Learning constraints on *wh*-dependencies by learning how to efficiently represent *wh*-dependencies: A developmental modeling investigation with Fragment Grammars

Niels Dickson **Lisa Pearl** **Richard Futrell**
Department of Language Science
University of California, Irvine
{nielsd, lpearl, rfutrell}@uci.edu

1 Introduction

It's hotly contested how children learn constraints on the allowed forms in their language, such as constraints on *wh*-dependencies (these constraints are sometimes called *syntactic islands*: Chomsky 1973; Pearl and Sprouse 2013). When learning this knowledge, a prerequisite is knowing how to represent *wh*-dependencies – constraints can then be hypothesized over these dependency representations. Previous work (Pearl and Sprouse, 2013; Liu et al., 2019) explained disparate sets of syntactic island constraints by assuming different *wh*-dependency representations, without a unified dependency representation capturing all these constraints. Here, we implement a modeled learner attempting to learn a Fragment Grammar (**FG**) representation (O'Donnell et al., 2011; O'Donnell, 2015) of *wh*-dependencies—a representation comprised of potentially different-sized fragments that combine to form full dependencies—that best accounts for the input while being as compact as possible. In particular, FG implements a theory of efficiency that balances the size of the fragments in the resulting grammar while also maximizing the probability of the dependency structures comprised of these fragments. So, when deciding on the fragments to represent from linguistic input, a learner can choose between smaller fragments of the input that may be reused often in different contexts and larger fragments that can be accessed without building up the structure from smaller pieces. The resulting fragment-based *wh*-dependency representation can then be used to generate any *wh*-dependency's probability on the basis of its fragments, and so predict acceptability patterns for stimuli sets that reveal syntactic island knowledge. We find that the identified FG, learned from a realistic sample of *wh*-dependencies from English-learning children's input, can generate the attested acceptability judgment patterns for all syntactic islands previously investigated, highlighting how implicit knowledge of *wh*-dependency constraints can emerge from trying to learn to efficiently represent *wh*-dependencies more generally. We additionally compare the FG representation's performance against baselines inspired by previous proposals, finding that one baseline also yields equivalent performance. We discuss how this baseline is similar to and different from the FG representation.

2 *Wh*-dependency representation

We assume *wh*-dependencies are represented as sequences of phrase structure nodes that indicate the path from the gap to the *wh*-word (Pearl and Sprouse, 2013) (1a)-(1b). However, it's unknown whether the phrasal categories (e.g., CP, VP) in this representation need to be lexically subcategorized. For instance, does the dependency path for a *wh*-dependency with *claim* need to include that the verb is *claim* (1d) or not (1e)?

(1) What did Lily claim that Jack forgot?

a. What did [$_{IP}$ Lily [$_{VP}$ claim [$_{CP}$ that [$_{IP}$ Jack [$_{VP}$ forgot $\_\_{}_{what}$]]]]]?

b. phrase-structure nodes in syntactic path: IP-VP-CP-IP-VP

c. lexical information for those nodes: IP=*past*, VP=*claim*, CP=*that*, IP=*past*, VP=*forget*

d. possible representations with lexically-subcategorized VP *claim*: IP-VP$_{claim}$-CP-IP-VP, IP$_{past}$-VP$_{claim}$-CP-IP$_{past}$-VP, IP-VP$_{claim}$-CP$_{that}$-IP-VP$_{forget}$, ...

e. possible representations without lexically-subcategorized VP *claim*: IP-VP-CP-IP-VP, IP-VP-CP$_{that}$-IP-VP, IP$_{past}$-VP-CP$_{that}$-IP$_{past}$-VP$_{forget}$, ...

