ACQUIRING SYNTACTIC VARIATION: REGULARIZATION IN WH-QUESTION PRODUCTION

by
An Nguyen

A dissertation submitted to Johns Hopkins University in conformity with the

requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
March 2023

© 2023 An Nguyen
All rights reserved

Abstract

Children are often exposed to language-internal variation. Studying the acquisition of variation allows us to understand more about children's ability to acquire probabilistic input, their preferences at choice points, and factors contributing to such preference. Using wh-variation as a case study, this dissertation explores the acquisition of syntactic variation through corpus analyses, behavioral experiments, and computational simulation.

In English and some other languages (e.g., French, Brazilian Portuguese, etc.), information-seeking wh-questions allow for at least two variants: a wh-in-situ variant and a fronted-wh variant. How do English-speaking children acquire wh-variation, and what factors condition their course of acquisition? Experimental results show that 3-to-5 year-old children regularize to fronted wh-questions in their production even in contexts that allow for both variants to be used interchangeably. Based on the characteristics of the variants, two factors are identified to potentially contribute to the preference for fronted wh-questions: frequency and discourse restrictions. Two artificial language learning (ALL) experiments are then conducted so that the effect of discourse can be studied separately from frequency. The results show that learners prefer the variant with fewer or no discourse restrictions (i.e., the fronted-wh variant) when frequency is controlled. Thus, regularization in language acquisition is conditioned by both domain-general factors, such as frequency, and language-specific factors, such as discourse markedness.

The dissertation also looks into the motivation for regularization. One prominent hypothesis is that regularization serves as a means to reduce the cognitive burden associated with learning multiple variants at once. Instead of mastering all the variants, learners can simplify the learning process and minimize their chance of violating a

constraint by producing the dominant variant. This work provides additional evidence for the hypothesis in three ways. First, we replicate the findings that tasks that are more cognitively taxing induce more regularization. Second, we present new evidence that participants with a lower composite working memory score tend to have a higher regularization rate. Third, we provide a computational simulation showing that regularization behavior only happens when an intake limit (reflecting limited working memory capacity) and a parsimony bias to reduce the cognitive burden are incorporated in the model.

Readers: Géraldine Legendre (advisor), Kyle Rawlins, Lisa Feigenson, Steven Gross, Jennifer Culbertson (external)

Acknowledgment

I still remember the excitement I felt six years ago, back in 2017, when I visited Johns Hopkins as a prospective graduate student. After the faculty's research presentations, I excitedly talked to Géraldine about all the potential research projects I felt I could do at Hopkins. Of course, none of the projects discussed on that day was even slightly related to my current work on syntactic variation -- as people often say, you never really know where your research may take you. The past five years have been an incredible journey, and I wouldn't make it to where I am today without the support of numerous people.

I must begin by thanking my advisor, Géraldine Legendre. She is the best advisor I could ever ask for. She is always there for me when I need her, yet also provides me the trust and autonomy to explore my interests and grow as an independent researcher. Working with Géraldine is smooth and easy, as we share many similar traits. I have felt supported every step of the way in this journey thanks to Géraldine.

I am grateful to other members of the Cognitive Science department. I thank Kyle Rawlins for all the discussions about the pragmatics of wh-questions as well as helpful feedback on running acceptability judgment studies. I also wish to thank Paul Smolensky, who has always been so kind and considerate that every brief interaction with him brightens my outlook.

I gratefully acknowledge the members of my committee who took time out of their busy schedules to read and discuss this work. I also thank Jenny Culbertson for all her valuable feedback on the design of my artificial language learning tasks and the statistical analysis of my experiments.

I would like to extend my thanks to the members of the Language Acquisition Lab, Jane Lutken, Renee Cong, and my two research assistants, Will Howe and Lily Zhu. As my senior, Jane has provided me with guidance and helped me navigate graduate school. Renee, Will, and Lily have assisted me with identifying and collecting wh-questions in English and French child-directed speech. Special thanks go to Will, who played an excellent role of the "storyteller" in my first behavioral experiment with young children.

The past five years have been special largely thanks to the presence of my fellow graduate students. Thanks for all the late-night talks, get-togethers, game nights etc. that keep me sane! I thank Natalia for being a great listener to my random ranting. I also want to thank my officemates, Suhas and Jane Li, for always coming to my practice talks and putting up with me distracting them from doing actual work in the office. Special shout out to the members of my cohort, Rennie and Suhas, who have gone through so many things with me in my first two years. When I think about the first few classes I took at Hopkins, I will always remember our hours-long discussion about the semantics of pizza.

Finally, thank you, Mike, for sticking with me through all the ups and downs. I'm forever grateful to have you in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Variation can be found across natural languages as well as within a language, posing a series of challenges to young learners. Cross-linguistic variation raises the question of how children learn the universal versus specific grammatical properties of their native language(s). One prominent explanation relies on an innate system of Universal Grammar, or UG principles (Chomsky, 1981), including some that are active on a language-specific basis only. Variation across languages can be viewed as a result of the differences in UG parameter settings, and thus, language acquisition is the task of selecting from a range of UG parameter values based on language input (Yang, 2010). For example, English-speaking children need to learn that English has a [-] value for the null subject parameter, as it is ungrammatical to drop the subject in a finite clause (e.g., *~~*She~~ *Is eating)*. In contrast, Italian-speaking children will learn that the language has a [+] value for the null subject parameter, based on grammatical sentences lacking an overt subject in the input (e.g., *Parla Italiano* – (*she) speaks Italian)*. The parameter theory of cross-linguistic variation was prominent in the 1980s and 1990s and remains one of the most influential theories of acquisition. An alternative universalist model is provided by Optimality Theory (Prince & Smolensky, 1993/2004), which proposes that variation in languages arises instead from differences in rankings of violable universal constraints. The acquisition process then involves learning the language-particular rankings of relevant constraints based on language input (Legendre, Hagstrom, Vainikka & Todorova, 2002).

Language-internal variation or alternative grammatical forms for a given meaning presents a different problem for acquisition. To a certain extent, language-internal variation offers learners options to choose from, and this allows us to re-examine whether children's

production closely reflects or diverges from the input they receive. Mismatches between children's production and their input are of particular interest to researchers, as they show that language acquisition is more than mere imitation. Furthermore, such choice points also allow researchers to investigate learners' preferences and factors that influence their acquisition path: when there are multiple ways to express the same idea, which variant is learned first, what drives the preference of the "chosen" variant, and why?

In that spirit, this dissertation aims to further the understanding of children's acquisition of language-internal variation. In the next section (1.1.), I will distinguish two "types" of language-internal variation, before focusing this dissertation on one type, namely, variation that applies at both the rule and the construction level (these terms are mere pre-theoretic labels for syntactic phenomena such as passive, wh-questions, etc., respectively). Section 1.2. discusses the acquisition of variation in general, while section 1.3. presents a discussion of wh-question variation as a case study. Finally, in section 1.4., I establish the research questions and the main claims of this work.

1.1. Language-internal variation

1.1.1. Two types of variation

There are two types of language-internal variation: stylistic variation across speakers, and language-internal variation. The former, not considered in this dissertation, happens due to extralinguistic factors, such as gender, age, and social background. This type of variation is often studied by sociolinguists, who investigate how contextual and social factors can contribute to speakers' choice of variant (e.g., Barbu et al., 2013; Chevrot & Foulkes, 2013; Lacoste & Green, 2016).

My focus in this dissertation instead is on language-internal variation, which can be found at two different levels. In one case, variation may apply at the rule level but not at the construction level. An example of this is the English past tense system. There are many rules that can be applied to a verb to express its past tense, including but not limited to:

- Change the vowel from -i to -a: ring -> rang, sing -> sang, spring -> sprang
- Change the vowel from -i to -u: swing -> swung, fling -> flung, cling -> clung
- Change the vowel from -o to -e: blow -> blew, grow -> grew, know -> knew
- Add -ed: talk -> talked, laugh -> laughed, lift -> lifted

Aside from about 20 exceptions (a given verb allowing both a regular and an irregular past tense form, e.g., dived/dove, dreamed/dreamt), each construction/verb only works with one rule, resulting in a single grammatical form. Even though the last rule (add -ed) is the most productive rule (as it is statistically predominant compared to the others), it is ungrammatical to say "flinged" or "growed". In other words, it is (typically) *ungrammatical* to alternate between rule variants within a construction to form a past tense.

In the other case, variation may apply at both the rule and the construction level, e.g., French wh-questions. French allows for considerable variation in wh-questions, with some variation associated with a register difference (Shlonky, 2012). Among the multiple variants of the wh-question-forming rule, there are the in-situ rule (the wh-phrase remains in its original position, such as in example (1)) and the wh-fronting rule (the wh-phrase is moved overtly to the beginning of the sentence, such as in (1b)). Putting pragmatic and discourse factors aside for now (I return to these in Chapter 2), it is generally *grammatical*

to use an in-situ wh-question in place of a fronted wh-question and vice-versa. In this case, within a construction, it is possible to alternate between the two rule variants.

(1) a. Tu as quitté qui? (in-situ object wh)
you have left who
Who have you left?
b. Qui tu as quitté? (fronted object wh without inversion)

There is also a "hybrid" type, exemplified by the English dative system (one construction, two rules, with restrictions). To express a double object dative construction, speakers may use the ditransitive rule (V -> NP NP) or a prepositional dative rule (V -> NP PP). In many cases, constructions based on both rules are grammatical and can be used alternatively (e.g., "Mary gives a book to John" and "Mary gives John a book"). However, there are also exceptions where only one rule can be used. For example, one can say "Mary donated a book to the library" but not "Mary donated the library a book".

From an acquisition perspective, these two types of language-internal variation could result in rather different learning patterns. In the rule-only case, learners need to know not only the variants of the rule but also when they can apply the rule. Research on this type of variation often focuses on how children learn which rule applies when. For example, Prasada & Pinker (1993) propose a dual-route model for English past-tense inflection, in which children learn regular inflection via a rule (add *-ed*) and irregular inflection (e.g., "exceptions") through lexical memory. Yang's Tolerance Principle (2016) spells out a more formal proposal to the problem by calculating the number of exceptions relative to the number of items that can be generated by a rule, with too many exceptions leading to a tipping point when forming a productive rule is no longer efficient for learners. On the other hand, in the rule-and-construction case, learners have more freedom about

when to use which rule. They may even get away with learning only one rule or a subset of the rules, for example, one child may use mostly French wh-in-situ in her production while another may use mostly fronted wh-questions in his, and both produce grammatical questions. Thus, research on this type of variation often studies whether (and how much of) children's production is similar to their input, and if not, what factors drive the divergence. For example, studies on French-speaking children's wh-questions have focused on comparing the production of wh-in-situ versus fronted questions (e.g., Crisma, 1992; Zuckerman & Hulk, 2001; Hamann, 2006; Strik, 2007; Gotowski, 2017), leading to a prominent hypothesis in acquisition of syntax such as the Derivational Complexity Hypothesis (DCH) (Jakubowicz, 2005, 2011), which claims that children prefer learning structures that are syntactically simpler. The DCH hypothesis is discussed in more detail in Chapter 2: A case study of variation: wh-question.

It is possible that early on, children treat the two types of variation similarly despite their differences. Studies have reported that at an early point in the acquisition course, children show a tendency to over-rely on a single variant of rule in both types (e.g., Kuczaj (1977) on the over-generalization of the English past tense *add -ed* rule for rule-only variation; Hendricks, Miller, & Jackson (2018) on the overregularization of Fering gender markings for rule-and-construction variation, which is discussed in section 0.). This results in a divergence from the input in both cases, but since one divergence is ungrammatical (the rule-only case) while the other is grammatical (the rule-and-construction case), they are rarely discussed together. Generally speaking, "overgeneralization" in rule-only variation (i.e., when children incorrectly extend a variant of the rules to cases that do not apply) and "overregularization" in rule-and-construction variation (i.e., when children

reduce variation in the input by using mainly one or a subset of variants) share at least the same tendency towards simplicity, that is, children prefer working with a single uniform rule. While this dissertation mostly focuses on variation at the rule and construction level, in particular wh-question variation, future work can explore the relationship between the acquisition processes of the two variation types.

