filler effects in *whether*-clauses that are comparable to or even larger than filler effects in embedded declaratives. This holds for both *wh*- and RC-dependencies (left and middle panels in Figure 3). On the other hand, in English, filler effects are significantly reduced inside embedded *whether*-questions compared to the embedded declaratives. There are some notable model differences in English: First, effect sizes are, on average, smaller for both comparisons in the LSTM model than the GPT-2 model. Second, the GPT-2 model shows non-zero filled-gap and unlicensed gap effects inside whether-clauses. Despite this, the overall reduction in effect size seems comparable between the two models given the baseline differences in effect size.

For statistical analysis, we used sum-coded fixed effects of CONDITION (0.5 for DECL-COMP and -0.5 for WHETHER-COMP). The results of the statistical analysis are presented in Table 2.

In Norwegian, effect sizes in *whether*-clauses were not smaller than in embedded declaratives for either dependency or model type. The few significant differences observed reflect *larger* filler effects in *whether*-clauses.[8] In English, filler effects were significantly smaller in the island condition for both models.

To supplement the within-language comparisons, we also conducted a between-language comparison of Norwegian and English *wh*-dependencies using a model with sum-coded effects of LANGUAGE (-0.5 for English, 0.5 for Norwegian), CONDITION (0.5 for DECL-COMP and -0.5 for WHETHER-COMP) and their interaction. A main effect of LANGUAGE (both ps<.001) indicated that mean filler effects were smaller in English than in Norwegian. Most importantly, we found a significant LANGUAGE×CONDITION interaction (both ps<.001) that reflected that filler effects were reduced in English *whether*-clauses, but not in Norwegian.

Taken together, these results suggest that while the English models are sensitive to the *whether*-island constraint, the Norwegian models treat dependencies into whether-clauses on par with or even more probable than dependencies into embedded declaratives.

	Norwegian - RC		Norwegian - Wh		English - Wh	
	FGE	UGE	FGE	UGE	FGE	UGE
				LSTM		
condition	-0.2	-0.04	-0.4***	0.05	0.4***	-0.5***
				GPT-2		
condition	-0.7***	0.1	-0.5***	0.3*	0.4***	-0.6***

Table 2: Output of the linear mixed-effects models for Experiment 3 testing object extraction from embedded polar questions/‘whether-islands’. Reported values are model coefficients and diacritics represent significance levels (*p<.05; ***p<.001).

[8] In line with the present findings, Kobzeva and Kush (2025) found that RC-dependencies into embedded whether-clauses were more frequent than RC-dependencies into embedded declaratives in the corpus of child-directed text that they studied.

3.4. Experiment 4: Embedded Adjunct Questions/ 'Wh-islands'

Experiment 3 tested whether the models would establish FGDs into embedded polar questions. Consistent with human judgment patterns, we found that the Norwegian models established *wh*- and RC-dependencies into embedded polar questions, but the English models did not. In Experiment 4 we probe the generality and the robustness of the models' ability to recover the cross-linguistic difference in the island status of embedded questions by testing a different construction. As discussed above, Norwegian differs from English in that it allows filler-gap dependencies into embedded questions introduced by other interrogative question words like *hvem* 'who', *hva* 'what', *hvordan* 'how', *hvor* 'where', etc. Moreover, alongside object gaps tested in Experiment 3, Norwegian also allows subject gaps in embedded questions, see (6-b), repeated in (18).

(18) Det er studentene$_i$ som jeg ikke vet [hvor $\_\_{i}$ kommer fra].
It is students.DEF REL I NEG know where come from
lit. 'Those are the students$_i$ that I don't know [where $\_\_{i}$ come from].'

To match human judgments, the Norwegian models should learn that such dependencies are possible. The English counterparts of sentences like (18) are judged unacceptable and not produced by native speakers (Morgan, 2022; Kush and Dahl, 2020; McDaniel et al., 2015; Kush et al., 2023). They are considered ungrammatical because they violate at least two constraints: (i) the prohibition on FGDs into embedded questions (*wh*-islands) and (ii) the prohibition on having a gap immediately adjacent to an overt phrase in the complementizer domain (a so-called Comp/that-trace configuration, see Perlmutter 1971; Chomsky and Lasnik 1977; Morgan 2022, a.o.). As such, a successful English model should not allow fillers to be related to subject gaps in the sentences.

3.4.1. Experiment 4: Materials

We created 50 experimental items by crossing the basic factors FILLER and GAP, where the critical gap was located in an embedded subject position. We crossed the 2×2 design with a third factor, CLAUSE, which varied properties of the embedded clause. CLAUSE had three levels: ZERO-COMP, in which the embedded clause was a declarative with a zero complementizer (i.e., no complementizer), DECL-COMP, in which the embedded clause was headed by the declarative complementizer *at* 'that' in Norwegian, and WH-COMP, where the embedded clause was an embedded adjunct question. Embedded questions were introduced by four different interrogative question words: *hvor* 'where', *når* 'when', *hvordan* 'how' and *hvorfor* 'why'. The different clause types are exemplified in (19).

(19) a. ZERO-COMP (+FILLER, +GAP)
Han fant ut hva$_i$ de bekreftet $\_\_{i}$ er planlagt til neste uke.
He found out what they confirmed is scheduled for next week

b. DECL-COMP (+FILLER, +GAP)

Han fant ut hva$_i$ de bekreftet at __$_i$ er planlagt til neste uke.
He found out what they confirmed that is scheduled for next week

c. WH-COMP (+FILLER, +GAP)
Han fant ut hva$_i$ de bekreftet når$_k$ __$_i$ er planlagt __$_k$.
He found out what they confirmed when is scheduled

We chose to include both the zero complementizer (19-a) and declarative complementizer (19-b) comparisons as controls to determine what effect, if any, having an overt complementizer immediately before the gap would have on the model's behavior (following the results of a similar manipulation in Experiment 1).

When creating the English items we dropped the condition containing an overt complementizer, as including the complementizer would have created a that-trace configuration, which is unacceptable in English (Perlmutter, 1971; Chomsky and Lasnik, 1977; Sobin, 1987). To minimize the effect of other potential sources of ungrammaticality on our conclusions, CLAUSE only had two levels in the English sub-experiment: ZERO-COMP and WH-COMP.

(20) a. ZERO-COMP (+FILLER, +GAP)
He found out what they confirmed __ is scheduled for next week.
b. WH-COMP (+FILLER, +GAP)
*He found out what they confirmed when __ is scheduled __.

The 50 lexically distinct test items were adapted to all language-dependency test pairs.

3.4.2. Experiment 4: Results and Discussion

The results of Experiment 4 are presented in Figure 4. Beginning with unlicensed gap effects in Norwegian, we find that the LSTM exhibits comparable effects across all three conditions with both wh- and RC-dependencies. The Norwegian GPT-2 shows large unlicensed gap effects in both control conditions, but reduced effect sizes in the wh-complementizer condition. Nevertheless, the unlicensed gap effect is still different from zero. Norwegian filled-gap effects are relatively large in the zero-complementizer condition for wh- and RC-dependencies, drastically reduced in the declarative complementizer condition and near zero in the wh-complementizer condition.

Turning to English, we see an identical pattern across LSTM and GPT-2 models: unlicensed gap effects are large with a zero complementizer and much smaller inside the embedded question. Importantly, the unlicensed gap effects are not zero in the embedded question. In fact, they are comparable in size to the unlicensed gap effects in the declarative control conditions from Experiment 3, which were taken as evidence that the model could establish filler-gap dependency. Finally, filled-gap effects are large in the zero-complementizer condition, but negligible inside the embedded question.

