Journal of Memory and Language 144 (2025) 104663

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

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

ARTICLE INFO

Dataset link: https://osf.io/2wjcm/

Keywords:
Filler-gap dependencies
Island constraints
Learnability
Neural Language Models
Norwegian

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.

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 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; Crain & Pietroski, 2001; Lasnik & Lidz, 2016; Phillips, 2013a). 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 (Christiansen & Chater, 2016; Clark & Lappin, 2010, 2012; Landauer & Dumais, 1997; Perfors et al., 2011; Reali & Christiansen, 2005). 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

* Corresponding author.
E-mail address: anastasia.kobzeva@ntnu.no (A. Kobzeva).

[1] Others hold that NLMs can even implement genuine theories of language (Piantadosi, 2023) — a view that has recently received much critique (Cuskley et al. 2024, Katzir 2023, Kodner et al. 2023, 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.

https://doi.org/10.1016/j.jml.2025.104663
Received 30 October 2023; Received in revised form 5 June 2025; Accepted 6 June 2025

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 (Bernardy & Lappin, 2017; Gulordava et al., 2018; Linzen et al., 2016), many studies have uncovered their impressive performance on certain structure-dependent linguistic phenomena (Ahuja et al. 2024, Hu et al. 2020, Lake and Baroni 2023, Linzen and Baroni 2021, 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 (Bhattacharya & van Schijndel, 2020; Chaves, 2020; Chowdhury & Zamparelli, 2018; Howitt et al., 2024; Lan et al., 2024b; Ozaki et al., 2022; Suijkerbuijk et al., 2023; Wilcox et al., 2023, 2019a, 2019b). 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.

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
Which topic said you that Marit thought that Odd knew that
… du skrev artikkelen din om _$_i$?]]]]
… you wrote 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 (Kobzeva et al., 2022; Kush et al., 2018, 2019; Sprouse et al., 2016, 2012).

(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; Kobzeva et al., 2022; Kush & Dahl, 2020; Kush et al., 2023, 2021).

(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
That was book.DEF REL I wondered on whether he had
lest _$_i$].
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; Pearl, 2022; Phillips, 2013b) 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 & Roeper, 2014; Pearl & 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 Dickson et al. 2022, Gulrajani and Lidz 2024, Pearl and Bates 2022).

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., 2023, 2019a, 2019b, 2018). 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 (Bhattacharya & van Schijndel, 2020; Chaves, 2020; Da Costa & Chaves, 2020; Howitt et al., 2024; Lan et al., 2024b). Moreover, NLMs' performance has been shown to vary depending on the type of FGD tested (Howitt et al., 2024; Ozaki et al., 2022), 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 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 models use to extract their 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 is 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] In particular, their modeled learner was trained on syntactically annotated child input and hardwired 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.

Method

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 & 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 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 & 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 & 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 & Bowman, 2022).

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 (Michaelov et al., 2024; Shain et al., 2024; Smith & Levy, 2013). In our models, surprisal values were calculated over the models' respective vocabularies calculated from their softmax layer.

[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.

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 model is 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 model's 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 & 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. −FGD
My brother wanted to know if Ruth will bring us home to Mom at Christmas.
b. +FGD
My brother wanted to know who Ruth will bring us home to _ at Christmas.

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.

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 (Kobzeva et al., 2022a, 2023; Wilcox et al., 2023, 2018), 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 is 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 & Kush, 2024).

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).

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. (2023, 2018), 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

'He found out from someone that the student used AI on the exam.'

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

'*He found out about something that the student used AI on the exam.'

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.'

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. −FGD
He got to know from someone that the student used AI on the exam.

b. +FGD
He got to know about something that the student used _ on the exam.

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.

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.