1.1.2. The problem of True optionality

From this section onwards, for simplicity, I will refer to language-internal syntactic variation at the rule-and-construction level as "variation". In other words, the term "variation" in this dissertation is used to mean "having more than one *grammatical* way to express a meaning within a language". Such variation is reported in many languages, for example:

(2) *Finite auxiliary placement in Afrikaans* (Biberauer & Richards, 2006)
a. Ek weet dat sy dikwels Chopin gespeel het.
I know that she often Chopin played has.
'I know that she has often played Chopin.'
b. Ek weet dat sy het dikwels Chopin gespeel.
I know that she has often Chopin played

(3) *Stylistic fronting in Icelandic* (Poole, 1996)
a. Petta er versta bok sem hefur verid skrifud.
this is the-worst book that has been written
'This is the worst book that has been written.'
b. Petta er versta bok sem skrifud hefur verid
this is the-worst book that written has been

In many cases, it is completely acceptable to use one variant in place of another. However, the general assumption is that there is most likely a subtle difference between the variants in terms of pragmatics. This can be expressed through a difference in sentiment, inference, or certain contexts in which one variant seems more acceptable than

the other. For example, Bolinger (1977) claims that while the two examples below express the same truth conditions, (4b) seems to imply that the lines are in, while this is not necessarily true for (4a). Similarly, Clark & MacWhinney (1987) claims that listeners can infer from (5a) that Rob has learned some French, while such an outcome is unspecified in (5b).

(4) a. They hauled in the lines.
b. They hauled the lines in.

(5) a. Jan taught Rob French.
b. Jan taught French to Rob.

Modern syntactic theory (such as the Minimalist Program (Chomsky, 1995) and Optimality Theory (Legendre et al., 2001)) rejects the idea of true optionality through a general notion of "Economy", which calls for simplicity and a tendency to reduce computation to a minimum in grammar. The principle of Economy states that syntactic operations are necessarily motivated. Syntactic variation only happens if such variation contributes to interface interpretations (MP) or faithfulness to semantics/pragmatics (OT). In other words, there is no *true* optionality. Zuckerman (2001) further defines true optionality as follows:

(i) Optionality: S and S' are optional structures if and only if:
1. a specific numeration set N yields both S and S'
2. both S and S' converge at the interface
3. the derivations leading to S and S' yield identical LF representations.

While Economy has been challenged by examples of two structures claiming to have no interpretative differences (e.g., Shlonsky, 1997; Biberauer & Richards, 2006, Suzuki, 2012), Zuckerman argues that all such structures violate at least one of the requirements. Due to the limited scope of this dissertation, I do not consider this debate

further, and choose to follow Chomsky (1995) and others in endorsing the view that true optionality does not exist.

Assuming no true optionality means that even when all the variants are grammatical and refer to the same truth-conditional meaning, there are constraints at the interface (e.g., discourse, style, prosody) that make one variant more felicitous or preferable. Thus, to master a form that allows for variation, the child needs to learn not only the syntax-semantics mapping of the form but also the constraints on the variants. This is certainly not an easy task for young children, especially considering that the contrast between two variants can be rather subtle, as shown in examples (4) and (5). Moreover, the contrast between variants can stem from the context they are in, as one variant is more felicitous than the other in certain contexts. Children may not be exposed to sufficient evidence to infer the contrast between two variants, since child-directed speech can be quite limited and specific in terms of contexts. Thus, it is possible that early on, children may mistreat some cases of variation as if they express true optionality.

Zuckerman (2001) argues that children avoid optionality because it is problematic for language learnability: if a child assumes that his language allows for an overly general optional rule, no amount of positive evidence can lead to the revision of this assumption. However, the problem Zuckerman describes only happens if optionality is children's initial or default hypothesis. A similar argument can be applied to the acquisition of pro-drop. The pro-drop parameter setting allows for input both with and without overt subjects, while the non pro-drop setting only allows for input with overt subjects. If children's default setting were pro-drop, they would never encounter evidence contradicting this, and as a result would never be able to learn a non pro-drop generalization (according to the Subset

Principle (Berwick, 1985), which states that learners "only assume a grammar sufficient to generate the sentences they hear"). To handle this, I do not reject the notion of pro-drop or the ability to learn pro-drop, but simply propose that the default setting is non pro-drop. Similarly, to avoid the problem in Zuckerman's argument, it's possible to assume that children's initial hypothesis does not contain optionality, but nevertheless optionality can be learned or inferred through input.

Allowing for the possibility of children mis-analyzing variation as true optionality does not conflict with the theoretical claim about the (non)existence of true optionality in language. Such flexibility is not unreasonable given the limitations children face (in terms of input and cognitive ability), and it also enables us to entertain a larger range of hypotheses to account for children's behavior. In particular, children who assume two variants as being fully interchangeable may use them differently from their parents. Even when children's production closely matches their input, it is possible that they do so not based on the subtle difference in interpretations but rather on other general factors such as frequency. In sum, while I assume that true optionality does not exist in mature grammars, I leave open the possibility of children mis-analyzing variants as an instance of true optionality.

In the next section, I will briefly review notable studies on the acquisition of morpho-syntactic variation and discuss remaining issues in the field. Using wh-variation (discussed in section 1.3.) as a case study, I aim to address some of these issues in this dissertation.

1.2. The acquisition of variation

This section provides a brief literature review on the acquisition of variation. Section 1.2.1. discusses theoretical frameworks while section 1.2.2. reviews previous studies on variation acquisition. Most studies focus on morphological variation and a few studies touch on syntactic variation. Section 1.2.3. and 1.2.4. discuss some aspects of variation in terms of current issues in language acquisition, such as the debate about domain-general versus domain-specific.

1.2.1. Theories of acquisition

Within the Principles and Parameters framework (Chomsky, 1981), language acquisition is often described as a process of choosing the right parameters of cross-linguistic variation. A child is endowed biologically with an innate system of UG principles including some that are parametrized to apply or not in a given language, and to acquire language is to set the parameter values correctly so that they correspond to the properties of the input language. Instead of building a grammar from the ground up, the child's task is now reduced to a selection task. Thus, parameter setting is often used to explain children's seemingly quick, overwhelmingly uniform, and effortless acquisition process despite not being exposed to rich enough language data (poverty of stimulus), as well as children's ability to acquire a native language regardless of what that language is. While it can account for cross-linguistic variation, parameter setting is discussed less frequently when it comes to variation within a language. Van Kampen (2004) argues that two values of a parameter being mutually exclusive does not necessarily mean that two variants within a grammar cannot exist. According to van Kampen, a parameter can withhold or add a grammatical feature to a category in the lexicon, which results in either a default value or

a marked value, respectively. The optionality in variation is due to the marked value being added to a lexical item as an optional value.

On the other hand, Westergaard (2009) proposes that grammar competition does not apply at the macro-parameter level; instead, Westergaard focuses on the micro-level where variation depends on fine linguistic distinctions (such as the pragmatic differences between two variants). The Micro-cue model is developed based on the cue-based model by Lightfoot (1999), and it is originally proposed to account for variation. Westergaard develops the model based on studies on children's acquisition of Norwegian V2. In particular, Norwegian wh-questions typically display V2, but there is variation depending on many factors, including for example the length of the wh-word (monosyllabic wh-phrases can license non-V2) and information structure (discourse-given subjects favor non-V2). Westergaard (2009) finds that children produce both variants and correctly use non-V2 In appropriate cases, showing that children are sensitive to the fine context distinctions between V2 and non-V2. Westergaard claims that a standard syntactic parameter-based approach will lead to overgeneralization of one variant, while in the Micro-cue model, information structure is integrated into the syntactic structure. A micro-cue contains the context for a particular word order, and such cue is language-specific and needs to be learned. While micro-cues are not provided by UG, they are built according to UG principles (assuming UG is necessary because it enables children to parse the linguistic data, select relevant primitives and build up syntactic structure). In the Micro-cue model, there is no parameter, and acquiring a particular type of movement means acquiring a set of micro-cues leading to that movement. For example, children acquiring inversion in wh-questions need to learn numerous cues such as the different clause types (declaratives

versus interrogatives), different classes of verbs (auxiliaries versus lexical verbs), and so on.

A distinct line of acquisition research that has been increasingly studied in the past two decades is statistical learning. The core idea is that children acquire language by tracking its statistical structure (or raw frequencies). This idea is supported by research showing that infants as young as 8-month-old can already track the transitional probabilities between adjacent syllables to segment words in an artificial language (Saffran, Aslin, & Newport, 1996) as well as natural language (Pelucchi, Hay, & Saffran, 2009). Going beyond learning words, Thompson & Newport (2007) have demonstrated that learners can also rely on transitional probabilities to segment sentences into phrases to acquire the syntactic structure of a language. However, most of the experiments in statistical learning tap into comprehension, i.e., more passive knowledge of the language via preferential looking or 2-alternative forced-choice tasks. Bridging the idea of statistical learning and UG, Yang proposes the Variational Learner (Yang, 2002, 2010), which accounts for variation at the rule level by allowing for different grammatical rules in the grammar space, and the probabilities for each rule changing, based on whether an adult input supports it or not. These grammatical rules compete with one another instead of operating in tandem. The model predicts that compared to consistent input, variation in the input will result in acquisition taking longer, but eventually the learners will learn that the target grammar is the one with the highest probability. Westergaard (2014) agrees that when there is variation there will also be some kind of competition (i.e., preference), but criticizes Yang's approach as insensitive to the linguistic contexts the variants appear in.

In general, to account for variation at the rule and construction level, a theory should be specific enough to consider the fine linguistic distinctions between variants, but also general enough to capture the quick and seemingly effortless acquisition process of children.

1.2.2. Previous studies on the acquisition of variation

Within the studies on the acquisition of multiple variants, there seems to be a general consensus: young children's production of variants typically does not perfectly align with their input. Studies in both natural and artificial languages have suggested that when there are multiple variants of the same grammatical item in the input, children tend to only produce ("regularize to") the dominant variant (or a subset of variants). This section summarizes previous studies on the acquisition of variation and discusses different aspects of regularization.

Conditioned variation

a. Pragmatically conditioned – Pozzan & Valian, 2016

Grammatical choices are usually conditioned by some factors, be they lexical, phonological, discourse-based, or other. An interesting question for the acquisition of conditioned variation is whether children are sensitive to the conditioning factors. Note that, to be consistent with the definition of variation in section 1.1, conditioning factors are different from "rules" when variant #1 is allowed only when factor A is satisfied and variant #2 is allowed only when factor B is satisfied (e.g., using the determiner "a" for nonspecific and "the" for specific nouns). Instead, conditioning factors are factors that when present or satisfied, a variant is *more likely* to be used (though there are still neutral situations that license all variants).

Unfortunately, there have not been many studies on this topic, and the most relevant one is a study from Pozzan & Valian (2016) on children's production of polar (or yes-no) questions. Although the original study does not focus on variation, the report on children's production of polar questions in the study fits nicely into the literature on the acquisition of multiple variants. Overall, the study investigates the prevalence of two polar question variants, inverted (e.g., "is the dog here?") versus non-inverted/in situ (e.g., "the dog is here?"), in children's input and their own production. In English, an inverted polar question is the default variant, while a non-inverted question can be used when certain pragmatic factors (such as a presupposition) are present. Despite the presence of 38% of non-inverted polar questions in child-directed speech (based on corpus analysis), they show that children almost never produced non-inverted polar questions (only 1 in 264 occurrences or ~0.4%) in their experimental setting.

Interpreting such results is tricky. First of all, this result may be due to a "lab effect" that makes children behave differently from how they use language in a natural setting, as there are a few corpus-based studies reporting non-inversion in spontaneously produced children's polar questions (e.g., Rowland, 2007). It is possible that young children are conscious of being in a "lab experiment" and strive to produce a more formal or commonly used variant. For example, results from studies on wh-question production with French-speaking children vary based on their methodology (i.e., corpus analysis versus elicited experiment), a point which I will return to in section 4.2. Another possibility is that the pragmatic conditions in the experiment do not motivate non-inverted questions, which are typically used when the speaker has some preexisting belief (Gunlogson, 2002). This explanation would assume that children have acquired the pragmatic factor(s) conditioning

the two polar question variants. To be certain whether children regularize or not, we have to test their production in contexts that are pragmatically plausible for both variants, which, to the best of my knowledge, has not been done.

Overall, it is not clear whether this result reflects children's behavior in a specific environment (e.g., lab setting) or their general acquisition and usage pattern, and only in the latter case can we claim regularization. However, for the sake of the discussion, let us assume here that this is indeed children's general acquisition pattern of polar questions. One then can say, based on the results, that children have regularized their production by exclusively producing only one of the two polar question variants they are exposed to. Exactly why they do this is another tricky question with no clear answer yet. There are multiple factors that can come into play, based on the characteristics of inverted polar questions: children may have regularized to the more frequent variant (62% in the input), they may have regularized to the variant that is less pragmatically-constrained and discourse-marked, or they may have regularized to the variant that they are exposed to first or most recently (primacy and recency effect).