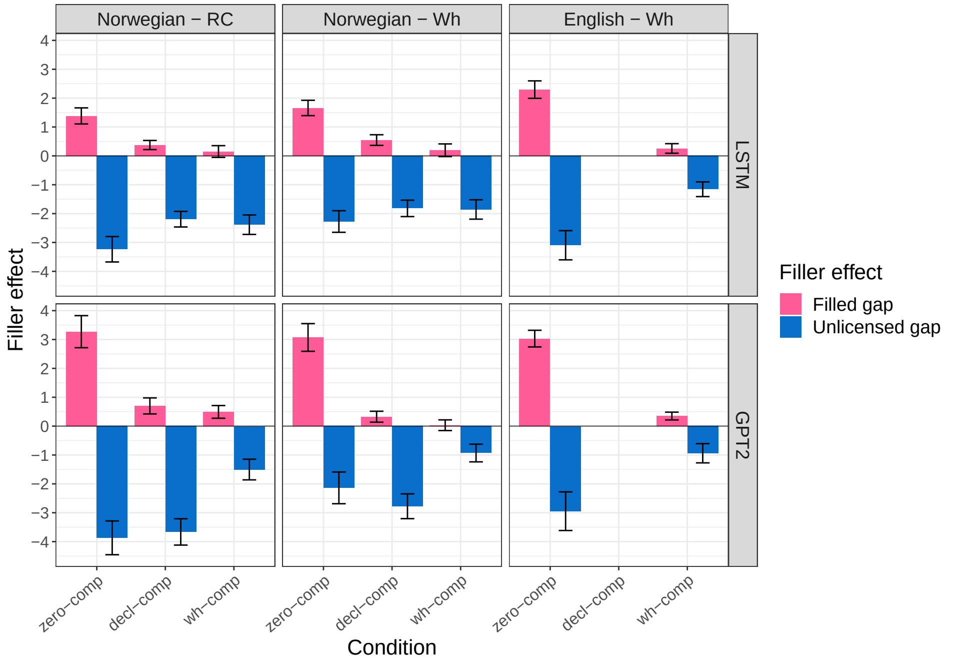


Figure 4: Results of Experiment 4 testing subject extraction from embedded adjunct questions/'wh-islands'. Error bars represent 95% confidence intervals.

A series of linear mixed effects models were used to compare the size of the effects across conditions. The output of the models is summarized in Table 3. Norwegian models employed forward difference-coded fixed effect of CONDITION to make two comparisons. The *declarative contrast* compared the mean filler effects in ZERO-COMP to DECL-COMP. The *'island' contrast* compared DECL-COMP to WH-COMP, using the former as a baseline. The English models had a fixed effects of CONDITION (0.5 for ZERO-COMP and -0.5 for WH-COMP), reported in the same line as the 'island' contrast in Norwegian.

	Norwegian - RC		Norwegian - Wh		English - Wh	
	FGE	UGE	FGE	UGE	FGE	UGE
	LSTM					
declarative contrast	1.5***	-1.3***	1.7***	-0.6**		
'island' contrast	1.0***	-0.4*	1.2***	-0.3	2.0***	-1.9***
	GPT-2					
declarative contrast	3.6***	-1.7***	3.9***	-0.4		

‘island’ contrast	2.0***	-3.0***	2.2***	-2.0***	2.7***	-2.0***

Table 3: Output of the linear mixed-effects models for Experiment 4, which tested subject extraction from embedded adjunct questions/‘wh-islands’. Declarative contrast compared the two declarative conditions in Norwegian only. Island contrast compared filler effects between decl-comp and wh-comp in Norwegian and zero-comp and wh-comp in English. Reported values are model coefficients and diacritics represent significance levels (*p <.05, **p <.01, ***p <.001).

Confirming our qualitative observations, statistical analysis revealed that filled-gap effects were significantly reduced in the wh-comp condition compared to decl-comp across all language-dependency combinations (all ps<.001). As for the unlicensed gap effects, the results are mixed. Starting with the Norwegian LSTM, unlicensed gap effects in wh-comp are similar to or slightly larger than the effects in decl-comp, in line with the results from Experiment 3. Filled-gap effects, on the other hand, are smaller in wh-comp than in decl-comp for both dependencies. For the Norwegian GPT-2 model, filler effects were consistently smaller in wh-comp than in decl-comp (ps<.001). A similar pattern was observed in English, where, irrespective of the model, the unlicensed gap effects were significantly reduced in the wh-comp condition compared to the zero-comp control condition.

As in Experiment 3, we also conducted a between-language comparison of filler effects with Norwegian and English *wh*-dependencies. Statistical models included sum-coded effects of language (-0.5 for English, 0.5 for Norwegian), condition (0.5 for zero-comp and -0.5 for wh-comp) and their interaction. For filled-gap effects, we observed main effects of condition (p<.001), reflecting reduced effects inside embedded questions, and language (p<.05), reflecting slightly larger effects in English. The language×condition interaction was not significant, indicating comparable patterns of reduction in English and Norwegian. As for unlicensed gap effects, the analysis revealed a significant main effect of condition qualified by a significant language×condition interaction (both ps<.001), which reflected that the unlicensed gap effects were significantly more reduced in English embedded questions than in Norwegian. The main effect of language was not significant.

The asymmetry between filled-gap effects and unlicensed gap effects observed here provides valuable insights. In Norwegian, filled-gap effects decrease in size across conditions, indicating that the model assigns lower probability to a gap as the intervening lexical material increases in complexity (from zero to a declarative complementizer to a *wh*-word). Expectation for a gap becomes ‘less active’, ultimately extinguishing in the embedded question.

In contrast, unlicensed gap effects are more robust across conditions, suggesting that even if active expectation for a gap is extinguished, the models still ‘recognize’ that it is possible to link a filler to a gap in all three environments.

4. Corpus Analysis

Experiments 3 and 4 above indicate that the NLMs can represent wh- and RC-FGDs into embedded questions in Norwegian. We sought to identify whether the models received direct evidence of such dependencies, and if so, how much direct evidence, in order to better understand how the model generalized.

4.1. Method

We parsed the Norwegian Wikipedia corpus that our models were trained on using the dependency parsing module in Stanza (Qi et al., 2020). After parsing, we queried the corpus for sentences containing a verb that could introduce an embedded question (e.g., *lure på* 'wonder') and a *wh*-word that depended on that verb.[9] This search resulted in 42482 candidate sentences. The first and the last authors of the paper manually checked 6400 (∼15%) of the sentences to identify any examples of the relevant dependencies into embedded questions. Among the sentences that were checked, we discovered 756 (∼12%) false positives, i.e. sentences that did not contain embedded questions, which can be attributed to misparses and the non-restrictive nature of the search queries used.

After discarding false positives, we first manually categorized the remaining embedded questions by the grammatical function of the *wh*-word introducing the question. Counts by type can be found in Table 4. Adjunct embedded questions introduced by *wh*-words like *hvor* 'where' and *hvordan* 'how' were by far the most common type of EQs, followed by polar embedded questions introduced by *om* 'whether'. Embedded subject and copular predicate questions were the next most common types. Together, these four question types constitute approximately 90% of all embedded questions in the sample.

EQ type	Count	Percentage
Adjunct	2843	50.0
Polar	1099	19.3
Subject	792	13.9
Copular predicate	406	7.1
Object	391	6.9
Oblique	157	2.7
Total:	5688	100

Table 4: Descriptive statistics for different types of EQs in the Norwegian Wikipedia corpus.