Proceedings of the Society for Computation in Linguistics (SCiL) 2022, pages 220-224.
Held on-line February 7-9, 2022

Figure 1 illustrates the consequences of lexical subcategorization and the related balance of fragment size mentioned earlier, showing two extremes of how to represent the example dependency path from (1). The leftmost representation uses minimal-sized fragments (phrasal-only like IP-VP, and lexicalized like IP$_{past}$) that may be reused often because they can appear in many different dependencies. This representation has no lexical subcategorization because the lexical information is separate from the phrasal structure. The rightmost representation uses a maximal-sized fragment (representing the entire dependency with both its phrasal structure and lexical pieces) that will only be reused if this exact dependency occurs. This representation has complete lexical subcategorization because all the lexical information is included in this phrase structure fragment. In terms of maximizing the probability of the dependency, each extreme has its drawbacks: the representation relying on minimal-sized fragments requires combining many individual fragments, which can lead to a lower probability even if the individual fragments have higher probabilities; the representation relying on the maximal-sized fragment likely has a fairly low probability unless this particular dependency happens to occur very frequently (and even if it does, this won't be true for all dependencies). To maximize the probability of a dependency in general, a better approach is to find some intermediate representation, such as the middle one in Figure 1, that involves some larger phrasal fragments incorporating lexical subcategorization (e.g., IP$_{past}$-VP), as well as some lexical-only fragments (e.g., VP$_{forget}$). In this example intermediate representation, there is thus a tradeoff between larger fragments that don't have to be built every time from smaller fragments (e.g., IP$_{past}$-VP from IP$_{past}$ and IP-VP) and smaller, more frequently-reused fragments (e.g., VP$_{forget}$). Of course, there are many possible intermediate representations, and the goal for a learner is to identify the best one that maximizes this tradeoff and so yields high probabilities collectively for the dependencies in the input.

3 Previous representation proposals

Previous developmental modeling work by Pearl and Sprouse (2013) predicted attested adult judgment patterns for 4 islands (Complex NP, Subject, Adjunct, Whether)—see Figure 2a—by assuming only CPs were lexically subcategorized (i.e., only the lexical information of CPs was included with the phrasal structure). Previous empirical work by Liu et al. (2019) predicted attested judgment patterns for 14 bridge (e.g, *say*), factive (e.g., *know*), and manner-of-speaking (e.g., *whisper*) verbs—see Figure 2c—in terms of the lexical frequency of the main-clause VP. While Liu et al. didn't explicitly propose a theory of representation, their results are compatible with a representation that lexically subcategorizes main-clause VPs (i.e., only the lexical information of the main verb is included with the phrasal structure). Yet, these are only two of many possible types of hypotheses for how the phrasal structure of *wh*-dependencies could be represented (i.e., different intermediate representations). Using an FG, we can explore the entire hypothesis space that investigates not only which lexical information should be included (e.g., CPs or main VPs), but also what size fragments are the most efficient for the phrasal structure of the dependency to be built from. Importantly, instead of telling the learner beforehand what phrase structure nodes are lexicalized and what size fragments to use, the learner using FGs infers both on the basis of its input.

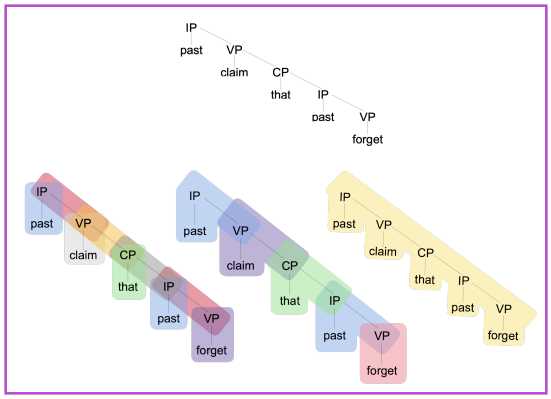
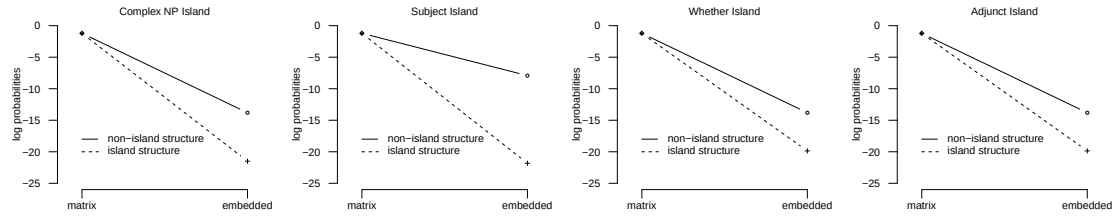