To summarize, Pozzan & Valian investigate children's production patterns when being exposed to two variants of polar questions. The study is not designed to study variation and lacks control conditions to inform whether such behavior is an actual instance of regularization. However, children consistently only produced one variant in their elicited experiment, a pattern that is similar to the regularization behavior typically found in studies focused on the acquisition of multiple variants.

b. Semantically conditioned – Schwab, Lew-Williams, & Goldberg (2018)

Schwab et al. (2018) have investigated adult and child learners' ability to learn grammatical variants of classifiers conditioned by semantics. Specifically, in their experiment, two-thirds of the nouns were animate and appeared with a classifier that is conditioned based on natural gender (e.g., *dax* for female and *po* for male). One-third of the nouns were paired with a classifier randomly (e.g., *dax/po* arbitrarily). Additionally, one classifier appeared more frequently than the other (e.g., *dax* appeared twice as frequently as *po*). Participants were asked to produce the classifiers for 30 test trials that include familiar items as well as novel gendered, animate items. Half of the test items required *dax* classifiers and half required *po.*

Adults in the experiment performed at ceiling-level accuracy, using each classifier 50% of the time. However, 17 out of 20 (85%) of the child participants displayed regularizing behavior: 7 of them exclusively produced the more frequent classifier 100% of the time, 4 produced the more frequent classifier 69% of the time, and 6 of them produced the less frequent classifier 69% of the time. Only 3 children regularly produced both classifiers. Interestingly, in a follow-up experiment with a 2-alternative-forced-choice task, children demonstrated a good understanding of which classifier should go with which noun, and none of them regularized to a classifier.

This experiment shows that children are able to learn the semantic factor conditioning the two classifier variants, as demonstrated in the 2AFC task. Yet, children still regularize in their elicited production. This suggests that regularization is not due to the inability to learn the variants but is more likely due to other performance factors. Moreover, while the majority of children in the experiment regularized to the more frequent

variant, there were still a few children regularizing to the less frequent variant, suggesting that frequency is not the sole factor driving regularization.

Unpredictable variation

Unpredictable, inconsistent variation in natural language is rare. Nevertheless, studies on unpredictable variation allow us to explore children's acquisitional path with variable input in the absence of other (linguistic) cues. Do children still regularize? Is their regularization influenced by domain-general and cognitive factors such as frequency? Below is a review of two studies on the acquisition of unpredictable variation.

a. Fering – Hendricks, Miller, & Jackson (2018)

Hendricks, Miller, & Jackson (2018) report a study on the acquisition of gender in Fering, a dialect of (Germanic) North Frisian which is marked on determiners (*de* and *det*). In current Fering, it has been reported that speakers sometimes mark the gender based on the noun's animacy and natural gender instead of the grammatical gender (Hendricks, 2014). Moreover, data from interviews as well as experiments have indicated that there is both inter- and intra-speaker inconsistency in gender marking. This means that speakers may not consistently use the same gender on a noun but alternate between the two determiners, and across speakers there is no consensus about which gender should be used for a given noun. Thus, children are exposed to unpredictable and inconsistent use of the two determiner variants.

All the children in the experiment were exposed to both Frisian and Fering, with the exposure rate to Fering ranging from 33%-100% (M = 76.5%). Children and adult controls completed an elicited task in which they had to name a circled object in one of

two images. None (0%) of the adult participants showed consistency in gender marking, in line with results from previous studies. However, 10 out of 25 (40%) child participants regularized gender marking by producing it consistently. Interestingly, those who regularized, on average, were exposed to less Fering (M = 64.8%) than those who showed adult-like inconsistency (M = 86.5%). Hendricks et al. suggest that the amount of data is one factor that can condition regularization.

b. American Sign Language – Singleton & Newport (2004)

Singleton & Newport (2004) report on a case study of Simon, a deaf child whose parents were non-native signers. Simon's deaf parents learned American Sign Language (ASL) at a late age (mother 15, father 16) and did not achieve native-like fluency. The study reports that they only used motion and location morphemes correctly 70% of the time, and handshape morphemes correctly 45% of the time. However, Simon's production was much more regularized and almost indistinguishable from children learning from native signers. His scores exceeded his parents' by almost 20% in each category of motion/location morpheme (e.g., orientation, manner, location etc.) that they were tested on.

Strictly speaking, the inconsistency in Simon's parental input does not match the definition of variation described above, as the "variants" in this case are ungrammatical. Still, this study has been discussed frequently in the literature on regularization in language acquisition. The main factor separating this from other studies incorporating ungrammatical variants (which typically results in over-generalization) is that the inconsistencies here are present in the input, unlike the cases of English past tense or dative structure in which the child learner would almost never hear ungrammatical expressions

such as “goed” or “Mary made Sam a decision”. It is important to note that virtually all of Simon’s exposure to ASL came from his parents only. Thus, ungrammatical inconsistencies were present in Simon’s input, and it was unlikely for him to receive correction on his own production of such inconsistencies. For Simon to surpass his parents and arrive at native-like fluency, he most likely had regularized his production to reduce the inconsistencies.

c. Unpredictable variation in artificial language learning experiments

Hudson Kam & Newport’s (2009) study the acquisition of determiner variants in 5-to-7-year-old children and adults. They exposed participants to a language consisting of a main determiner and 2 to 4 noise determiners. The percentage of the main determiners was always 60%, while that of the noise determiners ranged from 10% to 20% each depending on the condition. Overall, while adults performed better than children in the grammatical judgment task, all of them did relatively well and demonstrated that they had learned the simple grammar. In the production task, adults probability-matched the input and reproduced the inconsistency. Most children (14 out of 15, or 93%), on the other hand, regularized their production to reduce the inconsistency, although the direction of regularization varied among children. Specifically, half of those who regularized (7 out of 14) boosted the frequency of the main determiner and always used only the main determiner 100% of the time. Interestingly, 1 child regularized to a noise determiner instead by only using the noise determiner, and 4 children reduced variation by completely dropping determiners.

Wonnacott (2011) also studies children’s sensitivity to input frequency through an artificial language learning experiment that includes two noun-attaching particles with no

semantics. One particle always appeared 3 times more frequently than the other particle, though each noun was followed by both articles alternatively. Wonnacott reports above chance regularization ($p = .035$), with 7 out of 20 children producing 100% of the higher-frequency particle.

Austin (2010) conducts a somewhat similar artificial language learning experiment to Hudson Kam & Newport. The language consisted of 5 nouns, 5 verbs, and 2 determiners. The dominant determiner was used 60% of the time and the minority determiner was used 40% of the time, but the usage was probabilistic and unpredictable. Participants included adults, 8 older children (7;6 to 8;5 years old), and 10 younger children (5;6 to 6;9 years old). In general, adults used the dominant determiner 61.8% of the time, similarly to the original distribution in the input, while children produced the dominant determiner 82.7% of the time. However, a closer look at the children's behaviors suggests that the older children produced the dominant determiner more frequently than the adults but less frequently than the younger children (at 68%). This study shows developmental changes with regard to regularization.

The studies above suggest that not all variants are equal – some may be easier to learn or simply preferred by learners. When being exposed to multiple variants in the input, children do not reproduce a probability-matched distribution of the variants but tend to regularize to a subset of variants. The tendency to regularize the language is an important indicator that children do not simply imitate their input but they modify and impose their own rules.

1.2.3. Domain-general and domain-specific regularization

So far, we have only seen regularization being discussed in language learning studies. Is regularization language-specific? Apparently, such behavior is also found in children exposed to non-linguistic stimuli. Derks & Paclisanu (1967) presented participants with two flashing lights, one activated on 70% of the trials while the other was activated on the remaining 30%. Participants were later asked to predict which light would activate. The adults in the experiment probability-matched what they saw, predicting one light 70% of the time and the other 30%. On the contrary, 4-year-old children always selected the more probable light. Interestingly, by regularizing to the more probable (higher frequency) light, children maximized their chance of having a correct guess and outperformed adults. Yurovsky et al. (2013) found similar behavior in a visual learning task with 11-month-old infants. Their task consisted of "cue" shapes and videos of cartoon characters. One shape is a Strong Cue in that, after its initial appearance, there was a 70% chance of seeing a cartoon video followed and a 30% chance of seeing a blank screen. The other shape is a Weak Cue, predicting a cartoon video only 40% of the time. Finally, when two shapes appeared at the same time, no video would follow. Seeing both cues allowed participants to know where *not* to look, and seeing either of the Weak or Strong cues allowed participants to form a prediction about the likelihood of a following video. Results showed that both adults and infants were able to rely on the cues to predict the box that the video would appear in and adjust their gaze to look at the correct location. However, adults relied more on the Strong Cue, and showed less predictive looking when they saw a Weak Cue or both cues. Infants, on the other hand, regularized, treating the Weak Cue just like the

Strong Cue, increasing their predictive looking when seeing either of the cues, and only showed less predictive looking when seeing both cues.

In both of the experiments described above, we see the same tendency to probability-match in adults and to regularize to a simpler pattern in children, even those as young as 11-month-old. Such regularizing tendency happens not only with linguistic stimuli but also with visual stimuli. Thus, it is possible that regularization is a domain-general response when young learners are faced with variation in their input. Consistent with this conclusion, regularization relies on a domain-general factor such as frequency. Indeed, in many of the studies described above (e.g., Schwab et al., 2018; Wonnacott, 2011), when there are multiple variants with different frequencies, children typically regularize to the more frequent variant. However, it would be an oversimplification to stop at frequency and claim that it explains all regularization patterns. In both Schwab et al. (2018) and Hudson Kam & Newport (2009), while the number of children regularizing to the more frequent variant is higher, some children do regularize to the less frequent variant, suggesting that they may have focused on some other factors. Moreover, children also sometimes impose their own rule in unpredictable variation: while the determiner variants in Hudson Kam & Newport alternate arbitrarily, two children only produced determiners in transitive but not in intransitive sentences, and one child produced only the main determiner only with objects but never with subjects. These children did not rely on frequency to choose their 'target' variant, but came up with their own linguistic rules to systematize the language. Ferdinand, Kirby, & Smith (2019) claim that participants in a variation-learning task are sensitive to frequency regardless of the learning domain, but linguistic domain may impact the encoding of frequency differently. They find that while

regularization happens both with linguistic and non-linguistic (i.e., visual) stimuli, learners regularize more in linguistic tasks, suggesting that communicative goals may play a role in motivating regularization. Unfortunately, there have not been a lot of studies on the domain-general versus domain-specific debate of regularization. I will return to this discussion in Chapter 4, where I explore linguistic factors that may play a role in regularization.

To sum up, while regularization is a domain-general behavior in that it happens when children learn variation in different types of information, the level of sensitivity to variation can vary across different domains and modalities. Children may well rely on both domain-general and language-specific factors to determine which variant to regularize to in language acquisition. In this dissertation, I will specifically test one previously overlooked domain-specific factor – discourse (un)markedness.

1.2.4. Other issues in the acquisition of variation

As we have seen above, many of the studies targeting variation are artificial language learning experiments (e.g., Hudson Kam & Newport, 2005, 2009; Schwab, Lew-Williams, & Goldberg, 2018; Wonnacott, 2011) although there are a few studies in natural languages as well (e.g., Pozzan & Valian, 2016; Hendricks, Miller, & Jackson, 2018). Among the studies on variation in language acquisition, the general claim is that children tend to strongly prefer producing (i.e., regularize to) only a subset of the variants they are exposed to. Given the relatively small number of studies on regularization, there are still several aspects to be explored. For example, with regards to factors conditioning regularization, a number of studies (e.g., Hudson Kam & Newport, 2009; Wonnacott, 2011) have suggested a domain-general factor, namely frequency. Schwab et al. (2018)

echo this claim, adding that child learners appear to be oblivious to semantic conditioning and instead regularize to the higher-frequent variant. This opens up a question about the role of other language-specific factors[1] in regularization, such as syntactic economy, prosodic prominence, pragmatic principles, discourse markedness, and so on.

Another unresolved question is the motivation for regularization. Hudson Kam & Newport (2009) suggest memory limitations are the main motivation, while Perfors (2012) refutes the role of memory alone and instead proposes a regularization bias. While a number of studies (e.g., Hudson Kam & Newport, 2009; Perfors, 2012; Hudson Kam, 2019) have explored this issue by varying the conditions for learning (memory encoding) and testing (memory retrieval), none has attempted to test for a direct relationship between working memory ability and the degree of regularization in learners. These are questions that will be explored in this dissertation.

1.3. Variation in wh-questions

I now move from the discussion of general variation in (morpho)-syntax to the specific case of variation in wh-questions, which is the empirical focus of this dissertation. Based on their question formation strategy, wh-questions across languages are typically classified into wh-in-situ (e.g., Chinese, Vietnamese, Hindi) and wh-fronting (e.g., English, German). Interestingly, while English is traditionally characterized as a simple wh-fronting language, it does allow a type of information-seeking wh-in-situ questions in certain

[1] Culbertson, Smolensky, & Legendre (2012) suggest harmony in languages as a potential factor, though the term *harmony* is rather broad and its definition can vary depending on the linguistic scope.

contexts, such as legal questioning (Pires & Taylor, 2007), quiz shows (Comyn, 2013), classroom settings, and child-directed speech. In child-directed speech, parents often alternate between fronted wh-questions and wh-in-situ, as shown in example (6). Since this work focuses on language acquisition which takes child-directed speech as the main source of input, it is reasonable to use wh-questions as a case study for rule-and-construction variation. I do not make any claims about the status of wh-in-situ (whether it is well-formed or considered as an alternative option for fronted questions) in other contexts outside of child-directed speech.