[9] We looked for the following dependency relations `deprel` between the verb that could potentially introduce an EQ and a *wh*-word: clausal complement `ccomp`, open clausal complement `xcomp`, adverbial clause modifier `advcl`, and oblique `obl`. The search was non-restrictive: including `obl` into the list of relations led to the majority of false positives with prepositional phrases instead of embedded questions.

From 5688 sentences summarized above, we found 33 (0.6%) sentences that contain FGDs into embedded questions. All 33 sentences can be found in Appendix B. 30 of the dependencies were RC-dependencies. In line with findings from Kush et al. (2021), we found no examples of *wh*-dependencies into embedded questions. The remaining 3 examples of filler-gap dependencies were examples of long-distance topicalization, which is very prominent in Norwegian.

Table 5 summarizes the distribution of gaps by embedded question type.

Embedded *wh*-word	Gap position	Count
om 'whether'	subject	7
om 'whether'	object	3
hvor 'where'	subject	6
når 'when'	subject	1
hvordan 'how'	object	1
hvem 'who'	subject	6
hva 'what'	subject	9
Total:		33

Table 5: Summary of dependencies into different types of embedded questions from the manually checked portion of the Norwegian Wikipedia corpus.

When it comes to the prevalence of sentences with the specific structural configurations that we tested in Experiments 3 and 4, we find relatively few examples. Relevant to the results of Experiment 3, we find only two sentences in which an RC-filler is linked to an object gap in an embedded polar question. An example is below:

(21) På banen overrasker Luck motspillere med kommentarer som man ikke kan være
On field.DEF surprises Luck opponents with comments REF man NEG can be
sikker på om __ er frekt ment ...
sure whether are rudely meant ...
'On the field, Luck surprises the opposing players with comments$_i$ that one cannot be sure whether __$_i$ are rudely meant ...'[10]

Relevant to the results of Experiment 4, the sample contains 7 sentences with a subject gap inside an adjunct embedded question headed by *where* or *when* as in (22), but no examples with *why* or *how.* If we loosen the criteria to include sentences with subject gaps immediately following *any* wh-word, there are 28 potentially relevant sentences.

(22) Alt dette var en del av [tradisjonell kunnskap]$_i$ en vanskelig kan si [når __$_i$
All this was a part of traditional knowledge one difficult can say when

[10] Source: Andrew Luck Wikipedia page

oppsto.]
arose.
lit. ‘All of this was part of traditional knowledge that one can hardly say when __ arose.’[11]

4.2. Discussion

The sample suggests that direct evidence for the exact structures we tested — or near neighbor structures — is present, but not abundant, in the training corpus. The relative scarcity of the specific constructions and the divergence between our test items and the attested examples in terms of lexical content suggests that the Norwegian models have not just learned specific dependencies by rote.

The distribution of examples suggests a degree of cross-dependency generalization: Despite the conspicuous absence of *wh*-dependencies into any embedded questions, we nevertheless observed filled-gap effects and unlicensed gap effects for *wh*-dependencies into such constituents in Experiments 3 and 4. We speculate that the models may generalize from the distribution of gaps in RC-dependencies to possible gap positions for *wh*-fillers. The near uniformity in effect sizes between *wh*- and RC-dependencies in both experiments supports this claim. We also speculate that evidence could also be taken from another dependency type that we did not test: topicalization, which is well attested in naturalistic examples of dependencies into embedded questions.

The results also suggest a degree of cross-construction generalization. Although we observe relatively few examples of dependencies into embedded polar questions, for example, the Norwegian models assign roughly equal probability to object gaps in embedded polar questions and embedded declarative clauses. The parity of the effects suggests that the models treat both embedded clause types as ‘the same’ in some sense for the purposes of FGD formation.[12]

5. General Discussion

We investigated whether LSTM and Transformer models trained on Norwegian and English Wikipedia texts can recover generalizations about the broader distribution of filler-gap dependencies in English and Norwegian. We tested whether the models could learn that (i) FGD-formation is potentially unbounded in both languages, (ii) that subject phrases are islands for filler-gap dependency formation in both languages, and (iii) that embedded questions are islands in English, but not Norwegian. We assessed whether the models could establish FGDs by measuring whether they exhibited filled-gap effects and unlicensed gap effects in different positions, on the assumption that models should exhibit both kinds of effects in environments that allowed filler-gap dependency formation. Successful learning of island constraints would mean that both effect types would be extinguished in island environments. Our results suggest that

[11] Source: ‘Strikking [Knitting]’ Wikipedia page

[12] A similar claim could be made based on results from Experiment 4.

the models successfully approximate some of the target generalizations across dependency types and languages, particularly when their performance is evaluated against a relative metric, which simply asked whether the models assigned significantly lower probability to gaps in island environments than in non-island environments. However, according to a more stringent absolute metric, the models succeed only around half of the time. Despite the qualified successes, there were a few important areas in which the models' behavior was arguably not target-like. Below we consider what the models' successes and struggles tell us about the types of generalizations that they induce and the implications of our findings for debates surrounding the learnability of filler-gap dependencies and islands by statistical learners without language-specific biases (Wilcox et al., 2023; Lan et al., 2024b).

5.1. Successful Approximation of Target Generalizations

First, the models appear capable of relating fillers to gaps across multiple levels of hierarchical embedding under certain conditions (e.g., without declarative complementizers), partially aligning with the generalization that FGDs are unbounded. Second, the models exhibited filled-gap and unlicensed gap effects that were either at zero or very close to zero inside subjects, approximating the generalization that subject phrases are islands. Third, when trained on different languages, the models assigned different probability to dependencies crossing into embedded polar questions. The Norwegian models exhibited robust filled-gap and unlicensed gap effects in both declarative complement clauses and embedded polar questions. This was true for both *wh-* and RC-dependencies. In contrast, the English models showed reduced or near-zero filler effects inside embedded *whether*-questions.

5.1.1. (Some) Cross-linguistic Variation is Learnable

Our findings show that NLMs can recover patterns of cross-linguistic variation in the island status of embedded polar questions. One possible explanation for how the Norwegian models learned that embedded questions are not islands is via direct evidence. Our corpus analysis revealed that the Norwegian training data indeed contained a small number of examples of RC-dependencies into *whether*-clauses, which we conjecture the models were able to leverage to learn that dependencies into embedded questions should be treated equivalently to dependencies into embedded declaratives. The importance of such direct positive evidence for learning infrequent FGDs has recently been demonstrated by Lan et al. (2024b), who found that NLMs' performance on double-gap phenomena (parasitic gaps and across-the-board extraction) improved significantly after the training corpus has been augmented with examples of relevant constructions. Kobzeva and Kush also concluded that the non-island status of embedded polar questions could be learned from direct evidence (around 20 relevant examples) when evaluating a more traditional symbolic cognitive model in Norwegian. Their computational learner received as input structured representations from a corpus of child-directed text (28 times smaller than the Norwegian Wikipedia corpus) and was trained to estimate the probability of FGDs based on frequencies of n-grams of their constituent 'building blocks' (phrase structure nodes such as IP, VP and lexically annotated CPs). Taken together, the findings highlight a likely trade-off between learner's representational biases and the power of the learning mechanisms that are needed to arrive at the target state. While NLMs, which are powerful domain-general learners without in-built language biases, could induce the non-island status of dependencies into polar embedded questions from exposure to text, a symbolic model with very simple learning mechanisms could reach the same conclusion when supplemented with very strong representational biases for hierarchical structure of language.