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
He knows what she thought that parents.DEF found out that
skolen mistenkte [$_4$at læreren visste [$_5$at studenten
school.DEF suspected that teacher.DEF knew that student.DEF
brukte _ på prøven.]]]]]
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
He knows what she thought parents.DEF found out school.DEF
mistenkte [$_4$læreren visste [$_5$studenten brukte _ på prøven.]]]]]
suspected 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'.

[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.

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.

Experiment 1: Results and discussion

The results of the Unboundedness experiment are presented in Fig. 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. (2023, 2018) and extend it to Norwegian *wh*- and RC-dependencies.

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 Fig. 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 (Da Costa & Chaves, 2020; Wilcox et al., 2023).

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. Statistical Analysis: Experiment 1", 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).

[6] A recent eye-tracking study (Chow & 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).

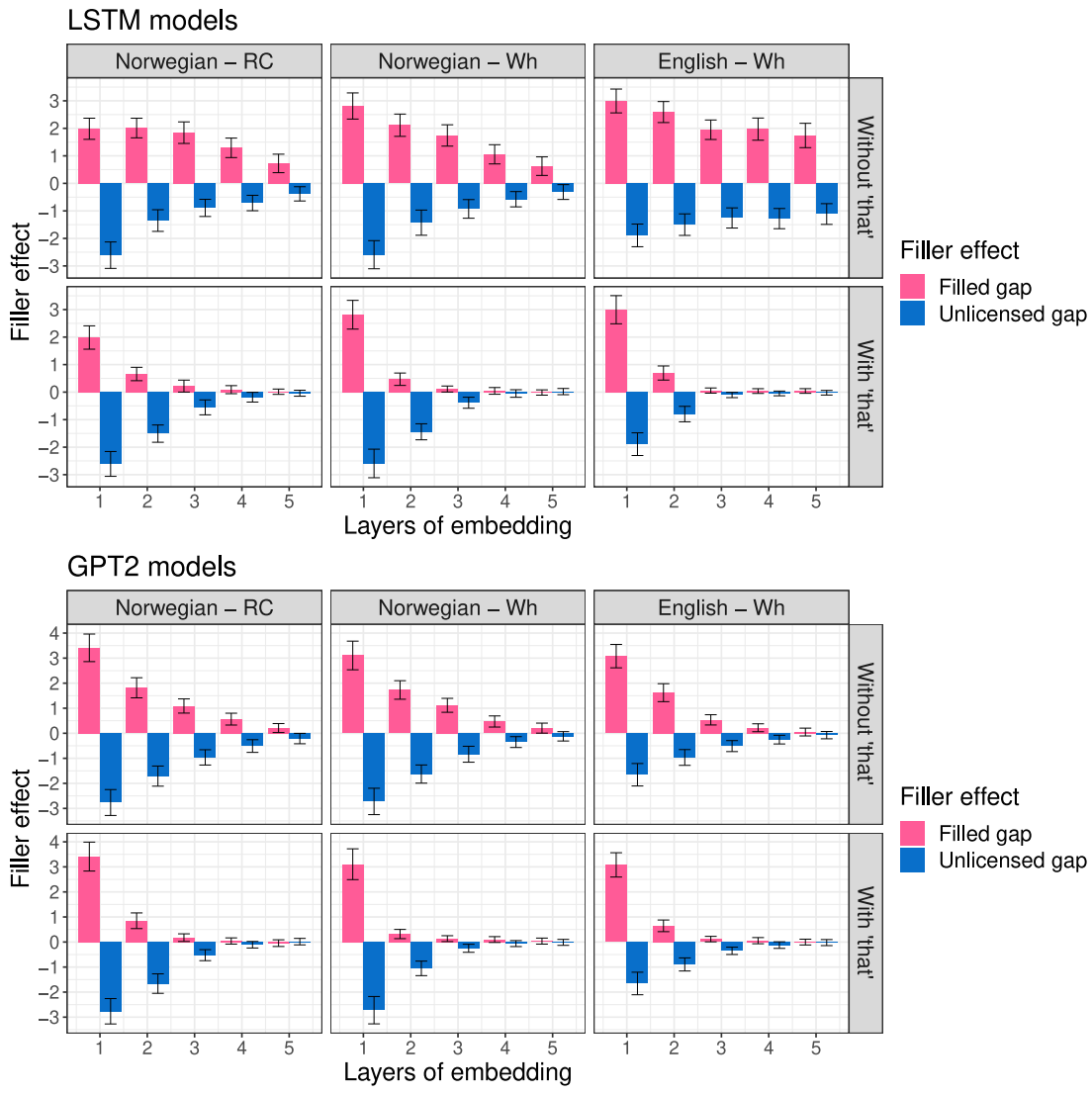

Fig. 1. Results of Experiment 1 testing unboundedness of filler-gap dependencies. Error bars represent 95% confidence intervals. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

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.

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.

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
She discovered that information.DEF in letter.DEF will confirm
mistanken under rettssaken.
suspicion.DEF 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
She discovered that information.DEF in _ will confirm
mistanken under rettssaken.
suspicion.DEF during 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
She discovered what information.DEF in letter.DEF will
bekrefte mistanken under rettssaken.
confirm suspicion.DEF 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
She discovered what information.DEF in _ will confirm
mistanken under rettssaken.
suspicion.DEF during 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
She discovered what C will confirm suspicion.DEF during
rettssaken.
trial.DEF
'She discovered what _ will confirm the suspicion during the trial.'

b. EMBEDDED CONTROL (+FILLER, +GAP)

Hun oppdaget hva han trodde _ vil bekrefte
She discovered what he believed will confirm
mistanken under rettssaken.
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).

Experiment 2: Results and discussion

A breakdown of filler effects by condition and dependency is presented in Fig. 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.

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).

	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***

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.

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 (Kobzeva et al., 2022; Kush et al., 2018, 2019, 2021), whereas English speakers consistently exhibit island effects when judging such constructions (Pañeda et al., 2024; Sprouse et al., 2016, 2012). Recent work has suggested that NLMs trained on English corpora exhibit *whether*-island sensitivity (Wilcox et al., 2023, 2018). We tested whether NLMs trained on Norwegian data would arrive at a different conclusion.

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

[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.