(6) Example from Weist corpus (Weist & Zevenbergen, 2008)
Father: hey Roman, if the dinosaur roars what's the baby gonna do?
Child: it gonna roar and it's gonna say like this (roar).
Father: yeah but if the dinosaur roars the baby is gonna be what?
Child: scared.
Father: no the baby's scared what's it gonna do?
Child: it gonna eat the thing.
Father: no no the baby, the baby's gonna what?

Monoclausal wh-questions are chosen as the target case for the investigation of variation acquisition for a number of reasons. First of all, wh-questions are syntactically complex enough to observe potential regularization in syntax (as opposed to morphology or morphosyntax). Secondly, in many languages (e.g., French, Brazilian Portuguese, Spanish, Malay, Bàsàa), wh-questions allow for syntactic variation. Thus, there is a large number of available studies on wh-question acquisition cross-linguistically, making it easier to compare and verify our results against the existing literature. Lastly, wh-question variation in English, specifically the in-situ variant (which I will refer to as *probe question*), has not been discussed much, especially in language acquisition (aside from Chapter 4 largely published as Nguyen & Legendre, 2021). Thus, in addition to the discussion on

multiple variants acquisition, the work here also contributes to the empirical understanding of wh-question acquisition in English. This section presents a brief literature review on the analysis of wh-questions in general, while a more detailed analysis and discussion of English wh-in-situ will be presented in Chapter 2.

1.3.1. Different syntactic strategies for question formation

Syntactic movement characteristic of the fronted strategy is standardly understood to be technically driven by head features. Fronted wh-questions have traditionally been analyzed in terms of [+Q] and [+WH] features on C (Cheng, 1991; Rizzi, 1996; Adger, 2003; for a Minimalist implementation in terms of formal features see Sobin, 2010). These encode interrogative illocutionary force and operator-status of wh-phrases, respectively. Assuming X'-theory, the [+Q] feature on C (the head of CP) triggers overt T to C head movement in information-seeking questions and the [+WH] feature triggers overt phrasal wh-movement to SpecCP.

Other analyses have introduced additional information structure features to the understanding of wh-questions, typically on the basis of languages that exploit information structure properties in various syntactic phenomena. In multiple fronting languages, it has been argued that wh-movement is motivated by a focus requirement. Puskas (1992) further assumes that the wh-criterion (Rizzi, 1990), which requires that a wh-phrase carrying the feature [+WH] move to the C system in interrogatives to instantiate the Spec-Head relation, is transmitted from C to Foc (the head of FocusP) in languages like Hungarian. When a wh-phrase is moved to the focus position to satisfy the focus requirement it also satisfies the wh-criterion. In other words, the functional head Foc contains both the features [+WH] and [+FOC]. However, Puskas also demonstrates that the focus requirement is independent

of the wh-criterion, and that wh-movement in Hungarian is motivated by both features. Choi (1996) proposes an alternative approach by introducing two information structure features, [±NEW] (representing new information) and [±PROM] (representing topic-like information). Building on Choi, Mycock (2013) states that overt wh-movement is mainly motivated by [+PROM].

The analysis for wh-in-situ is more complex. Based on the analysis of Chinese, Huang (1982) proposes that all wh-phrases are quantifiers that undergo movement, overtly in English-like languages but covertly in Chinese. There are many motivations for such a proposal. From a semantic analysis, wh-in-situ questions take wide scope, as shown in (7) and (8). In (7b), the answer focuses on *which book* and the wh-in-situ phrase takes scope across clause boundaries, suggesting movement to the matrix. In (8), the wh-phrase *what* has to take *everyone* in its scope to yield the correct interpretation, suggesting that *what* is fronted to specCP at LF.

(7) Q: Which student knows where Mary bought which book?

a. Single-pair reading: Bill knows where Mary bought which book.
b. Multiple-pair reading: Bill knows where Mary bought Ulysses and John knows where Mary bought Moby Dick.

(8) Meige ren dou mai-le shen-me?
Everyone all buy-ASP what
'What did everyone buy'?

From a syntactic analysis, both fronted and in-situ wh-questions display similar locality effects (Cheng, 2009), suggesting that wh-in-situ undergoes wh-movement, at least in languages like Japanese:

(9) a. *What do you remember where we bought?
b. *Doko-de nani-o katta ka oboete-iru no? (Japanese)
where-at what-ACC bought Q remember

However, there are arguments against positing LF wh-movement cross-linguistically as well. Some evidence includes data showing insensitivity to islands (example (10), taken from Pires & Taylor, 2007) and asymmetry reflected in binding (example (11), taken from Cheng, 2009). If the wh-phrase undergoes movement, *himself* in (11b) is incorrectly predicted acceptable as anaphoric to John, similarly to (11a). These examples suggest that wh-in-situ does not involve LF movement.

(10) a. (Mandarin Chinese)
Hufei xihuan nei-ben shei xie de shu
Hufei like that-CL who write POS book
'Who is the person x such that Hufei likes the book that x wrote?'
b. (Brazilian Portuguese)
E aí, você vai entrevistar o homem que ganhou na loteria quando?
So, you will interview the man that won in the lottery when?

(11) a. $John_i$ wondered [which pictures of $himself_{i/k}$] $Bill_k$ liked t_{wh}.
b. *$John_i$ wondered when $Mary_j$ saw [which pictures of $himself_i$]

There is still no full consensus about the covert movement status of wh-in-situ (see Cheng, 2009 for a full review). In recent years, researchers have paid more and more attention to pragmatic requirements of wh-in-situ and how they potentially interact with the syntactic structure of the question (e.g., Pires & Taylor, 2007; Hamlaoui, 2011; Faure & Palasis, 2020). Setting aside the question of covert movement, I believe that a syntax-pragmatics integrated approach is necessary to account for wh-variation. The pragmatics of wh-variation is discussed in more detail in Chapter 2.

1.3.2. The acquisition of wh-questions

Wh-questions have been extensively studied in the acquisition literature (e.g., Bellugi, 1965; Kuczaj & Brannick, 1979; Valian, 1991; Rowland & Pine, 2000; Rowland, 2007). Much of the earlier work on the acquisition of wh-questions contributes to the

classic debate between two lines of thought: nativists versus constructivists. The main assumption of the nativist account is that children, just like adults, have access to full grammatical knowledge, but may not be able to utilize such knowledge because of various cognitive constraints (Valian, 1986). In the case of wh-questions, "Continuity" nativists (e.g., Valian, 1986) have attributed typical errors found in children's production (e.g., omission errors, failure to invert the auxiliary, double auxiliary etc.) to problems in memory and processing load, i.e., due to limited memory and planning capacity, children are unable to process all the rules such as fronting the wh-word and inserting do-support or inverting the auxiliary. "Competence/Maturation" nativists instead propose that some aspects of the relevant knowledge are not available to children until later in development through maturation. For example, Weinberg (1990) and Valian et al. (1992) argue for the existence of a UG hypothesis space which includes grammatical structures of all possible languages. Upon accumulating input from their native language, children learn to set the proper parameter(s) for the target structure(s). The errors they make are thus a result of confusion about how wh-question formation rules work in their particular language. For example, English-speaking children may make wh-in-situ errors (as is allowed in Chinese, Japanese, Vietnamese etc.) or inversion errors (as is allowed in French).

Constructivists reject the need to posit a specific genetic adaptation for grammar or language and instead propose that children learn to construct the language from their input. In children's early production, there is no syntactic structure, but they learn item-based chunks in which they can insert different lexical items (Tomasello, 2005). For example, instead of learning the syntactic rule of inverting the auxiliary, children acquire the structure through using semiformulaic wh-word + auxiliary frames (Rowland, Pine,

Lieven, & Theakston, 2005). Constructivists often criticize nativist theory for not being able to account for (1) the co-occurrence of correct and incorrect utterances and (2) the differences in error rate specific to the wh-word and the auxiliary word (Lieven & Tomasello, 2008). Instead, they attempt to remedy such problems by looking into the frequency rate of each wh-aux combination in the input.

In general, constructivists tend to rely on (pure) frequency while nativists rely on (transformational) target grammars and syntactic sources of possible discrepancies when analyzing children's production. When it comes to variation in wh-questions, constructivists predict children's production to be in line with their input, and if there is a preference it should be towards the variant children hear more frequently (Rowland & Pine, 2000). On the other hand, nativists would predict, for example, that children prefer fewer transformational operations at first (i.e., a preference for economy) (Platzack, 1996; van Kampen, 1997).

While our study on the acquisition of multiple wh-variants is not originally designed to verify general theories of syntactic acquisition proposed by constructivists or nativists, it nevertheless can make a novel contribution to this debate. Do children simply imitate their input by reproducing all variants at a similar distributional frequency? More importantly, if there is regularization, is it simply a result of frequency redistribution of the input (i.e., boosting the frequency of the higher-frequent one), or is regularization also conditioned by linguistic factors? The first case is relatively uninteresting, as this may simply be an aftereffect due to children producing less data than their parental input, and therefore cannot support or argue against constructivism. If, however, children rely on

certain linguistic properties to regularize their production, then this would potentially support some aspects of nativism.

Summary

In this chapter, we have seen that language contains variation, and children are exposed to different types of variation in the language input during acquisition. Studying the acquisition of variation helps us understand more about children's ability to acquire probabilistic input, their preferences at choice points as well as factors contributing to such preferences. The literature review has introduced that wh-questions allow for variation in expression in many languages, and there is a large number of studies on wh-question acquisition cross-linguistically. However, there is a lack of discussion on English wh-variation outside of the semantics literature. In particular, in language acquisition, there is a strong assumption that English-speaking children are never exposed to information-seeking wh-in-situ (e.g., Takahashi, 1991; Yip and Matthews, 2000, 2007; Becker & Gotowski, 2015). Thus, using wh-variants as an in-depth case study not only furthers our broad understanding of the acquisition of variation but also contributes to the study of English wh-question acquisition specifically. With regard to multiple variants acquisition, there are still remaining questions about the motivation of and factors conditioning the tendency to reduce variation through regularizing to one variant. Overall, this dissertation has a narrow goal of understanding the acquisition of wh-variants in English and a broader goal of gaining more insights into the acquisition of variation in general. The main questions this dissertation strives to answer therefore are:

- Do English-speaking children produce wh-question variants in a way that matches with the distribution in their input, or do they regularize to one variant?
- If there is regularization, what factors condition it? Are the factors domain-general (such as frequency or internal consistency) or linguistic-specific (such as syntactic economy or pragmatic unmarkedness)? Or a combination of both?
- How and why does regularization in variation acquisition happen?

On the basis of corpus studies and behavioral studies of comprehension and production, this dissertation makes the following claims about the acquisition of wh-questions:

1. Besides fronted wh-questions, English also allows for another information-seeking question variant: in-situ questions (which I refer to as probe questions to differentiate them from in-situ echo questions). Parents sometimes use probe questions interchangeably with fronted questions in child-directed speech.

2. English-speaking children demonstrate good comprehension of probe questions, but regularize to fronted questions in their production.

3. Traditional hypotheses like structural economy-based accounts or frequency-based accounts are not sufficient to explain children's tendency to regularize to fronted questions. I propose instead an explanation that invokes both frequency and discourse requirements.

On the basis of artificial language learning studies and a computational model, this dissertation also makes claims about regularization as follows:

1. Regularization in language acquisition is conditioned by both domain-general factors such as frequency and language-specific factors, such as discourse markedness.

2. Tasks that are more cognitively-taxing are more likely to induce regularization.

3. There is a correlational relationship between working memory and regularization.

The dissertation proceeds as follows: Chapter 2 provides more information on wh-question variation in general. Besides an overview of wh-variation cross-linguistically, this chapter will also discuss several aspects (pragmatics, prosody, and syntax) of English wh-in-situ. Briefly speaking, English wh-in-situ are similar to fronted questions in terms of prosody and the information they request, but are more contextually constrained. Chapter 3 investigates the acquisition of wh-question variants in English-speaking children through a set of corpus analyses and behavioral experiments with 3-to-5 year-old children (which has resulted in a 2022 Best Original Research Article by an Untenured Scientist journal article in *Language Acquisition*). I find that while children are exposed to both variants in the input and demonstrate a good understanding of probe questions as information-seeking questions (in contrast to echo questions), they strongly prefer producing fronted questions. This result cannot be explained by traditional syntactic economy-based accounts, which predict that children prefer producing structurally simpler constructions (in this case, wh-in-situ). While its higher frequency in child-directed speech seems to motivate the preference for fronted questions, cross-linguistic results suggest that frequency alone is not a sufficient explanation. I propose that the discourse constraints on wh-in-situ also play a role in the acquisition of wh-variation. Through the use of two artificial language learning (ALL) experiments that model after the existing wh-variation in English, Chapter 4 further investigates the role of discourse factors in conditioning regularization. The results from these experiments confirm that when controlling for frequency, learners prefer producing ("regularize to") the un- or less-marked variant. Discourse-markedness is thus a domain-

specific factor conditioning the learning of multiple variants. Chapter 5 explores another aspect of regularization, i.e., the role of cognitive ability in motivating regularization. This chapter includes an ALL experiment along with a computational model using data from the behavioral experiments reported in Chapter 3. Results from the ALL experiment suggest that there is a relationship between regularization and working memory, in particular, participants with lower working memory scores tend to regularize more. In line with that, the computational model finds that regularization behavior only appears when certain assumptions about working memory limitations are made. Chapter 6 summarizes and concludes the dissertation.