One important question to ask is how such positive results add to the POS debates surrounding islands. It would appear that the input may be rich enough to support the learning of the relevant generalization through direct positive evidence. This is a welcome conclusion for both parameter-setting generativist accounts and empiricist accounts, since both camps predict that the patterns of cross-linguistic variation should be recoverable from the input. The accounts differ in how this input maps onto the developing linguistic representations — be they innately pre-defined or shaped by domain-general learning procedures. Although the positive results presented here are important, they alone do not provide empirical support for or against either account.

5.1.2. Cross-dependency Generalization?

Relevant to arguments from the POS, there is evidence that the models appropriately extrapolated beyond the fine-grained statistics of the input to approximate the broader generalization that Norwegian embedded questions are not islands for different types of FGD. The primary evidence for some degree of abstract generalization is that the models showed filled-gap effects and unlicensed gap effects with *wh*-dependencies into embedded questions, even though we found no examples of such dependencies in our corpus. We hypothesize that the models inferred that such *wh*-dependencies are licensed via indirect evidence, using examples of RC-dependencies into embedded questions (and perhaps other dependencies like topicalization). The idea that NLMs can utilize indirect evidence found in the input is supported by recent work (Patil et al., 2024; Misra and Mahowald, 2024; Potts, 2023; Leong and Linzen, 2024), and such cross-dependency generalization is consistent with a kind of shared underlying representation that treats the two FGDs as an equivalence class. This conclusion is in line with previous work that suggests that NLMs induce abstract representations (Gulordava et al., 2018; Hu et al., 2020; Linzen and Baroni, 2021), that might track linguistically interpretable classes of constructions (Prasad et al., 2019).

Our conclusion that the models can generalize across FGD types differs from those of Howitt et al. (2024), who investigated if an LSTM developed a shared representation for four types of FGDs typically analyzed as movement dependencies: *wh*-dependencies, clefts, topicalizations, and *tough*-movement. The authors tested whether augmenting their training corpus with examples of otherwise infrequent types of FGDs (clefts or topicalizations) improved model performance across all four FGD types, under the assumption that training effects should transfer under a shared representation account. The authors found that training did not yield systematic improvement of the model's performance on other FGD types (and in some cases the performance was even degraded). The authors concluded that their LSTM did not have a shared representation underlying all four dependencies and relied on superficial contingencies in the input.

The results of Howitt et al. (2024) do not rule out the broader possibility of cross-dependency generalization (in Norwegian or English). A narrower interpretation is that models tested in Howitt et al. (2024) failed to generalize across the specific set of dependencies tested in English, perhaps due to frequency. Howitt et al. (2024) showed that their model performed best on *wh*-dependencies, which are relatively frequent, as compared to three relatively infrequent dependencies (as estimated by Ozaki et al. 2022). It is possible that even if the English models have adopted an abstract representation of *wh*-dependencies, they did not receive enough evidence of the other three dependencies to extend that representation. Under this interpretation, models would be expected to generalize more readily across *wh*- and RC-dependencies, which are rather frequent (Kobzeva and Kush (2025) show that RC-dependencies are even more frequent than *wh*-dependencies in the kinds of written texts used to train our models). Moreover, there may be even more evidence for cross-dependency generalization in Norwegian, given the prevalence of fronting and topicalization in the language.

It is of course possible that our models, too, fail to generalize across dependencies in any meaningful way, and instead exploit a constellation of superficial piecemeal generalizations, shallow heuristics or lexical co-occurrences to arrive at correct superficial predictions (Kam et al., 2008; McCoy et al., 2019; Kodner and Gupta, 2020; Vázquez Martínez et al., 2024). For example, the models' performance in Experiments 3 and 4 could to some extent be explained by frequency of collocations between verbs introducing embedded questions and the following *wh*-words: there is some correlation between the magnitude of filler effects and the frequency of the corresponding type of embedded question (i.e., the filler effects in Experiment 4 are larger than the ones in Experiment 3, and embedded adjunct questions are more frequent than embedded polar questions). Moreover, Norwegian 'om' has more meanings than English 'whether': it can function as both a complementizer (*if/whether*) and a preposition (*about/around/during*), and therefore appears in more distributional contexts. It has been shown that homonyms can lead to what appear to be correct predictions (Kam et al., 2008), with the models being right for the 'wrong reasons' (McCoy et al., 2019). It is therefore important to examine what features of the input are driving the models' generalizations, and future work leveraging augmented/filtered corpus training could shed light on the exact nature of the models' generalizations (Leong and Linzen, 2024; Patil et al., 2024; Misra and Mahowald, 2024). For example, it would be informative to see how manipulating the presence or absence of non-complementizer examples of 'om' impacts the models' performance on dependencies into embedded polar questions.

5.2. Failures to Approximate Target Generalizations

We discuss below two important instances where the NLMs we evaluated arguably fail to approximate target human generalizations.

First, the results of Experiment 1 indicate that the models' ability to relate fillers to gaps across multiple layers of hierarchical embedding depends on the presence or absence of an overt declarative complementizer (*at/that*). When test sentences did not contain overt complementizers, models showed large filled-gap and unlicensed gap effects up to 4 layers of embedding. However, when test sentences included complementizers, effect sizes dropped precipitously with each new layer. In most cases, any evidence of filler-gap association was absent by the third layer. Thus, the models seem to have induced two separate generalizations: (i) FGDs are unbounded when intervening clauses do not contain overt complementizers, and (ii) FGDs are bounded to 2 or 3 clauses in the presence of overt complementizers. Inasmuch as complementizer presence does not affect human judgments the same way (Ritchart et al., 2016), it appears that the models have, in this case, *undergeneralized* from the input relative to the target state.

Second, although the English models display smaller unlicensed gap effects in subject position inside embedded adjunct questions compared to the control condition (Experiment 4), the models still seem to predict gaps in those positions according to our absolute metric: The size of the unlicensed gap effect ($\approx$ -1 bit of surprisal) was comparable to effects observed in grammatical gap locations in other experiments. Taken at face value, it would seem that the English models have extrapolated to a less restrictive generalization than the human target.

One interpretation of the models' performance in this case bears on their ability to challenge POS arguments and the need for domain-specific biases in acquisition. As discussed above, biases are assumed to guide generalization when the data in the input is equivocal, i.e. compatible with multiple candidate generalizations. They are supposed to prevent both under- and over-generalization. Insofar as the models' failures are taken to represent cases of undergeneralization (complementizer-dependent boundedness) and overgeneralization (*wh*-islands), it seems that the general biases of the NLMs tested here are insufficient to guarantee success, at least when trained on Wikipedia corpora. That is, learning the acceptable distribution of filler-gap dependencies in human language still represents a POS problem (see also Howitt et al. 2024, Lan et al. 2024).

Could the model's failure be attributed to our choice of Wikipedia text as input instead of input that is more representative of child-directed language? It is clear that the distribution of structures differs between Wikipedia text and child-directed speech. Wikipedia text could, in principle, contain fewer cues to the correct generalizations, which could in turn impact model performance. For example, written texts vastly underrepresent the quantity and range of *wh*-questions that are frequent in child-directed speech (Noble et al., 2018). In general, we do not know whether models trained on more realistic input would arrive at the correct conclusions but note that evidence of success with more realistic input is mixed. Though studies show that language models are sensitive to the size and style of their training (van Schijndel et al., 2019; Arehalli and Linzen, 2024) and might learn more efficiently when trained on smaller-scale child-directed language (Mueller and Linzen, 2023; Huebner et al., 2021), models trained on developmentally plausible corpora still fall short in replicating patterns of human judgments (Yedetore et al., 2023). We also point out that, at least for the types of generalizations that the models fail on in our experiments (unboundedness, island sensitivity), child-directed input is unlikely to contain more examples of relevant direct evidence. It is not the case that child-directed input contains significantly more examples of multi-clausal embedding than Wikipedia texts (Pearl and Sprouse, 2013b,a). Moreover, the evidence that embedded questions are islands in English is the *absence* of FGDs into these constituents, so there could not possibly be more *direct* evidence for island-sensitivity. Thus, it seems that the same models are likely to face similar indeterminacy regarding the generalizations with a different corpus.