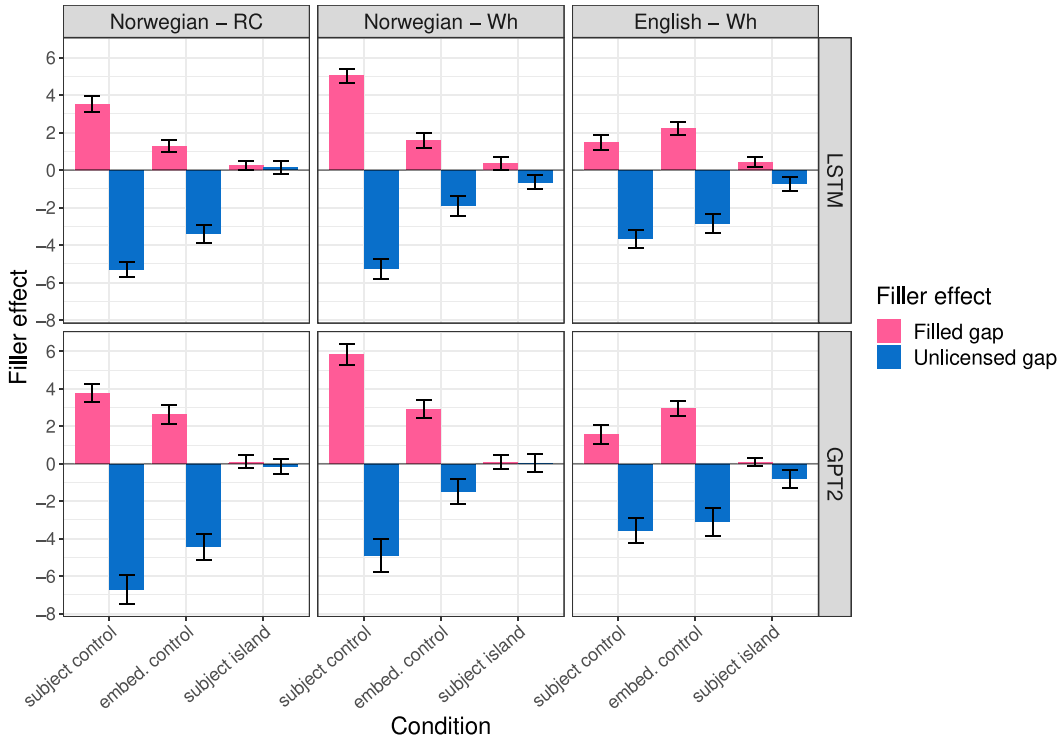

Fig. 2. Results of Experiment 2 testing subject islands. Error bars represent 95% confidence intervals.

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
He knows what professor.DEF could tell that student.DEF
brukte _ på prøven.
used _ on 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
He knows what professor.DEF could tell whether
studenten brukte _ på prøven.
student.DEF used _ on 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.

Experiment 3: Results and discussion

Average filler effects across comparisons in Experiment 3 are presented in Fig. 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 Fig. 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.

[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.

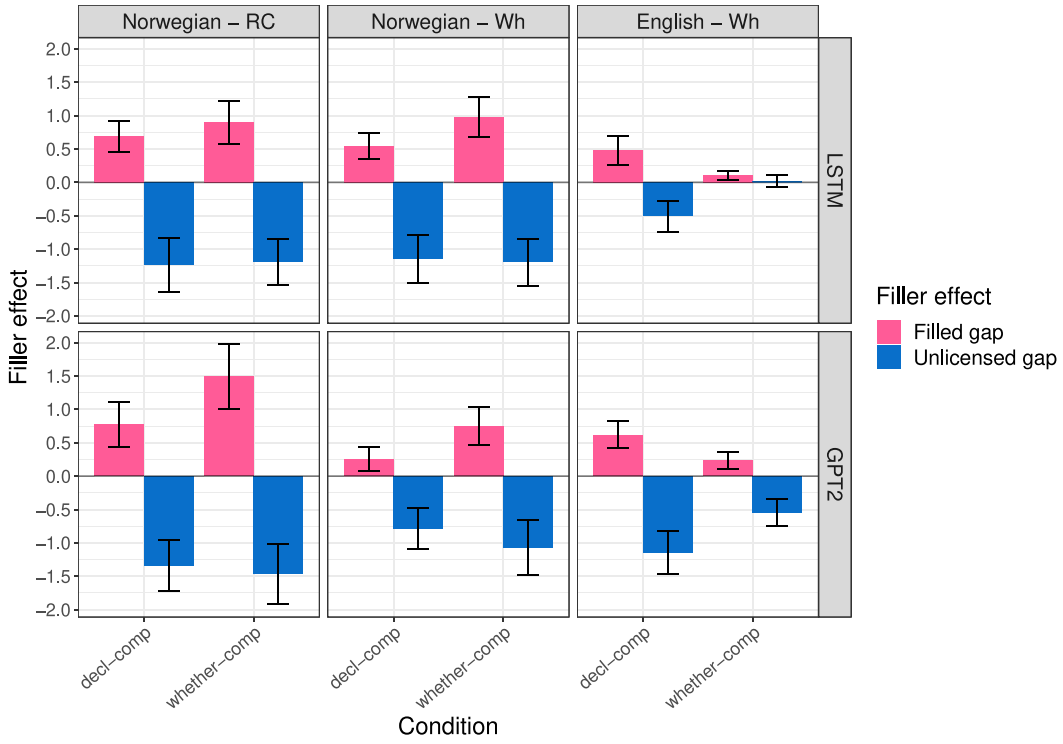

Fig. 3. Results from Experiment 3 testing object extraction from embedded polar questions/'whether-islands'. Error bars represent 95% confidence intervals.