Chapter 2: A case study of variation: wh-questions

This chapter presents a more detailed discussion of variation in wh-questions. I will start with English wh-questions and introduce a less commonly discussed variant of information-seeking questions, i.e., in-situ probe questions (section 2.1.). I will show that probe questions are similar to fronted wh-questions in terms of prosody and the answer set they define; however, they are more contextually constrained. Section 2.2. shows that wh-in-situ is also observed to be more constrained than the fronted question variant in languages that allow for both options, such as French and Brazilian Portuguese. However, constraints on wh-questions are not necessarily tied to the position of the wh-word, as there are languages in which fronted wh-questions are the constrained variant.

2.1. English wh-questions

While English wh-questions typically involve overtly fronting the wh-word, the language also allows for a subset of questions where the wh-phrase remains in situ (see example (13)). One type of in-situ wh-questions that is frequently discussed in the literature is echo questions (EQs; e.g., Sobin, 1990; Blakemore, 1994; Noh, 1998; Artstein, 2002; Iwata, 2003), and some have claimed that this is the only type of grammatical in-situ questions in English (e.g., Takahashi, 1990; Yip and Matthews, 2000, 2007; Becker and Gotowski, 2015; Park-Johnson, 2017). However, there is another type of information-seeking in-situ wh-questions, in particular, probe questions (PQs), that appear quite often in certain contexts, such as legal questioning (Pires & Taylor, 2007), quiz shows (Comyn, 2013), classroom settings, and child-directed speech. Some researchers view probe

questions as an independent category, while others suggest that they belong to a sub-category of echo questions (Beck and Reis, 2018). On the basis of largely novel evidence, I will show below that probe questions are separate from echo questions, and are typically used to request new information, similar to fronted wh-questions. Thus, PQs in child-directed speech and fronted wh-questions are two variants of information-seeking questions in English.

2.1.1. Probe questions: an understudied variant of information-seeking questions

Separating probe questions from echo questions

PQs are information-seeking questions, while EQs are repetition-seeking questions. Pragmatically, EQs obey a strict linguistic context requirement. As Banfield (1982) has observed, an EQ can only occur as a reaction to a prior utterance. The strict context requirement, plus the specific purpose of EQs as a request for clarification or repetition, leads to a strong presupposition that the addressee knows the answer and can provide it when asked. For example, (12b) and (12c) are infelicitous responses to the EQ in (12a). The response to an EQ must be the original utterance, a synonym of the original utterance, or at the very least, a description that is co-referential with the original utterance (Blakemore, 1994).

(12) a. A: Jimmy just bought an accordion yesterday.
B: Jimmy just bought a WHAT?[2]
b. A: #I don't know.
c. A: #maybe an accordion?

On the other hand, instead of asking for repetition or clarification of a previous utterance, PQs can be thought of as a "fill-in-the-blank" type of question: the addresser

[2] EQs will be systematically represented with a wh-phrase in caps for ease of identification

prompts the addressee for a piece of information by providing the base structure of the answer with a blank slot to fill in. Since the addresser is requesting new information that has not been previously mentioned in discourse, it is perfectly acceptable if the addressee does not know the answer, unlike the case of EQs. In (13), the first answer the child provides is wrong, and the final answer is "I don't know". There is no restriction on the possible set of answers that the child can consider, as long as the answers do not digress from the main question. This is similar to information-seeking questions. We can easily replace the in-situ questions with fronted wh-questions and still get the same answers.

(13) (excerpt from Adam, Brown corpus)
Mother: and he had a sister named what?
Child: Tony.
Mother: no, Tony was the little baby. His sister's name was what?
Child: […] I don't know.
Mother: Her name is Sheila.

In Hamblin semantics (Hamblin, 1973), a question denotes a set of possible answers. A question is essentially a request to identify the true alternative from the set of alternatives. In that spirit, example (14a) is interpreted as (14b).

(14) a. And he had a sister named what?

b. $[[\text{And he had a sister named what?}]]_o = \{\lambda w.\text{Her name}_w \text{ is Mary}, \lambda w.\text{Her name}_w \text{ is Anne}, \lambda w.\text{Her name}_w \text{ is Sheila}, \ldots\}$

$= \lambda w.\text{Her name}_w \text{ is } x \mid x \in D\}$

Beck & Reis (2018) propose that EQs presuppose that a particular answer to the question is already available in the context. In other words, the discourse constraint of EQs requires that a propositional alternative has to be given, which is unusual for questions.

(15) a. A: John bought a book

B: John bought WHAT?

b. $[[\text{John bought } [\text{what}_F]]]_{Alt} = \{ \lambda w.\text{John bought}_w\ z\}$ where z is the unique contextually relevant object of the appropriate type

z := a book

In Beck & Reis's analysis, what separates EQs from typical wh-questions is the focus on the wh-word. A wh-word, by itself, is an alternative trigger. The general function of focus is also to evoke alternatives (Rooth, 1992), and a focused entity has two semantic values: the ordinary value and the focus semantic value. The effect of focus on the wh-word in EQs leads to the availability of the alternative semantic value z. In (15), the unique contextually relevant object of the appropriate type is "a book".

The focus on EQs' wh-word is evidenced by the rising intonation and focal stress it receives. While a PQ and an EQ can contain exactly the same words in the same order (e.g. *this is a what? / this is a WHAT?*), it is not difficult to differentiate them based on their prosody. It is well-acknowledged that EQs have a distinctive intonational pattern, consisting of a rising pitch accent, and specifically, they have a L+H* intonation with a HH% boundary tone (Pierrehumbert, 1980; Bolinger, 1987; Artstein, 2002). In contrast, PQs have a flat or even falling pitch accent and the wh-word does not receive heavy stress (Reis, 2012).

An informal survey with 16 adult participants was conducted to test if people can differentiate EQs and PQs solely based on their prosody. At the beginning of the survey, participants were presented with one written example of an EQ and one of a non-echo in-situ question in context. After that, participants were instructed to listen to 20 short audio files (10 for each type of questions). The questions were presented in random order. Participants were asked to determine whether the question they heard was an echo or a

non-echo question. The audio files were extracted from three different audio corpora (HSLLD, VanHouten, and Weist), all of which instantiate child-directed speech. Each of the sound files lasted between one and two seconds, and no other context information was given. The questions of each type were controlled so that they had similar lengths (examples of an EQ from the audio file*: we gotta do WHAT?;* and of a PQ: *so you were the what?).* The average accuracy of the task was moderately high at 77.8%, which is significantly above chance level ($t(16) = 8.696$, $p < .0001$). This suggests that the two types of questions have different phonological make-ups, and people can distinguish each type of question solely on the basis of their prosodic properties.

Following the informal survey, a more detailed acoustical analysis of EQs and PQs was conducted. I examined the duration and F0 characteristics of the wh-word's vowel in 50 EQs and PQs extracted from three CHILDES audio corpora in English: HSLLD (Dickinson & Tabors, 2001), VanHouten (Van Houten, 1986), and Weist (Weist & Zevenbergen, 2008), to confirm this result. The questions were forced-aligned using the Montreal Forced Aligner (McAuliffe et al., 2017) and analyzed using the PRAAT software (Boersma & Weenink, 2019). As shown in Figure 1, the wh-word pitch contours of EQs and PQs follow opposite directions.

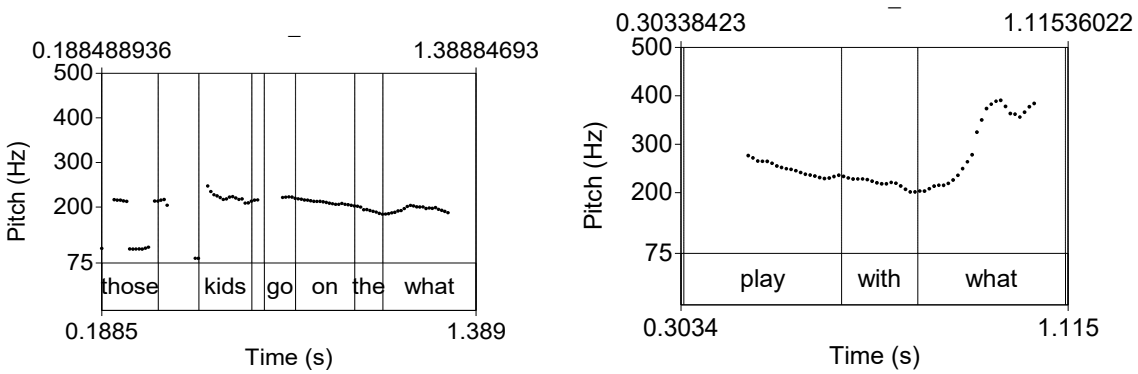

Figure 1. 1a: Example pitch track of a PQ; 1b: Example pitch track of an EQ.

In brief, PQs should not be grouped together with EQs, as they differ in the following properties:

(1) PQs request new information while EQs request repetition or clarification of a previously mentioned information.

(2) PQs do not need an antecendent, while EQs can only occur as a response to the immediately preceding utterance.

(3) The wh-word in PQs has a flat/falling pitch and does not receive focal stress, while the wh-word in EQs has a rising pitch and bears heavy stress.

Other properties of PQs

a. Discourse-pragmatic constraints

While PQs are found in spontaneous speech, they are used less frequently than the fronted wh-variant as information-seeking questions, most likely due to the discourse-pragmatic constraints of PQs. In general, it is established that PQs are heavily discourse linked and need to satisfy Common Ground conditions (Pires & Taylor, 2007), where Common Ground is defined as the shared information given in the discourse or in the extralinguistic context (Stalnaker, 1978). PQs are infelicitous when used out of the blue, as shown by the contrast in (16).

(16) a. A: I need to go to California.
B: And you are leaving when?
b. A: I need to go to California.
B: I see, oh, and where did you buy this bag?
B': #I see, oh, and you bought this bag where?
c. (Seeing somebody reading): You're reading what? (Pires & Taylor, 2007)
d. (Walking past a fellow student in the hallway)
#You're reading what for your Syntax class?

There are a few contexts in which PQs are most frequently found, such as quiz shows (Comyn, 2013) (17a), courtrooms (17b), classrooms (17c), and child-directed speech (17d). Comyn (2013) suggests that these are situations[3] in which the addresser tends to have more information or know something about the answer, suggesting that PQs are used when the addresser is more interested in assessing the addressee's knowledge than in the answer itself.

(17) a. (from Who Wants to Be a Millionaire)
During WWII, U.S. soldiers used the first commercial aerosol cans to hold what?
b. (in a court setting)
You were informed of the fact on what day?
c. (from the Corpus of Contemporary American English)
Teacher: I need to know about displacements. They have a what?
2nd student: Distance.
Teacher: They have a fixed distance and fixed what?
3rd student: Direction.
Teacher: And fixed direction. Fixed distance and fixed direction. Kim, number three. Tell us what you have, Kim. A displacement of how many?
d. (Talking about a family friend)
Mother: and he had a sister named what?
Child: Tony.

According to Farkas & Roelofsen (2017), when two forms have the same semantic content, one form is considered more marked than the other if it is more formally complex or induces more special discourse effects. In that sense, between the two information-seeking wh-question variants, wh-in-situ is the marked form while fronted wh-question is the unmarked form.

[3] Biezma (2020) proposes that these are situations where the addresser has more authority than the addressee, as PQs need to satisfy a power-dynamic requirement. I will discuss Biezma's judgment in more detail in section 2.3.

b. Syntactic properties of probe questions

As briefly mentioned in section 1.2., the debate about covert LF movement in wh-in-situ is still unresolved, as there are conflicting pieces of evidence.

Fronted wh-words in English typically give rise to syntactic locality effects. For example, in (18a), the overt movement of the lower (object) wh-phrase across the higher (subject) wh-phrase in a multiple wh-question violates the Superiority Condition (Chomsky, 1973). (18b) is also ungrammatical because it contains a wh-island, as an embedded wh-phrase overtly moves across another embedded wh-phrase (Ross, 1967) in violation of Subjacency (Chomsky, 1973). Finally, (18c) violates the Coordinate Structure Constraint (Ross, 1967) by extracting a single conjunct from the coordinate structure, resulting in ungrammaticality.