Warstadt and Bowman (2022) suggest that other differences between the input to children and models could be responsible for the difference. For example, they note that children's input is multi-modal and *grounded*. They argue that information from these extra dimensions could conceivably play a role in correct generalization that our models would be unable to identify.[13] As such, they contend that a model's failure does not clearly *support* POS arguments. We concede the general point, but note that absent a theory of how the additional information exerts this influence, it is a relatively weak and promissory counterargument.

Are there other explanations for the models' suboptimal performance apart from data limitations? Could it be due to random chance, architectural limitations, or the choice of the training objective function? As we have not tested a wider variety of models with different parameters or objectives, we cannot say for certain. For example, it has been shown that the choice of the training objective affects Transformers' preferences for hierarchical generalizations over linear rules (Ahuja et al., 2024), and that different objectives might be required to capture recursive patterns (relevant to the unboundedness generalization) in formal language learning (Lan et al., 2024a). It is therefore possible that different NLM implementations could show better results on problematic cases discussed here, and thus overcome potential POS challenges related to filler-gap dependency acquisition. However, current evidence does not support the claim that FGDs and island constraints on them can be learned without domain-specific biases.

6. Conclusion

In this work, we tested if LSTM and Transformer models trained on Norwegian and English Wikipedia texts can induce major generalizations about the distribution of acceptable filler-gap dependencies in the two languages. Our findings show that although such models do acquire some sophisticated generalizations about filler-gap dependencies in the two languages, their overall predictions still diverge from patterns

[13] See Vázquez Martínez et al. (2024) for a different perspective on the utility of grounding for NLMs.

characteristic of human judgments: In some cases — when tested on structurally complex environments — the models either adopted a narrower generalization than humans do or overgeneralized beyond their input in non-human-like ways. We conclude that current evidence does not support the claim that FGDs and island constraints on them can be learned without domain-specific biases.

Data Availability

Models checkpoints and training data, together with test materials, corpus search data, and analysis scripts are available at the following OSF repository: https://osf.io/2wjcm/.

Acknowledgments

We thank Roni Katzir and two anonymous reviewers for their thoughtful feedback, which significantly improved this paper. All remaining errors are our own. Dave Kush was supported in part by funding from the Social Sciences and Humanities Research Council (Grant #430-2022-00118).

Appendix A. Statistical Analysis: Experiment 1

To compare the differences between different levels of NUMBER OF LAYERS, we used the backward difference coding contrast scheme that compares the mean filler effect at one layer to the mean filler effect for the prior adjacent layer (2 v. 1, 3 v. 2 and so on). The output of the linear mixed-effects models is presented in Table A.6 below.

	Norwegian - RC		Norwegian - Wh		English - Wh	
	FGE	UGE	FGE	UGE	FGE	UGE
	LSTM, without ‘that’ in the embedding layer					
2 layers v. 1	-1.6+	-0.6	0.04	-1.0	0.1	0.05
3 layers v. 2	-1.5	-1.3	-1.2	-0.5	1.4	-0.9
4 layers v. 3	0.4	0.07	0.3	0.2	-2.0*	0.4
5 layers v. 4	1.4	4.1***	-1.3	3.5***	-0.8	1.2+
	GPT-2, without ‘that’ in the embedding layer					
2 layers v. 1	1.4	-1.1	1.6	-1.5	2.2*	-0.7
3 layers v. 2	0.6	-0.3	-0.3	-0.2	1.4	-0.3
4 layers v. 3	-0.4	1.2	-0.2	1.4	-2.4*	0.7
5 layers v. 4	-4.8***	2.7**	-4.0***	2.9**	-4.3***	1.9*
	LSTM, with ‘that’ in the embedding layer					
2 layers v. 1	1.6*	-1.8*	2.5**	-2.3*	3.0***	-1.7*
3 layers v. 2	1.3	-0.7	2.5**	-0.9	2.9**	-1.6+
4 layers v. 3	-0.2	2.2*	1.0	2.7*	-0.2	1.6+
5 layers v. 4	-4.7***	2.9**	-8.9***	3.1**	-8.5***	3.6***
	GPT-2, with ‘that’ in the embedding layer					
2 layers v. 1	3.0**	-2.2*	2.8**	-2.5**	2.8***	-1.1
3 layers v. 2	2.8**	-0.7	2.9**	-1.8+	2.8**	-0.5

4 layers v. 3	0.08	2.9**	2.2*	1.3	0.4	1.1
5 layers v. 4	-9.4***	2.7**	-10.9***	5.6***	-9.1***	2.2**

Table A.6: Output of the linear mixed-effects models for Experiment 1 that tested unboundedness. Reported values are model coefficients and diacritics represent significance levels (+p <0.1, *p <.05, **p <.01, ***p <.001).

Appendix B. Corpus Findings

Table B.7: Dependencies into embedded questions found in the Wikipedia corpus.

Sentence text or fragment	EQ verb	Wh-word	Gap	Source (clickable)
Compsognathus er en av de få dinosaurene vi vet hva spiste.	vite	hva	subj	Compsognathus
[...] ”Sordello”, som ingen begrep hva handlet om [...]	begripe	hva	subj	Robert Browning
[...] en [...] idé, som man sliter med å forstå hva dreier seg om.	forstå	hva	subj	No Wikipedia source found
[...] pengesummene [...], har ikke medlemmene fått vite hva er brukt på.	vite	hva	subj	Mohammad Tahir ul-Qadri
Heller ikke elektrisitet kunne en forklare hva var.	forklare	hva	subj	Kristian Birkeland
[...] seksuell aktivitet som de ikke samtykker til og kanskje ikke forstår hva er.	forstå	hva	subj	Barnemishandling
[...] noen rare lyder som han ikke skjønner hva er.	skjønne	hva	subj	After.Life
[...] det som du overhodet ikke vet hva er?	vite	hva	subj	Menon
[...] der de fant forskjellige ting som de ikke helt vet hva er.	vite	hva	subj	Fimlene
Men det står en ved siden av han, som de ikke helt ser hvem er.	se	hvem	subj	Milliardæren
[...] det er tøft gjort å gå rett inn i et rom med menn man ikke vet hvem er [...]	vite	hvem	subj	Disturbed
[...] sportsutøvere og ulike samfunnsaktører som svært mange vet hvem er.	vite	hvem	subj	Kjendis
[...] det var en person han visste hvem var.	vite	hvem	subj	Pengegaloppen
[...] et band heavy metal-tilhengere visste hvem var.	vite	hvem	subj	Sodom
[...] en ny gjest som bare den ene av programlederne viste hvem var.	vite	hvem	subj	Par-i-bol
[...] et par barnesko han ikke kan huske hvor kommer fra [...]	huske	hvor	subj	Jul i Skomakergata
[...] de mystiske haukakarane, som ingen vet hvor kom fra [...]	vite	hvor	subj	Rau’e Aarhanen spelle
Den siste er det ikke kjent hvor ble levert [...]	kjenne	hvor	subj	Volkswagen Transporter
[...] ”hemmelige” benker som man helst ikke skal røpe hvor er.	røpe	hvor	subj	Godliaskogen
Disse maleriene forsvant og det er få av dem man vet hvor er i dag.	vite	hvor	subj	Nikolaj Ge
[...] det også finnes en annen gravstatue som ingen vet hvor er.	vite	hvor	subj	Tordivelen flyr i skumringen
[...] tradisjonell kunnskap en vanskelig kan si når oppsto.	si	når	subj	Strikking
[...] en situasjon regjeringen ikke visste hvordan de skulle håndtere.	vite	hvordan	obj	Holocaust i Slovakia
[...] å gi ham komplimenter han er usikker på om han fortjener.	være usikker	om	obj	Knøttene
[...] ei setningsknute: ”den boka veit jeg ikke om jeg har lest”.	vite	om	obj	Setningsknute
[...] en rolle vi ikke vet om han har spilt.	vite	om	obj	Lukket avdeling
[...] kommentarer som man ikke kan være sikker på om er frekt ment [...]	sikker	om	subj	Andrew Luck
[...] Silver som van Onselen spekulerer om kunne ha vært Jack the Ripper.	spekulere	om	subj	Charles van Onselen
[...] forutsetninger som domstolen selv plikter å undersøke om er på plass [...]	undersøke	om	subj	Norsk sivilprosess
[...] de første gjerningsmenn som myndighetene undersøkte om var tilregnelig.	undersøke	om	subj	Wozzeck
[...] misjonærer som [...] man er usikker på om faktisk kom dit [...]	være usikker	om	subj	Liste over kinamisjonærer...
Disse særtrekkene, som det ikke er visst om eksisterte i uraustroasiatisk [...]	vite	om	subj	Vietnamesisk
[...] nisjer i veggene, som man lurte på om kunne ha inneholdt de kremerte restene av [...]	lure	om	subj	Hettittene