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).

	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***

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 do not 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 (Kush & Dahl, 2020; Kush et al., 2023; McDaniel et al., 2015; Morgan, 2022). 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 Chomsky and Lasnik 1977, Morgan 2022, Perlmutter 1971, a.o.). As such, a successful English model should not allow fillers to be related to subject gaps in the sentences.

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 (Chomsky & Lasnik, 1977; Perlmutter, 1971; 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.

Experiment 4: Results and discussion

The results of Experiment 4 are presented in Fig. 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.

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.

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.

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).

	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***

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.

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 models generalized.

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

[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.

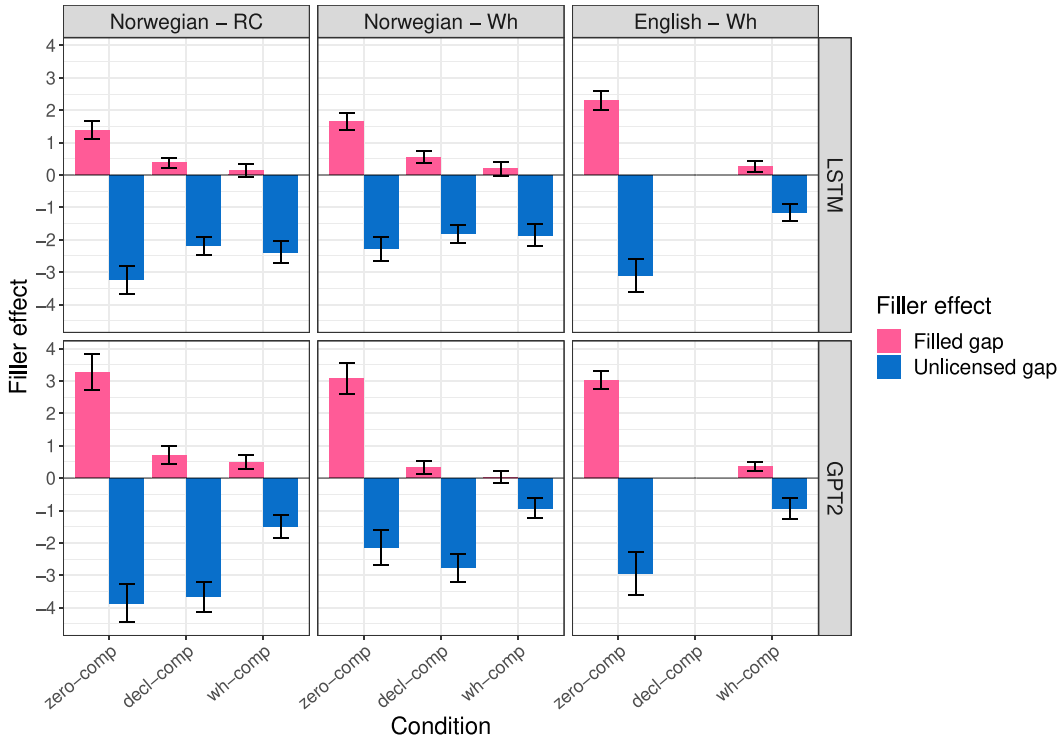

Fig. 4. Results of Experiment 4 testing subject extraction from embedded adjunct questions/'wh-islands'. Error bars represent 95% confidence intervals.

Table 4
Descriptive statistics for different types of EQs in the Norwegian Wikipedia corpus.

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 5
Summary of dependencies into different types of embedded questions from the manually checked portion of the Norwegian Wikipedia corpus.

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

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.

From 5688 sentences summarized above, we found 33 (0.6%) sentences that contain FGDs into embedded questions. All 33 sentences can be found in Table B.7 in the Appendix. 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.