(18) a. *What did who give to Mary?
b. *Which book does Tom know where Jill borrowed from?
c. *What did Mary buy and a candle?

Interestingly, the in-situ EQ in (19a) is grammatical in contrast to (18a), and the EQ in (19b) is grammatical in contrast to (18c), despite a similar surface structure violating the Superiority Condition and Coordinate Structure Constraint. This has been interpreted as solid evidence that in-situ EQs do not involve covert movement at LF, hence they are not constrained by locality conditions (e.g., Kuno and Robinson, 1972; Artstein, 2002; Sobin, 2010; Beck and Reis, 2018).

(19) a. A: What did John give to Mary?
B: What did WHO give to Mary?
b. A: I know Mary bought cakes and candles.
B: You know Mary bought WHAT and candles?

A PQ parallel to (19a), in (20a), sounds much less acceptable. However, (20b) contains a similar violation of the Coordinate Structure Constraint but is better, and in fact is often found in quiz shows. Thus, the problem of (20a) is not entirely due to the in-situ status of the wh-phrase, and is not sufficiently good evidence for PQs' sensitivity to the constraint.

(20) a. Teacher: *Now class, tell me Swan invented what and the light bulb?
b. Teacher: ?Now class, tell me Swan invented the light bulb and what?

Pires & Taylor (2007) observe that information-seeking wh-in-situ in both English and Brazilian Portuguese show insensitivity to islands, as shown in (21). This is in contrast to Lasnik & Saito's (1992) claim about wh-adjuncts being impossible within an island.

(21) A: A man won the lottery this year. Another one did it last year.
(BP) B: E aí, você vai entrevistar o homem que ganhou na loteria quando?
(English) B: So, you will interview the man that won in the lottery when?

Furthermore, PQs cannot license Parasitic Gaps. Parasitic Gaps are licensed by a gap left by A-bar movement (Engdahl, 1983), which includes traces of wh-movement (Chomsky, 1986). Example (22a) illustrates a Parasitic Gap construction licensed by wh-movement, while (22b) shows that a Parasitic Gap is not possible when there is no A-bar movement.

(22) a. Which article did Mary file __ without reading $_{pg}$?
b. *Mary filed the article without reading $_{pg}$.

Assuming covert movement works in a similar way to overt movement, we expect that PQs will license Parasitic Gaps. Yet native speakers tend to reject the PQ in (23):

(23) ?Mary filed which article without reading $_{pg}$?

On the other hand, PQs seem to be able to license Antecedent-Contained Deletion, which is a diagnostic test for covert movement (Pesetsky, 2000). An example of ACD is presented in (24). The elided VP is interpreted as the VP [read t], hence it must have an

antecedent in the form [read t] in the linguistic context. This is resolved by covert movement of the DP containing [$_{VP}$ read t] to a VP-external position (such as specAgrOP):

(24) John read [every book that Mary did [$_{VP}$ Δ]].

Now consider the following scenario: A mother is teaching her son the concept of equivalence. She asks her son: "John knew that Bill read a certain number of books, and he also thought that Fred read the same amount of books as Bill. How many books did John think that Fred read?". When the son fails to answer the question, she asks it again as a probing question, adding a hint about the books that Bill read (recall that parents frequently switch to a probing question after children fail to answer an information-seeking question):

(25) John thought that Fred read how many books that Bill did?

Eight native speakers of English consulted report that the question in (25) is grammatical, suggesting that covert movement is involved.

In sum, the covert status of PQs is still unclear, as there are contradicting sources of evidence both supporting and arguing against movement in PQs. The insensitivity to islands and the inability to license Parasitic Gap suggest that PQs involve no covert movement, while the ability to license Antecedent-Contained Deletion suggests otherwise. I leave this question unresolved within the limits of this dissertation but note that this does not affect our most general claim about PQs being syntactically simpler to fronted questions, as both no movement and covert movement are less computationally costly than overt movement (Procrastinate Principle, Chomsky, 1993)[4].

[4] Movement happens to license formal feature. Covert movement, or movement at LF, only needs to carry formal features that need licensing, while overt movement has to carry full lexical items for pronunciation.

2.2. Wh-question in other languages

2.2.1. French

As is well-known, French allows for considerable variation in wh-questions, with some variation associated with a register difference. The examples in (26), taken from Shlonsky (2012), illustrate some main possibilities. To simplify the matter, I will limit the discussion of the variants to the difference in the placement of the wh-word only, i.e., whether wh-movement is present or not.

(26) a. Tu as quitté qui? (in-situ object wh)
you have left who
Who have you left?

b. Qui tu as quitté? (fronted object wh without inversion)
who you left

c. Qui as-tu quitté? (fronted object wh with inversion)
who have you left

d. Qui est-ce que tu as quitté? (fronted object wh with Q-marker)
who Q you have left

e. C'est qui que tu as quitté? (clefted wh)
It is who that you have left

f. Qui c'est que tu as quitté? (movement of the cleft pivot)
who it is that you have left

In general, the presence versus absence of subject-auxiliary/verb inversion in (26b) versus (26c) marks a register difference (with inversion being characteristic of the formal register). In many cases, the French wh-in-situ position of the wh-phrase can be seen as an alternative to wh-fronting. In terms of prosody, Cheng & Rooryck (2000) propose that French wh-in-situ is licensed by an intonation morpheme. However, other studies have found that rising intonation is not required for in-situ-wh in French (Adli, 2004; Déprez, Syrett, and Kawahara, 2013).

While Chang (1997), Cheng & Rooryck (2000), and others have claimed that French in-situ wh-questions are associated with a strong presupposed context not found in fronted questions (as shown in the negative answer contrast in (27a)-(27b)), such a claim is controversial. In at least the colloquial register of the language, there is strong evidence that the pragmatic constraint attributed to wh-in-situ does not hold. Mathieu (2004) provides the following in-situ examples that elicit perfectly acceptable negative answers – the colloquial register is identifiable from the absence of the negative scope marker *ne* in the answer (*j'ai pas faim* "I am not hungry" in (28)). This is the register relevant to child language acquisition, where *ne* is also characteristically absent from many child-directed utterances (Culbertson, 2010).

(27) a. Q: Marie a acheté quoi? — A: *Rien.
Mary has bought what — Nothing
What has Mary bought?

b. Q: Qu'est-ce que Marie a acheté? — A: Rien.
What Q Mary has bought — Nothing

(28) a. Q: Tu fais quoi dans la vie? (Mathieu, 2004:18)
You do what in the life
What do you do for a living?
A: Rien. Je suis au chômage.
Nothing. I am unemployed.

b. Q: Tu veux manger quoi ce soir? (Mathieu, 2004:18)
You want to eat what tonight
What do you want to eat tonight?
A: Rien. J'ai pas faim.
Nothing. I am not hungry

In certain contexts, one structure may be preferred over the other. Boucher (2010) claims that fronted wh-questions have a less restricted distribution than wh-in-situ and can be used in most discourse contexts. Beyssade (2006) claims that fronted questions are typically used to introduce a new topic while wh-in-situ extends the discourse topic.

Hamlaoui (2011) extends the claim to suggest that the non-wh portion in fronted questions is typically non-given, while that in wh-in-situ is discourse-given. In other words, fronted questions tend to have broad focus where the focus set consists of the entire sentence, while the focus in wh-in-situ is on the wh-word. Example (29) illustrates the differences between the two question types:

(29) a. *Context: Two friends are in now and had planned to visit many cities, including Edinburgh. They are discussing the next steps but B realized that they haven't discussed the specific plan to Edinburgh yet.*

B: Quand est-ce qu'on va à Edimbourg?
When is-it that-we go to Edinburg?
When are we going to Edinburg?

b. *Context: A, B, and C have been discussing a possible trip to Edinburgh, so they have mentioned the idea of "going to Edinburgh". B and C are side-tracked and A wants to bring their attention back to Edinburgh.*

A: On y va quand à Edimbourg?
We there go when to Edinburg
(So), we go to Edinburg when?

In (a), "going to Edinburg" has not been previously mentioned and is therefore discourse-new. It is more appropriate then to use a fronted question. In (b), because the idea of going to Edinburg is a current topic of discussion, it is acceptable to use an in-situ question. In general, Hamlaoui claims that using a fronted question in a discourse-given context (i.e., when wh-in-situ is appropriate) is costly in terms of structural economy; however, using wh-in-situ in a discourse-new context (i.e., when a fronted question is called for) is costly in terms of processing and interpretability (for the listener), as the listener needs to accommodate why the question is asked.

We investigated the Madeleine files[5] from Paris corpus (Morgenstern & Parisse, 2007) to see whether adults in child-directed speech and children in their own production follow the discourse constraints Hamlaoui proposed. We extracted parents-produced and child-produced wh-questions between 1;10 and 4;01 and found that speakers are generally sensitive to the constraints. Out of all in-situ utterances Madeleine's parents produced, 80.6% of them (104 out of 129) were in a discourse-given context. For Madeleine, that percentage was even higher: 90.6% (135 out of 149) of all her wh-in-situ were discourse-given. This confirms Hamlaoui's observation.

However, the discourse requirement for French wh-in-situ does not seem to be a strict rule that *always* needs to be enforced, as wh-in-situ can still be used in out-of-the-blue contexts (Adli, 2006; Baunaz & Patin, 2011). An example is shown in (30), where it is acceptable to use French wh-in-situ as the first sequence of an out-of-the-blue exchange. Recent studies by Zimmermann & Kaiser (2019) and Conveney (2020) have shown that the usage of wh-in-situ seems to be on the rise, and wh-in-situ in contemporary French is less likely to be subject to many of the formal restrictions discussed in older literature.

(30) Pardon, il est quelle heure? (Adli, 2006: 184)
sorry it is what time
Sorry, what time is it?

Overall, wh-in-situ is a viable alternative option for fronted information-seeking questions, and these two variants fit under our definition of variation as spelled out in Chapter 1.

[5] I thank my advisor, Géraldine Legendre, for her tremendous help in coding all the French data mentioned in this dissertation.

2.2.2. Brazilian Portuguese

In Brazilian Portuguese, wh-in-situ is also an alternative way to fronted questions to ask for new information. Similar to French, the "fronted" structure also includes some variants (fronted with *que* in (31b) and fronted with *é que* in (31c)), but I will limit the discussion to the difference between fronted versus in-situ wh-word. The examples below are taken from Grolla (2009).

(31) a. O que o João comprou? (Moved-wh)
What the João bought
What did John buy?

b. O que que o João comprou? (Wh-que)
What that the João bought
What did John buy?

c. O que é que o João comprou? (Wh-é-que)
What is that the João bought
What is it that John bought?

d. O João comprou o quê? (Wh-in-situ)
The João bought what
What did John buy?

Wh-in-situ in Brazilian Portuguese has a falling intonation. In line with Hamlaoui's (2011) claim, DeRomma (2011) states that the non-wh portion in Brazilian Portuguese wh-in-situ tends to require more prominent Common Ground than that in fronted questions. Similarly, Pires & Taylor (2007) state that wh-in-situ is possible in Brazilian Portuguese when the information being requested is expected to be part of the Common Ground, i.e., "information that was previously given in the discourse or in the extralinguistic context". Grolla (2009) agrees that there exist some factors that favor one structure over another, though such factors have not been clearly defined, as native speakers have also reported that wh-in-situ can still be used out of the blue. This suggests that Brazilian Portuguese is

less restrictive in the licensing of wh-in-situ than English (Pires & Taylor, 2007). The example below shows that at least the Brazilian Portuguese *which*-type wh-questions are acceptable as wh-in-situ even out of the blue, while the English counterpart remains infelicitous.

(32) a. (out-of-the-blue): Anna, você está assistindo qual istributio TV essa semana?

b. (out-of-the-blue) #Anna, you're watching which program on TV this week?

2.2.3. Vietnamese

So far, we have seen that wh-questions in English, French, and Brazilian Portuguese share some common characteristics: wh-in-situ questions are discourse-marked and are restricted to a smaller subset of contexts compared to the fronted variant. However, it is important to note that the difference in discourse restriction between wh-in-situ and fronted wh-questions is not necessarily tied to the syntactic difference between the two variants. This section presents an overview of the wh-variation in Vietnamese, a wh-in-situ language.