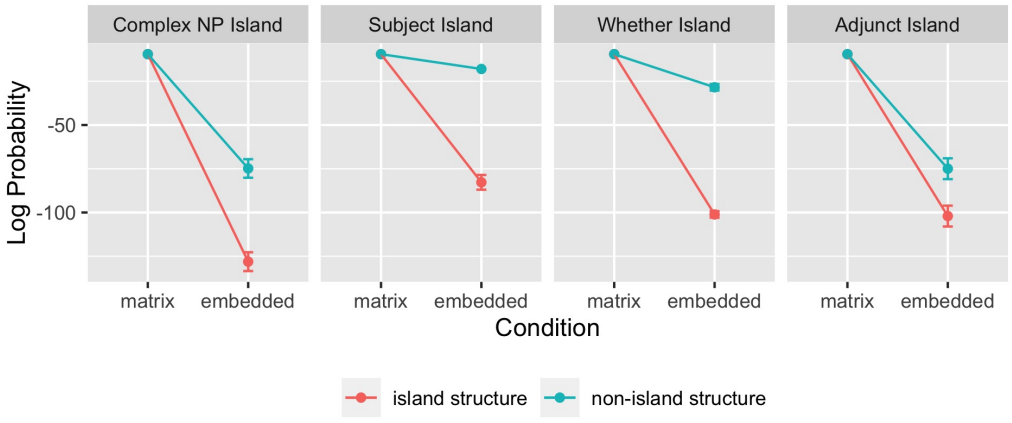
Figure 1: Example *wh*-dependency path as a syntactic tree and possible ways to build it from fragments.

4 Learning efficient representations that underlie *wh*-dependency constraints

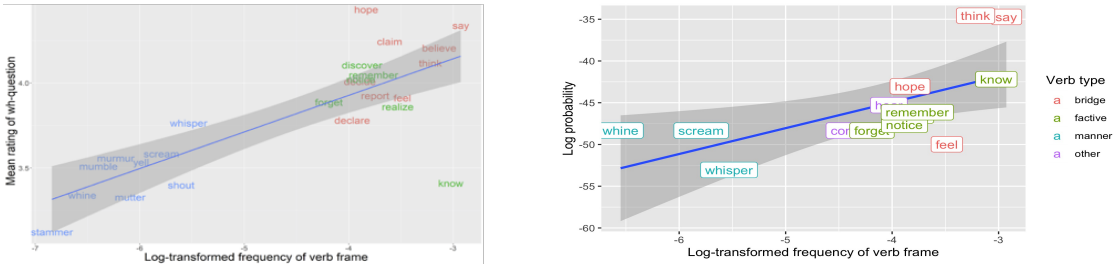
We implement a computational-level modeled learner that attempts to identify an FG encoding the most efficient dependency path representation. The model uses Bayesian inference to identify the best representation. In particular, the modeled learner uses a Metropolis-Hastings-based inference algorithm to find the set of fragments that best explains the input, by yielding a high probability for the dependencies in the input. To identify this FG representation, the modeled learner uses the Metropolis-



(a) From Pearl & Sprouse (2013).



(b) The same superadditive pattern.



(c) Left: From Liu et al. (2019). Right: The same positive correlation.

Figure 2: The top row shows (a) the modeled judgment patterns, matching empirical judgment patterns, from Pearl & Sprouse (2013), and (b) the judgment patterns (log probabilities) generated by the Fragment Grammar (FG) identified by the modeled learner, given realistic samples of child-directed speech. The bottom row shows (c) left: the empirical judgment patterns from Liu et al. (2019), and right: the judgment patterns generated by the FG.

Hastings algorithm to iteratively resample a potential FG representation for each item in the input and then accepts or rejects the representation to increase the probability of the input data.

To approximate the *wh*-dependency input that children learn from, we collected 12,704 *wh*-dependencies from the CHILDES Treebank (Pearl and Sprouse, 2013) and extracted the dependency path from each.[1] We then estimated the counts of the dependencies that children would encounter by four years old, when some syntactic island knowledge seems to be present (De Villiers et al., 2008).[2] From this input, the modeled learner infers the best fragments for the *wh*-dependencies in its input, which may or may not include lexically-subcategorized phrasal structures for any given fragment. This model allows us to explore all the possibilities of lexically subcategorizing different phrasal categories as opposed to implementing a particular hypothesis (i.e. main verbs are always lexicalized or CPs are always lexicalized).