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
On field.DEF surprises Luck opponents with comments REF
man ikke kan være sikker på om _ er frekt ment …
man NEG can be 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.

[10] Source: Andrew Luck Wikipedia page

(22) Alt dette var en del av [tradisjonell kunnskap]$_i$ en vanskelig kan
All this was a part of traditional knowledge one difficult can
si [når _$_i$ oppsto.]
say when arose.
lit. 'All of this was part of traditional knowledge that one can hardly say when _ arose.'[11]

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]

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 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 (Lan et al., 2024b; Wilcox et al., 2023).

[11] Source: 'Strikking [Knitting]' Wikipedia page

[12] A similar claim could be made based on results from Experiment 4.

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.

(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.

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 (Leong & Linzen, 2024; Misra & Mahowald, 2024; Patil et al., 2024; Potts, 2023), 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 & 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 & 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; Kodner & Gupta, 2020; McCoy et al., 2019; 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 & Linzen, 2024; Misra & Mahowald, 2024; Patil et al., 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.

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 (≈ -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 (Arehalli & Linzen, 2024; van Schijndel et al., 2019) and might learn more efficiently when trained on smaller-scale child-directed language (Huebner et al., 2021; Mueller & Linzen, 2023), 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 & Sprouse, 2013a, 2013b). 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.

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 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.

CRediT authorship contribution statement

Anastasia Kobzeva: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Suhas Arehalli:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Tal Linzen:** Writing – review & editing, Supervision, Conceptualization. **Dave Kush:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Conceptualization.

[13] See Vázquez Martínez et al. (2024) for a different perspective on the utility of grounding for NLMs.

Declaration of competing interest

The authors declare no competing interests. This research was conducted independently and was not influenced by any commercial, financial, or personal relationships that could be construed as a potential conflict of interest.

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.

Appendix B. Corpus findings

See Table B.7 below.

Data availability

Model 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/.

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 (+$p < 0.1$, *$p < .05$, **$p < .01$, ***$p < .001$).

	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 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 spilled.	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. http://dx.doi.org/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. http://dx.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. http://dx.doi.org/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*, URL https://aclanthology.org/2017.lilt-15.3/.

Bhattacharya, D., & van Schijndel, M. (2020). 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, http://dx.doi.org/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: Vol. 3*, (pp. 20–30). University of Massachusetts Amherst Libraries.

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 M. Halle, S. R. Anderson, & P. Kiparsky (Eds.), *A festschrift for Morris Halle* (pp. 232–286). New York: Holt, Rinehart and Winston.

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., & Lasnik, H. (1977). Filters and control. *Linguistic Inquiry*, *8*, 425–504, URL 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. http://dx.doi.org/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 E. Engdahl, & E. Ejerhed (Eds.), *Readings on unbounded dependencies in Scandinavian languages* (pp. 77–98). Stockholm: Almquist & Wiksell.

Christiansen, M. H., & Chater, N. (2016). *Creating language: Integrating evolution, acquisition, and processing*. The 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 R. Kempson, N. Asher, & T. Fernando (Eds.), *Philosophy of Linguistics* (pp. 445—-475). Elsevier, http://dx.doi.org/10.1016/b978-0-444-51747-0.50013-5.

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. R. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives* (pp. 94–128). Cambridge University Press, http://dx.doi.org/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. http://dx.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. http://dx.doi.org/10.7275/7fd4-fw49.

Frank, M. C. (2023). Bridging the data gap between children and large language models. http://dx.doi.org/10.31234/osf.io/qzbgx, PsyArXiv Preprints.

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. http://dx.doi.org/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). http://dx.doi.org/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). http://dx.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). http://dx.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* (pp. 179–219). Springer International Publishing, http://dx.doi.org/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 L. Barak, & M. Alikhani (Eds.), *Proceedings of the 28th Conference on Computational Natural Language Learning* (pp. 269–279). Miami, FL, USA: Association for Computational Linguistics, http://dx.doi.org/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* (pp. 1725–1744). Association for Computational Linguistics, http://dx.doi.org/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. http://dx.doi.org/10.5964/bioling.13153.