References

Ahuja, K., Balachandran, V., Panwar, M., He, T., Smith, N.A., Goyal, N., Tsvetkov, Y., 2024. Learning syntax without planting trees: Understanding when and why Transformers generalize hierarchically. arXiv preprint arXiv:2404.16367 .

Arehalli, S., Linzen, T., 2024. Neural networks as cognitive models of the processing of syntactic constraints. Open Mind 8, 558–614. https://doi.org/10.1162/opmi_a_00137, 10.1162/opmi_a_00137.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. Journal of memory and language 68, 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. Journal of Statistical Software 67, 1–48. 10.18637/jss.v067.i01.

Bernardy, J.P., Lappin, S., 2017. Using deep neural networks to learn syntactic agreement. Linguistic Issues in Language Technology 15. https://aclanthology.org/2017.lilt-15.3/.

Bhattacharya, D., van Schijndel, M., 2020. Filler-gaps that neural networks fail to generalize, in: Fernández, R., Linzen, T. (Eds.), Proceedings of the 24th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Online. pp. 486–495. 10.18653/v1/2020.conll-1.39.

Chaves, R.P., 2020. What don't RNN language models learn about filler-gap dependencies?, in: Society for Computation in Linguistics, University of Massachusetts Amherst Libraries. pp. 20–30.

Chomsky, N., 1965. Aspects of the theory of syntax. Cambridge: the MIT press.

Chomsky, N., 1971. Problems of knowledge and freedom: The Russell lectures.

Chomsky, N., 1973. Conditions on transformations, in: Halle, M., Anderson, S.R., Kiparsky, P. (Eds.), A Festschrift for Morris Halle. Holt, Rinehart and Winston, New York, pp. 232–286.

Chomsky, N., 2001. Derivation by phase, in: Kenstowicz, M. (Ed.), Ken Hale: A life in language. Cambridge: The MIT Press, pp. 1–52.

Chomsky, N., Lasnik, H., 1977. Filters and control. Linguistic Inquiry 8, 425–504. http://www.jstor.org/stable/4177996.

Chow, W.Y., Zhou, Y., 2019. Eye-tracking evidence for active gap-filling regardless of dependency length. Quarterly Journal of Experimental Psychology 72, 1297–1307. 10.1177/1747021818804988.

Chowdhury, S.A., Zamparelli, R., 2018. RNN simulations of grammaticality judgments on long-distance dependencies, in: Proceedings of the 27th international conference on computational linguistics, pp. 133–144.

Christensen, K.K., 1982. On multiple filler-gap constructions in Norwegian, in: Engdahl, E., Ejerhed, E. (Eds.), Readings on unbounded dependencies in Scandinavian languages. Almquist & Wiksell, Stockholm, pp. 77–98.

Christiansen, M.H., Chater, N., 2016. Creating language: Integrating evolution, acquisition, and processing. Mit Press.

Clark, A., Lappin, S., 2010. Linguistic nativism and the poverty of the stimulus. John Wiley & Sons.

Clark, A., Lappin, S., 2012. Computational learning theory and language acquisition, in: Kempson, R., Asher, N., Fernando, T. (Eds.), Philosophy of Linguistics. Elsevier, pp. 445—-475. `http://dx.doi.org/10.1016/B978-0-444-51747-0.50013-5`, `10.1016/b978-0-444-51747-0.50013-5`.

Crain, S., Fodor, J.D., 1985. How can grammars help parsers?, in: Dowty, D.R., Karttunen, L., Zwicky, A. (Eds.), Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives. Cambridge University Press, pp. 94–128. `http://dx.doi.org/10.1017/CBO9780511597855.004`, `10.1017/cbo9780511597855.004`.

Crain, S., Pietroski, P., 2001. Nature, nurture and universal grammar. Linguistics and philosophy 24, 139–186.

Cuskley, C., Woods, R., Flaherty, M., 2024. The limitations of large language models for understanding human language and cognition. Open Mind 8, 1058–1083. `https://doi.org/10.1162/opmi\_a\_00160`.

Da Costa, J.K., Chaves, R.P., 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. Proceedings of the Society for Computation in Linguistics 3, 189–198.

Dickson, N., Pearl, L., Futrell, R., 2022. Learning constraints on wh-dependencies by learning how to efficiently represent wh-dependencies: A developmental modeling investigation with Fragment Grammars. Proceedings of the Society for Computation in Linguistics 5, 220–224. `10.7275/7fd4-fw49`.

Frank, M.C., 2023. Bridging the data gap between children and large language models. PsyArXiv Preprints `10.31234/osf.io/qzbgx`.

Gilkerson, J., Richards, J.A., Warren, S.F., Montgomery, J.K., Greenwood, C.R., Oller, D.K., Hansen, J.H.L., Paul, T.D., 2017. Mapping the early language environment using all-day recordings and automated analysis. American Journal of Speech-Language Pathology 26, 248–265. 10.1044/2016_AJSLP-15-0169.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., Baroni, M., 2018. Colorless green recurrent networks dream hierarchically, in: Proceedings of NAACL 2018, pp. 1195–1205. https://aclanthology.org/N18-1108, 10.18653/v1/N18-1108.

Gulrajani, A., Lidz, J., 2024. Reassessing a model of syntactic island acquisition, in: Proceedings of the Society for Computation in Linguistics 2024, pp. 43–51. https://doi.org/10.7275/scil.2128.

Hale, J., 2001. A probabilistic Earley parser as a psycholinguistic model, in: Second meeting of the North American chapter of the Association for Computational Linguistics, pp. 1–8. https://doi.org/10.3115/1073336.1073357.