We find that the learned FG dependency representation can be used to correctly generate all previously-attested acceptability judgment patterns

[1] See Supplemental Section A.1 for more details.

[2] We drew on estimations by Bates and Pearl (2021) that consider waking hours, utterances per hour, and *wh*-dependency frequency in children's input between 20 months (when *wh*-dependencies are reliably processed) and 4 years.

(Figure 2b and d).[3] Notably, the FG representation's fragments lexically subcategorize phrasal structures only for some more-frequent items (e.g., VP_{think}, CP_{that}, VP_{say}-CP). This means the modeled learner automatically determined the best frequency threshold for lexically subcategorizing each individual phrase structure type, due to the goal of efficient representation.

5 Comparison representations

We compared the FG representation's performance against several trigram-based baseline representations (2), all of which used the same input as the FG model.[4] We chose trigram-based representations, as n-grams are common representations in language modeling (see Manning and Schutze 1999 for a review), and trigram-based representations have been used in prior successful models that predict adult judgement patterns of *wh*-dependencies (Pearl and Sprouse, 2013). A trigram-based representation also can (i) be paired with a straightforward learning algorithm (e.g., tracking frequencies of the trigrams in the input), and (ii) can transparently reflect different proposals for lexical subcategorization, as in (2).

(2) Baseline trigam representations

a. no-lexicalization: phrase labels only, e.g., "IP-VP-CP"

b. fully-lexicalized: subcategorized phrase labels, e.g., "IP_{past}-VP_{claim}- CP_{that}"

c. CP-lexicalized (from Pearl and Sprouse 2013): only CP is subcategorized, e.g., "IP-VP-CP_{that}"

d. main-V-lexicalized (in line with Liu et al. 2019): only main V is subcategorized, e.g., "IP-VP_{claim}-CP"

We selected the no-lexicalization and the fully-lexicalized representations as the two extremes of our hypothesis space; we can include no lexical information or all the lexical information for phrase structure nodes in a trigram-based dependency representation. The remaining two representations each implement a hypothesis about what lexical information should be included in the phrasal structure, inspired by previous work: the CP-lexicalized representation from Pearl and Sprouse (2013), and the main-V-lexicalized representation from Liu et al. (2019).

Most baselines failed to capture the full range of acceptability judgment patterns: the no-lexicalized failed to capture Adjunct and Whether islands, as well as the verb frequency effect; the fully-lexicalized failed to capture Adjunct islands; and the CP-lexicalized failed to capture the verb frequency effect. However, the main-V-lexicalized did capture all the acceptability patterns. We note that the FG representation also lexicalized main verbs (though only those that were more frequent), and so has this in common with the main-V-lexicalized baseline (which lexicalized all main verbs, irrespective of frequency). We note that one advantage of the inferred FG representation over the main-V-lexicalized representation is that the FG representation was automatically learned – including which parts are lexicalized and how large the pieces are that comprise a dependency path – rather than needing to be specified beforehand, as the trigram-based main-V-lexicalized baseline was.

6 Conclusion

Here we have explored how children could learn constraints on English *wh*-dependencies by focusing their learning efforts on how to efficiently represent *wh*-dependencies, rather than trying to explicitly learn the constraints. The specific approach we explored involved a modeled learner attempting to identify the best Fragment Grammar (FG) for efficiently representing the *wh*-dependencies encountered in English child-directed speech. The FG representation allowed the modeled learner to generate all the acceptability judgement patterns previously attested to reflect knowledge of different constraints on *wh*-dependencies, known as syntactic islands. Because the modeled learner learned from input that four-year-olds would encounter, one testable prediction that future behavioral work can investigate is that four-year-olds should in fact have acquired all the syntactic knowledge assessed via the acceptability judgment patterns used here if four-year-olds are in fact using an FG representation. Additionally, future work can investigate predictions for other *wh*-dependency constraints known to be acquired by children around age four (De Villiers et al., 2008), comparing the FG representation against other representational possibilities, such as the main-V-lexicalized baseline.