(33) a. Cậu đã đi đâu?
You PST go where
Where did you go?
b. *Đâu cậu đã đi?
Where you PST go
c. Cậu mua cái gì?
You buy what
What do you buy?
d. *Cái gì cậu mua?
What you buy

(34) a. Tại sao cậu đến đây?
Why you come here
Why do you come here?
b. Cậu tại sao đến đây?
you why come here

c. *Cậu đến đây tại sao?
You come here why

As shown in (33), argument wh-phrases (e.g, 'what', 'who') in Vietnamese questions remain in-situ, and fronting the wh-phrase would result in ungrammatical sentences. However, adjunct "why"- and "how"-questions can vary considerably (Bruening & Tran, 2006). The distinctive behavior of "why" from other wh-words has been observed in many languages, including Korean, Japanese, and Chinese (Pozzan & Valian, 2017), which has attracted a lot of discussions (see Stepanov & Tsai, 2008 for a full review). I will leave aside the "why"-question problem and focus on other wh-words only.

While wh-in-situ is the default option for argument wh-questions, fronted wh-questions are also allowed in certain contexts. The most common use of fronted wh-questions is as trivia or quiz questions, as shown in (35).

(35) a. Cái gì mà Lọ Lem đánh rơi?
What PRT Cinderella drop
What is the thing that Cinderella lost?
b. Lọ Lem đánh rơi cái gì
Cinderella drop what
What did Cinderella lose?
c. Cái gì ai cũng thích?
What everyone also like
What is x such that everyone likes x?

Another possible context is when there is a strong narrow focus on the wh-phrase, which happens when the non-wh portion is given. The role of focus is shown in the contrast between (36) and (37) in eliciting a long versus a short answer (which contains only the information corresponding to the wh-phrase).

Context: a group of friends is planning to watch fireworks together on New Year's Eve.

(36) Q: Xem pháo hoa từ đâu thì đẹp nhất?
watch fireworks from where PRT beautiful most
Where is the best place to watch fireworks?
a. A: Xem pháo hoa từ sân thượng thì đẹp nhất.
watch fireworks from rooftop PRT beautiful most
Watching fireworks from the rooftop is the best.
b. A: Từ sân thượng.
From the rooftop.

(37) Q: Từ đâu xem pháo hoa thì đẹp nhất?
from where watch fireworks PRT beautiful most
Where is the best place to watch fireworks?
a. A: ?#Xem pháo hoa từ sân thượng thì đẹp nhất.
b. A: Từ sân thượng.

For the regular wh-in-situ question in (36), both a long answer and a short answer are acceptable. For the fronted wh-questions in (37), however, speakers tend to prefer the short answer to the long answer. Short answers are claimed to be derived from focus movements (Nishigauchi, 2006).

The distribution of wh-variants in Vietnamese presents a picture contrasting with what we have seen so far in English, French, and Brazilian Portuguese: here, wh-in-situ is the neutral, unmarked variant. In a dominantly wh-in-situ language, fronted wh-questions would need to satisfy certain context requirements, while in a dominantly wh-fronting language, the context requirements would apply to wh-in-situ. This suggests that the discourse restriction(s) of a wh-variant is not necessarily tied to whether the wh-word is moved but more likely to other factors, such as the lower frequency or the later emergence compared to the default/neutral variant of the language. This claim will play an important role in the design of my artificial language learning experiments in Chapter 4, though future work is needed to further verify it.

2.3. Wh-in-situ: interrogatives or declaratives?

So far, I have treated wh-in-situ and fronted-wh questions as two variants of information-seeking questions. This is not unreasonable, given that wh-in-situ shares many similarities with fronted wh-question in form (e.g., containing a wh-phrase and a question mark) as well as in meaning (e.g., requires an answer that carries new information). At the same time, there are proposals to analyze wh-in-situ as declaratives with a wh-expression in focus. In this section, I will review different arguments on the status of wh-in-situ as declaratives, as well as discuss how such analyses may affect the claims in this dissertation.

2.3.1. Wh-in-situ as declaratives: from a syntactic viewpoint

Bobaljik & Wurmbrand (2015) argue that wh-in-situ are "questions with declarative syntax". In other words, despite the interrogative force, wh-in-situ should be analyzed syntactically as a declarative clause. Their claim applies to all the wh-fronting languages with "optional" wh-in-situ. Bobaljik & Wurmbrand observe that if a language has wh-movement, then wh-movement is obligatory in indirect questions. Similarly, if a language truly allows for wh-in-situ, then wh-in-situ is acceptable in indirect questions as well. This is observed in wh-in-situ languages like Chinese and Vietnamese:

(38) (Mandarin Chinese, example from Cheng, 2003)

a. Hufei mai-le shenme?
Hufei buy-ASP what
What did Hufei buy?

b. Botong xiang zhidao [Hufei mai-le shenme]?
Botong want know Hufei buy-ASP what
Botong wants to know what Hufei bought.

(39) (Vietnamese)

a. Cậu đã mua cái gì?
You PST buy what

What did you buy?

b. Tôi muốn biết cậu đã mua cái gì.
I want know you PST buy what
I want to know what you bought.

Yet wh-in-situ cannot occur in indirect questions in wh-fronting languages like English, French, German, Brazilian Portuguese:

(40) a. *He asked me the boy's name is what.

b. *Stark hat gefragt diese Teilhaber erreichen wir wie? (German)
Stark has asked these partners reach we how
Stark asked how we reached these partners.

c. *Je me demande tu parles de quoi. (French)
I me ask you talk of what
I wonder what you are talking about.

d. *O Pedro perguntou você viu quem? (Brazilian Portuguese)
The Pedro asked you saw who
Pedro asked who you saw.

This supports the conclusion that wh-in-situ in these languages is distinct from *true* wh-in-situ in Chinese and Vietnamese. For Bobaljik & Wurmbrand the optional wh-in-situ in these languages is instead a question with declarative syntax (DSQ). They propose that the difference between *true* wh-in-situ and DSQ lies in the interrogative complementizer $C_{WH.}$. DSQs do not carry C_{WH} and therefore the wh-in-situ questions cannot be selected by a higher predicate to form an indirect question.

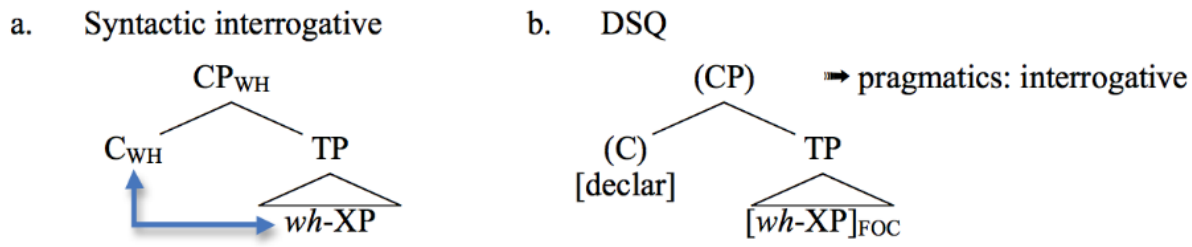

Figure 2. The difference between true wh-questions and wh-questions with declarative syntax. Figure taken from Bobaljik & Wurmbrand (2015).

Bobaljik & Wurmbrand's claim, however, has not gone unchallenged. Sato & Ngui (2017) provide survey data from Colloquial Singapore English showing that wh-in-situ questions in indirect questions are acceptable. Specifically, Sato & Ngui consulted 13 native speakers of Singapore English with two variants of fronted and in-situ wh-questions in indirect questions in (41)-(43):

(41) a. I wonder what Mary bought already.
b. I wonder Mary bought what already.

(42) a. I wonder what John bought for Peter.
b. I wonder John bought what for Peter.

(43) a. John asks who the rice is for.
b. John asks the rice is for who.

Out of 13 native speakers, 11 of them accepted both variants (though 5 speakers expressed a preference for the fronted variant), and 2 accepted only the fronted variant. Sato & Ngui take this result as evidence that Singapore English allows wh-in-situ in indirect questions. Additionally, Sato & Ngui present further data from Malay and Ancash Quechua showing that there are languages permitting both fronted and in-situ indirect wh-questions:

(44) *Malay* (taken from Cole & Hermon, 1998)
a. Dia tidak membeli apa-apa untuk saya.
he neg buy what-what for me
He did not buy anything for me.
b. Dia tidak membeli apa-pun untuk saya.
he neg buy what-also for me
He did not buy anything for me.

(45) *Ancash Quechua* (taken from Cole, 1982)
a. (Qam) tapurqonki ima-ta María munanqanta José rantinanta.
you asked what-acc María wants José buy
You asked what María wants José to buy.
b. (Qam) tapurqonki María munanqanta José ima-ta rantinanta.

You asked María wants José what-acc buy
You asked what María wants José to buy.

In sum, Bobaljik & Wurmbrand claim that wh-in-situ in wh-movement languages is blocked as indirect questions and functions differently from "true" questions as DSQ. However, data from Sato & Ngui show that in-direct wh-in-situ is acceptable in at least one wh-fronting language (Colloquial Singapore English). This weakens Bobaljik & Wurmbrand's claim, though additional studies are needed to fully explain why languages with both wh-in-situ and wh-fronting strategies like English, French, and Brazilian Portuguese do not allow indirect wh-in-situ.

2.3.2. Wh-in-situ as declaratives: from a semantic/pragmatic viewpoint

Biezma (2020) refers to English wh-in-situ as wh-declaratives (WhDec). Biezma claims that English[6] wh-in-situ questions are similar to declaratives in the sense that they function as a proposal to update the common ground, while interrogatives are a proposal to update the context's question stack. For instance, the wh-question *where is the man from?* is only a proposal about the issue to be addressed, which can be accepted (the listener agrees to pursue the issue in discourse) or rejected (leave the context as it was without any updates). The wh-question updates the context's question stack as follows:

[6] Similar claims to Biezma's are also found in other languages, such as French. Glasbergen-Pas (2021) observes that certain French wh-in-situ can be used in out-of-the-blue contexts, while at the same time there are wh-in-situ that are contextually restricted and require strong presupposition (see the discussion on Hamlaoui's claim in 2.2.). Glasbergen-Pas proposes that there are two mechanisms to interpret French wh-in-situ: out of the blue wh-in-situ can be interpreted via covert movement, while wh-in-situ questions that require strong presupposition are interpreted via choice function. Based on the claims from Glasbergen-Pas and Biezma, Doetjes (2021) further proposes that contextually restricted French wh-in-situ are uninformative declaratives.

Let ϕ be the interrogative sentence *where is the boyfriend?*

a. $c + \ulcorner \text{Question}(\phi) \urcorner = \langle cs_c, Q_c, l_c \oslash \ulcorner \phi \urcorner \rangle$

b. $l_c \oslash \ulcorner \phi \urcorner = \langle cs_{l_c}, push(Q_{l_c}, [\![\phi]\!]^o) \rangle$ $\quad (push(Q_{l_c}, [\![\phi]\!]^o) = Q'_c)$

c. Acceptance: $c_2 = \langle cs_{l_c}, Q'_c, \emptyset \rangle$

In contrast, to accept the declarative answer *the man is from Baltimore*, the presupposition that there is a question open in discourse of the form *where is the man from?* needs to be accepted. Similarly, the WhDec *the man is from where?* establishes that the presupposition about a question in discourse of the form *where is the man from?* has been accepted. Infelicitous WhDecs are similar to infelicitous declaratives, i.e., it happens because the listener cannot accommodate how they have accepted the presupposition. WhDecs update the common ground as follows:

Let $c = \langle cs, Q, \emptyset \rangle$ be the initial context and *WhDec* the WhDec *the boyfriend is where?*

a. $c + \ulcorner \text{Assert}(WhDec) \urcorner = \langle cs_c, Q_c, l_c \oplus \ulcorner WhDec \urcorner \rangle$

(i) Accommodate that the local context is $\langle cs_c, Q'_c \rangle = l'_c$
s.t. $Q'_c = push(Q_c, [\![\text{where is the boyfriend?}]\!]^o)$, i.e.
$top(Q'_c) = [\![\text{where is the boyfriend?}]\!]^o$

(ii) Propose the update of cs: (Assertion)
$\langle cs_c, Q'_c, l'_c \oplus \ulcorner WhDec \urcorner \rangle$
$l'_c \oplus \ulcorner WhDec \urcorner = \langle cs_c \cap contentProp(WhDec), Q'_c \rangle$

(recall that $cs_c \cap contentProp(WhDec) = cs_c$; cs_c is trivially updated)

Since the information update from the WhDec is trivial, the question presupposed is left to be addressed by the listeners, and hence it is viewed as a request for information.

Biezma then goes on to make the following empirical claims about WhDecs:

1) WhDecs are heavily discourse-linked.

2) The addresser of WhDecs typically has authority over the addressee as WhDecs force the listener to accept what the next topic to be addressed is, while no authority requirement is needed for WHQs as they only propose to address the topic.

3) Addresser tends to assume that the addressee knows the answer.

4) A mixed sequence of WhDec->WhQ or WhQ->WhDec is not acceptable because this would create an inconsistency about whether the question is proposed or presupposed.