Hart, B., Risley, T.R., 1992. American parenting of language-learning children: Persisting differences in family-child interactions observed in natural home environments. Developmental psychology 28, 1096.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural computation 9, 1735–1780.

Hollebrandse, B., Roeper, T., 2014. Empirical results and formal approaches to recursion in acquisition, in: Studies in Theoretical Psycholinguistics. Springer International Publishing, p. 179–219. http://dx.doi.org/10.1007/978-3-319-05086-7_9, 10.1007/978-3-319-05086-7_9.

Howitt, K., Nair, S., Dods, A., Hopkins, R.M., 2024. Generalizations across filler-gap dependencies in neural language models, in: Barak, L., Alikhani, M. (Eds.), Proceedings of the 28th Conference on Computational Natural Language Learning, Association for Computational Linguistics, Miami, FL, USA. pp. 269–279. https://aclanthology.org/2024.conll-1.21/, 10.18653/v1/2024.conll-1.21.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., Levy, R.P., 2020. A systematic assessment of syntactic generalization in neural language models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 1725–1744. https://aclanthology.org/2020.acl-main.158, 10.18653/v1/2020.acl-main.158.

Huebner, P.A., Sulem, E., Cynthia, F., Roth, D., 2021. BabyBERTa: Learning more grammar with small-scale child-directed language, in: Proceedings of the 25th conference on computational natural language learning, pp. 624–646.

Kam, X.N.C., Stoyneshka, I., Tornyova, L., Fodor, J.D., Sakas, W.G., 2008. Bigrams and the richness of the stimulus. Cognitive science 32, 771–787.

Katzir, R., 2023. Why large language models are poor theories of human linguistic cognition: A reply to Piantadosi. Biolinguistics 17, 1–12. https://doi.org/10.5964/bioling.13153.

Kobzeva, A., Arehalli, S., Linzen, T., Kush, D., 2022a. LSTMs can learn basic wh- and relative clause dependencies in Norwegian, in: Proceedings of the Annual Meeting of the Cognitive Science Society, pp. 2974–2980. https://escholarship.org/uc/item/012683gb.

Kobzeva, A., Arehalli, S., Linzen, T., Kush, D., 2023. Neural networks can learn patterns of island-insensitivity in Norwegian, in: Proceedings of the Society for Computation in Linguistics, pp. 175–185. https://doi.org/10.7275/qb8z-qc91.

Kobzeva, A., Kush, D., 2024. Grammar and expectation in active dependency resolution: Experimental and modeling evidence from Norwegian. Cognitive Science 48, e13501. https://doi.org/10.1111/cogs.13501.

Kobzeva, A., Kush, D., 2025. Acquiring constraints on filler-gap dependencies from structural collocations: Assessing a computational learning model of island-insensitivity in Norwegian. 10.1080/10489223.2024.2440340, arXiv:https://doi.org/10.1080/10489223.2024.2440340.

Kobzeva, A., Sant, C., Robbins, P.T., Vos, M., Lohndal, T., Kush, D., 2022b. Comparing island effects for different dependency types in Norwegian. Languages 7, 195–220. https://doi.org/10.3390/languages7030197.

Kodner, J., Gupta, N., 2020. Overestimation of syntactic representation in neural language models, in: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics. pp. 1757–1762. https://aclanthology.org/2020.acl-main.160/, 10.18653/v1/2020.acl-main.160.

Kodner, J., Payne, S., Heinz, J., 2023. Why linguistics will thrive in the 21st century: A reply to Piantadosi (2023). arXiv preprint arXiv:2308.03228 10.48550/arXiv.2308.03228.

Kush, D., Dahl, A., 2020. L2 transfer of L1 island-insensitivity: The case of Norwegian. Second Language Research , 1–32https://doi.org/10.1177/0267658320956704, 10.1177/0267658320956704.

Kush, D., Dahl, A., Lindahl, F., 2023. Filler–gap dependencies and islands in L2 English production: Comparing transfer from L1 Norwegian and L1 Swedish. Second Language Research http://dx.doi.org/10.1177/02676583231172918, 10.1177/02676583231172918.

Kush, D., Lohndal, T., Sprouse, J., 2018. Investigating variation in island effects: A case study of Norwegian wh-extraction. Natural Language & Linguistic Theory 36, 743–779. https://doi.org/10.1007/s11049-017-9390-z, 10.1007/s11049-017-9390-z.

Kush, D., Lohndal, T., Sprouse, J., 2019. On the island sensitivity of topicalization in Norwegian: An experimental investigation. Language 95, 393–420. https://doi.org/10.1353/lan.2019.0051, 10.1353/lan.2019.0051.

Kush, D., Sant, C., Strætkvern, S.B., 2021. Learning island-insensitivity from the input: A corpus analysis of child- and youth-directed text in Norwegian. Glossa: a journal of general linguistics 6, 1–50. https://doi.org/10.16995/glossa.5774, 10.16995/glossa.5774.

Lake, B.M., Baroni, M., 2023. Human-like systematic generalization through a meta-learning neural network. Nature 623, 115–121. http://dx.doi.org/10.1038/s41586-023-06668-3, 10.1038/s41586-023-06668-3.

Lan, N., Chemla, E., Katzir, R., 2024a. Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length, in: Ku, L.W., Martins, A., Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 13198–13210. https://aclanthology.org/2024.acl-long.713/, 10.18653/v1/2024.acl-long.713.

Lan, N., Chemla, E., Katzir, R., 2024b. Large language models and the argument from the poverty of the stimulus. Linguistic Inquiry , 1–56https://doi.org/10.1162/ling_a_00533.

Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review 104, 211.

Lasnik, H., Lidz, J.L., 2016. The argument from the poverty of the stimulus, in: Roberts, I. (Ed.), The Oxford Handbook of Universal Grammar. Oxford University Press, p. 220–248. http://dx.doi.org/10.1093/oxfordhb/9780199573776.013.10, 10.1093/oxfordhb/9780199573776.013.10.

Leong, C.S.Y., Linzen, T., 2024. Testing learning hypotheses using neural networks by manipulating learning data. arXiv https://arxiv.org/abs/2407.04593, https://doi.org/10.48550/arXiv.2407.04593, arXiv:2407.04593.

Levy, R., 2008. Expectation-based syntactic comprehension. Cognition 106, 1126–1177. https://doi.org/10.1016/j.cognition.2007.05.006.

Linzen, T., Baroni, M., 2021. Syntactic structure from deep learning. Annual Review of Linguistics 7, 195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035.

Linzen, T., Dupoux, E., Goldberg, Y., 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. Transactions of the Association for Computational Linguistics 4, 521–535.

McCoy, R.T., Pavlick, E., Linzen, T., 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3428–3448.

McDaniel, D., Cowart, W., McKee, C., Garrett, M.F., 2015. The role of the language production system in shaping grammars. Language , 415–441.

Michaelov, J.A., Bardolph, M.D., Van Petten, C.K., Bergen, B.K., Coulson, S., 2024. Strong prediction: Language model surprisal explains multiple N400 effects. Neurobiology of Language 5, 107–135. http://dx.doi.org/10.1162/nol_a_00105, 10.1162/nol_a_00105.

Misra, K., Mahowald, K., 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. arXiv preprint arXiv:2403.19827 .

Morgan, A.M., 2022. The that-trace effect and island boundary-gap effect are the same: Demonstrating equivalence with null hypothesis significance testing and psychometrics. Glossa Psycholinguistics 1. http://dx.doi.org/10.5070/G601140, 10.5070/g601140.