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). URL 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, Vol. 6* (pp. 175–185). http://dx.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*, Article e13501. http://dx.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. (pp. 1–44). http://dx.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*, 195–220. http://dx.doi.org/10.3390/languages7030197.

Kodner, J., & Gupta, N. (2020). Overestimation of syntactic representation in neural language models. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1757–1762). Association for Computational Linguistics, http://dx.doi.org/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, http://dx.doi.org/10.48550/arXiv.2308.03228.

Kush, D., & Dahl, A. (2020). L2 transfer of L1 island-insensitivity: The case of Norwegian. *Second Language Research*, 1–32. http://dx.doi.org/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.

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. http://dx.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*, 393–420. http://dx.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*, 1–50. http://dx.doi.org/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.

Lan, N., Chemla, E., & Katzir, R. (2024). Bridging the empirical-theoretical gap in neural network formal language learning using minimum description length. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Long papers: Vol. 1, Proceedings of the 62nd Annual Meeting of the Association for Computational linguistics* (pp. 13198–13210). Bangkok, Thailand: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2024.acl-long.713.

Lan, N., Chemla, E., & Katzir, R. (2024). Large language models and the argument from the poverty of the stimulus. *Linguistic Inquiry*, 1–56. http://dx.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 I. Roberts (Ed.), *The Oxford Handbook of Universal Grammar* (pp. 220–248). Oxford University Press, http://dx.doi.org/10.1093/oxfordhb/9780199573776.013.10.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177. http://dx.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. http://dx.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.

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.

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).

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*, http://dx.doi.org/10.16995/glossa.11164.

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, http://dx.doi.org/10.48550/arXiv.2405.15750.

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*, 1–34. http://dx.doi.org/10.1017/S0305000922000514.

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

Pearl, L., & Sprouse, J. (2013). 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.

Perlmutter, D. M. (1971). *Deep and surface structure constraints in syntax*. Holt, Rinehart & Winston.

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

Phillips, C. (2013). On the nature of island constraints II: Language learning and innateness. In *Experimental syntax and island effects* (pp. 132–158). Cambridge University Press, http://dx.doi.org/10.1017/CBO9781139035309.007.

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).

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 M. Bansal, & A. Villavicencio (Eds.), *Proceedings of the 23rd Conference on Computational Natural Language Learning* (pp. 66–76). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/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). URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

R Core Team (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, URL 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.

Ritchart, A., Goodall, G., & Garellek, M. (2016). Prosody and the that-trace effect: An experimental study. In *33rd West Coast Conference on Formal Linguistics* (pp. 320–328). Cascadilla Proceedings Project.

Rizzi, L. (1982). Violations of the WH island constraint in Italian and the subjacency condition. In *Issues in Italian syntax* (pp. 49–76). Dordrecht.

Ross, J. R. (1967). *Constraints on variables in syntax* (Ph.D. thesis), MIT, URL 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* (pp. 5831–5837). Hong Kong, China: Association for Computational Linguistics, http://dx.doi.org/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*, Article e2307876121. http://dx.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. http://dx.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. http://dx.doi.org/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. http://dx.doi.org/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. In *Conference on Neural Information Processing Systems: Vol. 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 J.-L. Mendivil-Giro (Ed.), *Artificial Knowledge of Language*. Vernon Press, URL 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. G., Futrell, R., & Levy, R. (2023). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–44. http://dx.doi.org/10.1162/ling_a_00491.

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

Wilcox, E., Levy, R., & Futrell, R. (2019). 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). http://dx.doi.org/10.18653/v1/W18-5423.

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.), *Long papers: Vol. 1*, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 9370–9393). Toronto, Canada: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2023.acl-long.521.