[3] See Supplemental Section A.4 for details.

[4] These baselines additionally had a "Start" and "End" symbol in their dependency paths to ensure each dependency created at least one trigram. For instance, a main clause subject dependency like "What happened?" would be represented with the trigram "Start-IP-End".

References

Alandi Bates and Lisa Pearl. 2021. When do input differences matter? using developmental computational modeling to assess input quality for syntactic islands across socio-economic status.

Noam Chomsky. 1973. Conditions on transformations. In S. Anderson and P. Kiparsky, editors, *A Festschrift for Morris Halle*, pages 237–286. Holt, Rinehart, and Winston, New York.

Jill De Villiers, Thomas Roeper, Linda Bland-Stewart, and Barbara Pearson. 2008. Answering hard questions: Wh-movement across dialects and disorder. *Applied Psycholinguistics*, 29(1):67–103.

Yingtong Liu, Rachel Ryskin, Richard Futrell, and Edward Gibson. 2019. Verb frequency explains the unacceptability of factive and manner-of-speaking islands in english. In *CogSci*, pages 685–691.

Christopher Manning and Hinrich Schutze. 1999. *Foundations of statistical natural language processing*. MIT press.

Timothy O'Donnell, Jesse Snedeker, Joshua Tenenbaum, and Noah Goodman. 2011. Productivity and reuse in language. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33.

Timothy J O'Donnell. 2015. *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Lisa Pearl and Jon Sprouse. 2013. Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, 20:19–64.

A Supplemental Material

A.1 Preprocessing from CHILDES

Using NLTK and Python, we extracted the *wh*-dependency trees from the CHILDES Treebank. Using the trace annotations in the corpora, we extracted the path from the gap position to the *wh*-word, including the phrase label (e.g., VP) and its lexical child (e.g., *think*) in the resulting sequences. When a VP followed the IP in a dependency path, tense was added as the lexical child of IP nodes (e.g., *thought* would yield IP_{past}-VP_{think}).

A.2 Preprocessing

Preprocessing for the FG grammar input: We created the tree structures of the dependency paths from these sequences in a form that the FG learner can process (e.g. the *wh*-dependency "What are you eating?" would be encoded as "((IP (LEX present) (VP (LEX eat))))"). We note also that IP-only dependencies like "*What happened?*" did not have the IP lexicalized with tense (e.g., the FG input representation would be (IP null)).

Preprocessing for the trigram baseline models: The baseline trigram models took the final dependency path, extracted from CHILDES and preprocessed to include tense (e.g., IP_{past}-VP_{think}), and extracted appropriate trigrams, depending on the baseline. We included a "Start" and "End" symbol in our dependency paths for the baselines in order for all paths to be one trigram at a minimum. This allowed IP-only dependencies to be handled by trigram-based models (i.e., the trigram would be Start-IP-End).

A.3 Inference of the best FG grammar

The inference algorithm used to identify the best FG was implemented using code provided by Tim O'Donnell. We used the default parameter values: pitman-Yor (PY) *a* set to 0 and PY *b* set to 1; sticky concentration parameter set to 1 and sticky distribution parameter set to 0.5; the Dirichlet-multinomial pseudo-counts (pi parameter) were set to 1; the model performed 1000 sweeps.

A.4 Generating predictions of acceptability using the FG representation

When generating predictions for *wh*-dependencies, based on the FG representation, we extracted the same form of the *wh*-dependency path from the Pearl and Sprouse (2013) and Liu et al. (2019) stimuli. Due to the design of the code, structures that required rules the FG did not hypothesize would yield no output (i.e., cause a code crash). To circumvent this and be able to generate predictions for structures like those that cross syntactic islands, we needed to add all possible phrase rules to the FG representation. So, we added all possible rules (in the form "Label1 – Label2 Label3", where a Label was a phrase structure node like IP or PP) that the FG representation did not create through inference. We then gave these rules "counts" of 0.5 (as opposed to any seen structure having a count of at least 1) and re-normalized the log probabilities of all rules.