Mueller, A., Linzen, T., 2023. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases . arXiv preprint arXiv:2305.19905 .

Noble, C.H., Cameron-Faulkner, T., Lieven, E., 2018. Keeping it simple: The grammatical properties of shared book reading. Journal of Child Language 45, 753–766.

Ozaki, S., Yurovsky, D., Levin, L., 2022. How well do LSTM language models learn filler-gap dependencies?, in: Proceedings of the Society for Computation in Linguistics 2022, pp. 76–88.

Patil, A., Jumelet, J., Chiu, Y.Y., Lapastora, Wang, L., Willrich, C., Steinert-Threlkeld, S., 2024. Filtered corpus training (FiCT) shows that language models can generalize from indirect evidence. arXiv preprint arXiv:2405.15750 10.48550/arXiv.2405.15750.

Pañeda, C., Kush, D., Villata, S., Sprouse, J., 2024. A translation-matched, experimental comparison of three types of wh-island effects in Spanish and English. Glossa: A journal of general linguistics 9. https://doi.org/10.16995/glossa.11164.

Pearl, L., 2022. Poverty of the stimulus without tears. Language Learning and Development 18, 415–454.

Pearl, L., Bates, A., 2022. A new way to identify if variation in children's input could be developmentally meaningful: Using computational cognitive modeling to assess input across socio-economic status for syntactic islands. Journal of Child Language , 1–34https://doi.org/10.1017/S0305000922000514.

Pearl, L., Sprouse, J., 2013a. Computational models of acquisition for islands, in: Sprouse, J., Hornstein, N. (Eds.), Experimental Syntax and Island Effects. Cambridge University Press, pp. 109–131. 10.1017/CBO9781139035309.006.

Pearl, L., Sprouse, J., 2013b. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. Language Acquisition 20, 23–68.

Perfors, A., Tenenbaum, J.B., Regier, T., 2011. The learnability of abstract syntactic principles. Cognition 118, 306–338. http://dx.doi.org/10.1016/j.cognition.2010.11.001, 10.1016/j.cognition.2010.11.001.

Perlmutter, D.M., 1971. Deep and surface structure constraints in syntax. Holt, Rinehart & Winston.

Phillips, C., 2013a. On the nature of island constraints I: Language processing and reductionist accounts, in: Sprouse, J., Hornstein, N. (Eds.), Experimental syntax and island effects. Cambridge University Press, pp. 64–108. http://www.colinphillips.net/wp-content/uploads/2014/08/phillips2013_islands1.pdf, https://doi.org/10.1017/CBO9781139035309.005.

Phillips, C., 2013b. On the nature of island constraints II: Language learning and innateness, in: Experimental syntax and island effects. Cambridge University Press, pp. 132–158. https://doi.org/10.1017/CBO9781139035309.007.

Piantadosi, S.T., 2023. Modern language models refute Chomsky's approach to language, in: Gibson, E., Poliak, M. (Eds.), From fieldwork to linguistic theory: A tribute to Dan Everett, pp. 353–414.

Potts, C., 2023. Characterizing English preposing in PP constructions. LingBuzz lingbuzz/007495 .

Prasad, G., van Schijndel, M., Linzen, T., 2019. Using priming to uncover the organization of syntactic representations in neural language models, in: Bansal, M., Villavicencio, A. (Eds.), Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL), Association for Computational Linguistics, Hong Kong, China. pp. 66–76. 10.18653/v1/K19-1007.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D., 2020. Stanza: A Python natural language processing toolkit for many human languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 101–108. https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Reali, F., Christiansen, M.H., 2005. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. Cognitive Science 29, 1007–1028. http://dx.doi.org/10.1207/s15516709cog0000_28, 10.1207/s15516709cog0000_28.

Ritchart, A., Goodall, G., Garellek, M., 2016. Prosody and the that-trace effect: An experimental study, in: 33rd West Coast Conference on Formal Linguistics, Cascadilla Proceedings Project. pp. 320–328.

Rizzi, L., 1982. Violations of the wh island constraint in Italian and the subjacency condition, in: Issues in Italian Syntax. Dordrecht, pp. 49–76.

Ross, J.R., 1967. Constraints on variables in syntax. Ph.D. thesis. MIT. https://dspace.mit.edu/handle/1721.1/15166.

van Schijndel, M., Mueller, A., Linzen, T., 2019. Quantity doesn't buy quality syntax with neural language models, in: Inui (Ed.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 5831–5837. 10.18653/v1/D19-1592.

Shain, C., Meister, C., Pimentel, T., Cotterell, R., Levy, R., 2024. Large-scale evidence for logarithmic effects of word predictability on reading time. Proceedings of the National Academy of Sciences 121, e2307876121. https://doi.org/10.1073/pnas.2307876121.

Smith, N.J., Levy, R., 2013. The effect of word predictability on reading time is logarithmic. Cognition 128, 302–319. https://doi.org/10.1016/j.cognition.2013.02.013.

Sobin, N., 1987. The variable status of Comp-trace phenomena. Natural Language & Linguistic Theory 5, 33–60.

Sprouse, J., Caponigro, I., Greco, 2016. Experimental syntax and the variation of island effects in English and Italian. Natural Language & Linguistic Theory 34, 307–344. https://doi.org/10.1007/s11049-015-9286-8, 10.1007/s11049-015-9286-8.

Sprouse, J., Wagers, M., Phillips, C., 2012. A test of the relation between working-memory capacity and syntactic island effects. Language , 82–123.

Stowe, L.A., 1986. Parsing wh-constructions: Evidence for on-line gap location. Language and cognitive processes 1, 227–245. https://doi.org/10.1080/01690968608407062, 10.1080/01690968608407062.

Suijkerbuijk, M., de Swart, P., Frank, S.L., 2023. The learnability of the wh-island constraint in Dutch by a long short-term memory network, in: Proceedings of the Society for Computation in Linguistics 2023, pp. 321–331.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Conference on Neural Information Processing Systems (NIPS 2017) 30.

Vázquez Martínez, H.J., Heuser, A.L., Yang, C., Kodner, J., 2024. Evaluating the existence proof: LLMs as cognitive models of language acquisition, in: Mendivil-Giro, J.L. (Ed.), Artificial Knowledge of Language. Vernon Press. https://lingbuzz.net/lingbuzz/008277.

Warstadt, A., Bowman, S.R., 2022. What artificial neural networks can tell us about human language acquisition. Algebraic Structures in Natural Language , 17–60.

Wilcox, E., Levy, R., Futrell, R., 2019a. Hierarchical representation in neural language models: Suppression and recovery of expectations, in: Proceedings of the 2019 ACL Workshop BlackboxNLP, pp. 181–190. 10.18653/v1/W19-4819.

Wilcox, E., Levy, R., Futrell, R., 2019b. What syntactic structures block dependencies in RNN language models? arXiv preprint arXiv:1905.10431 .

Wilcox, E., Levy, R., Morita, T., Futrell, R., 2018. What do RNN language models learn about filler-gap dependencies?, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP, pp. 211–221. https://aclanthology.org/W18-5423, 10.18653/v1/W18-5423.

Wilcox, E.G., Futrell, R., Levy, R., 2023. Using computational models to test syntactic learnability. Linguistic Inquiry , 1–44http://dx.doi.org/10.1162/ling_a_00491, 10.1162/ling_a_00491.

Yedetore, A., Linzen, T., Frank, R., McCoy, R.T., 2023. How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 9370–9393. https://aclanthology.org/2023.acl-long.521, 10.18653/v1/2023.acl-long.521.