While I agree with Biezma's theoretical definition of WhDecs, let us review these four points in more detail. The first two points are similar to my own observations reported in section 2.1. It is true that wh-in-situ questions in English are rarely used out of the blue and typically satisfy a discourse-given requirement. Furthermore, wh-in-situ tends to occur in contexts that instantiate power dynamics: between parent-child in child-directed speech, teacher-student in classroom settings, judge-defendant in courtrooms, and host-participant in quiz shows. However, it is unclear whether this is due to a requirement of speaker authority as Biezma claims, or due to the fact that speakers of wh-in-situ tend to know something about the answer. While one can argue that having more information (i.e., knowing something about the answer) gives the speaker more authority, such an explanation is not compatible with Biezma's claim about the information update of WhDecs. This is because speaker authority allows the speaker to impose the topic of conversation but simply having more information about the answer is not sufficient to do so. One popular scenario for wh-in-situ questions in child-directed speech is during book reading, in which a parent points to an object in the book and asks about it (e.g., *this is what?*). Example (46) reports a similar scenario but with the roles flipped, which is sometimes found in CHILDES:

(46) *B is reading a book about futuristic designs of ordinary objects. He finds a picture of a coffee maker that looks nothing like how current coffee makers look. B wants to know if his mom can guess what the object in the picture is.*

B: Mom, look at this. This is what?

The wh-in-situ in this scenario is felicitous even though the power dynamics is not right.

Biezma also claims that WhDecs are not possible when it is likely that the addressee does not know the answer. This is illustrated in the contrast between (47) and (48):

(47) I want to repair the sink but I can't find your mother's toolbox. I don't seriously expect you to know this, but just in case,

a. #...your mother's toolbox is where?
b. ...where is your mother's toolbox?

(48) I want to repair the sink. You told me where the toolbox is but I forgot,

a. ...your mother's toolbox is where?
b. ...where is your mother's toolbox?

Since WhDecs presuppose that the question in discourse has been accepted by all participants, it is contradicting to assume that the addressee accepts the question proposal without having sufficient knowledge about it. Assuming that the power dynamics and discourse requirements are both satisfied, the infelicity in (47a) is due to the speaker's assumption about the addressee's lack of knowledge.

However, consider the following scenario:

(49) *Father and son are reading a book about space.*
Father: This is what?
Son: A plane.
Father: (pointing at a telescope) Alright, this one is harder. I'm not sure if you know this but let's give it a try. This is what?

The latter wh-in-situ question in (49) is similar to the wh-in-situ question in (47a), yet it is felicitous despite the speaker admitting that the addressee may not know the answer.

Furthermore, such assumptions about the addressee's knowledge may not hold in a quiz show context, where wh-in-situ is frequently used. Since the goal of a quiz show is to test the participant's knowledge, the host (i.e., the speaker) is not in a position to assume that the participant knows all the answers. Example (50), taken from the Weist child-directed corpus, also demonstrates that the assumption about the addressee's knowledge is

not the key requirement for wh-in-situ. In this example, we see that the child repeatedly gives the wrong (or undesired) answer. According to Biezma, the assumption that the addressee knows the answer is much stronger in WhDecs than in WhQs. However, while the father switches to an in-situ wh-question after initially asking a fronted question in (50), it is strange to think that the father's assumption about the child's ability to answer the question is somehow made stronger after the child has given a wrong answer.

(50) Father: hey Roman, if the dinosaur roars what's the baby gonna do?
Child: it gonna roar and it's gonna say like this (roar).
Father: yeah but if the dinosaur roars the baby is gonna be what?
Child: scared.
Father: no the baby's scared what's it gonna do?
Child: it gonna eat the thing.
Father: no no the baby, the baby's gonna what?

Finally, the example above also challenges the fourth point about the feasibility of a mixed sequence of WhDecs and WhQs, which Biezma's analysis does not allow. While the father's questions are not a straight sequence as they are interrupted by the child's answers, all of these questions target the same topic. Thus, it is strange to think that the father first *proposes* the topic, then *imposes* it with the wh-in-situ, before going back to proposing the same topic again.

Biezma draws differences between WhDec and WhQ based on how they update the context. In other semantics/pragmatics accounts, a question may be defined differently. For instance, Siemand (2001) characterizes an interrogative by its ability to define a set of answers or its usage to elicit information from the addressee. Groenendijk & Roelofsen (2009) defines a question as being uninformative and inquisitive (i.e., consisting of two or

more possibilities). Depending on how the terms "interrogative", "question"[7] are defined, a wh-in-situ question may or may not fit into those categories. It is also possible that the boundary between interrogative and declarative, or question and assertion, is not as clear-cut as it has been claimed. It could be the case that wh-in-situ is ambiguous between the two categories, and depending on several factors such as contexts and prosody, each wh-in-situ question is different.

2.3.3. Discussion

It is clear that the status of wh-in-situ questions is an ongoing debate that is made complicated by a) differences with respect to how "interrogative" and "declarative" are defined and b) differences in acceptability judgments across speakers with regard to wh-in-situ. I will leave this issue open for future research. Importantly, regardless of whether wh-in-situ questions are categorized as interrogative or declarative, the claims that are made in this work about the distribution of wh-variation are unharmed.

Recall that variation is defined in this work as "having more than one grammatical way to express a meaning". Thus, concerning Bobaljik & Wurmbrand's claim that wh-in-situ are questions with declarative syntax, even if wh-in-situ and fronted wh-questions belong to different syntactic categories, they are still variants of each other, and none of the acquisition claims are affected.

---

[7] According to Huddleston (1994), "interrogative" is a syntactic term used to describe the sentence type, contrasting with "declarative" as another syntactic category. Interrogative is marked by the presence of at least one interrogative word (such as *who, what, where* etc.). Declarative is unmarked in the syntactic system. On the other hand, "question" is defined in terms of semantic and pragmatic properties, contrasting with "assertion". The distinctive property of a question is that it defines a set of answers. However, "interrogative" and "question" are sometimes used in the literature to mean the same thing. I have reported them here in consistency with how the terms are used in their original papers, but it's worth noting the distinction between them.

As discussed in section 1.2.2., variants can have subtle differences. As long as speakers can use wh-in-situ and fronted wh-questions interchangeably to extract similar responses in some contexts, the differences pointed out by Biezma or indicated in section 2.2.1. should not disqualify their status as variants of each other. One potential problem, however, pertains to our cross-linguistic approach if such differences are language-specific. It is possible that when we compare the acquisition of wh-variation cross-linguistically, we are comparing the acquisition of wh-interrogative and wh-declarative (in English) with the acquisition of two wh-interrogative types (in other languages). However, as noted in footnote 6, there is also a claim about wh-in-situ in French being declaratives, in line with Biezma's proposal. Given that there are remaining issues with Biezma's proposal, I will leave this discussion aside for now and maintain that the clause type of wh-in-situ questions would not impact the overall acquisition results.

2.4. Chapter 2 summary

Chapter 2 has presented a brief overview of wh-variation in English and cross-linguistically. In English, there are two variants for information-seeking wh-questions: the unmarked wh-fronted form and the marked wh-in-situ form (PQs). While PQs can be used interchangeably with fronted questions in some contexts, they typically need to satisfy Common Ground requirements. Similar restrictions on wh-in-situ are also found in other languages like French and Brazilian Portuguese. However, it is not a universal pattern that wh-in-situ are more marked than fronted questions.

In general, the fact that wh-questions allow for different syntactic strategies in many languages makes them a suitable case study for the acquisition of variation. There are still

ongoing debates about some aspects of the wh-in-situ variant, such as its syntactic movement status (covert versus no movement) and category (declarative versus interrogative). While further research on these topics would help advance our understanding of wh-in-situ questions, they are not explored here further given the focus of this work on acquisition.

Chapter 3: The acquisition of English wh-question variation

In Chapter 2, I have established that there is variation in English wh-questions. Interestingly enough, such variation is often ignored in the language acquisition literature despite wh-questions being a prominent topic of discussion. This chapter seeks to fill that gap through a series of corpus analyses and behavioral experiments to study the acquisition of English wh-variation.

Since there is a lack of acquisition study on wh-variation in English, I start by reviewing research in languages that share some similarities with English, namely, French and Brazilian Portuguese (section 3.1.). The next two sections (3.2. and 3.3.) present corpus analyses of English wh-question variation in child-directed speech and child production, where the distribution and usage contexts of the wh-variants were investigated. To confirm the results as well as control for factors that could have biased the corpus analyses (e.g., the discourse-pragmatic contexts in which children can produce PQs), a behavioral experiment with 3;06-to-5;06 year-old children was conducted (section 3.4.). The experiment included a comprehension component and a production component, and was designed to elicit children's preference of wh-variant given an appropriate context.

3.1. Previous research on the acquisition of wh-question variation

3.1.1. French

One prominent hypothesis on the acquisition of French wh-questions is the Derivational Complexity Hypothesis (DCH) developed by Jakubowicz (2005, 2011), which draws on the idea of structural economy (Chomsky, 1995). This is an influential

hypothesis in the acquisition of syntax that is frequently discussed in the study of wh-questions acquisition (e.g., Yuan, 2015; Durrleman, Marinis, & Franck, 2016; Prévost, Strik, & Tuller, 2014; Hopp, Putnam, & Vosburg, 2019) and has been extended to other syntactic phenomena (e,g., acquisition of differential object marking: e.g., Cuza et al., 2019; acquisition of object and quantitative pronouns: e.g., van Hout, Veenstra, & Berends, 2011).

The hypothesis suggests that derivational complexity conditions the course of syntactic acquisition. The derivational complexity metric is defined as follows:

(ii) A. Merging α_i n times gives rise to a less complex derivation than merging α_i $(n + 1)$ times.

B. Internal Merge of α gives rise to a less complex derivation than Internal Merge of $\alpha + \beta$.

In particular, the DCH predicts that children will avoid structures involving more syntactic operations unless those are obligatorily required, resulting in a single syntactic option.

In support of the hypothesis, Jakubowicz (2011) shows that children with Specific Language Impairment tend to prefer wh-in-situ over fronted wh-questions compared to their typically developing peers at the same age. However, the empirical picture among typically-developing children is complicated, as reports on their acquisition of French wh-questions have been messy and vary across studies. For example, with regard to children's input, Becker and Gotowski (2015) report a wh-in-situ rate of 16.6% in child-directed speech, while our own corpus study of 2 children finds a rate of 55%. With regard to children's production, there is inconsistency both within and across two methods of data collection: spontaneous speech (corpus analysis) and elicited tasks. In studies relying on

corpus analysis, Crisma (1992) finds that Phillippe (from the Leveillé corpus) did not produce any wh-in-situ until 2;6 and had a much higher rate of fronted wh-question production compared to wh-in-situ. Palasis, Faure, & Lavigne (2019) report a comparable rate of wh-in-situ and fronted wh-questions (not counting clefted questions exemplified in (26e) in Chapter 2). On the other hand, Hamann (2006) finds a strong preference for wh-in-situ. Our research team's corpus study (Cong, 2021) of two children finds a similar overall preference for wh-in-situ at roughly 60%, but it is worth noticing that children start with fronted questions early on and wh-in-situ questions only come in later. In elicited production tasks, children tend to show a preference for fronted over in-situ wh-questions, though the degree of preference varies across studies. Table 1 summarizes the general distribution of French-speaking children's production of wh-questions from both experimental and corpus studies.

Table 1a. Elicited production of French wh-questions (experiments)

	Wh-in-situ[8]	**Fronted Wh**	**Age range**
Gotowski (2017)	12%	51%	3;09 – 5;08
Cronel-Ohayon (2004)	24.2%	64.6%	4;0 – 6;0
Prévost et al. (2017)	25.9%	74.1%	4;0 – 4;05
Strik (2007)[9]	~ 20%	no report	4;0 – 4;06

Table 1b. Spontaneous production of French wh-questions (corpus studies)

	Wh-in-situ	**Fronted Wh**	**Age range**
Crisma (1992)	26.19%	73.81%	2;01 – 2;07
Hamann (2006)	80.04%	19.96%	1;08 – 2;09
Palasis et al. (2019)	43.8%	42.7%	2;06 – 4;11
Our study (Cong, 2021)	90.48%	9.52%	2;0 – 4;0

Gotowski (2017) proposes that the lower frequency of wh-in-situ in elicited tasks compared to that in corpus studies is due to wh-in-situ being more informal. In the same vein, I hypothesize that the "preference" for wh-in-situ in children's spontaneous speech is due to an inherent bias for discourse-given contexts in child-adult interaction. To confirm this, we[10] went through the Paris corpus (Morgenstern & Parisse, 2007) and identified wh-question utterances produced by the child between 01;10 and 04;01. We also randomly extracted 10 child-directed utterances produced by the parents for each month between

[8] Zuckerman & Hulk (2001) report a lower production rate of French in-situ questions (6%, after omission of outliers (n=5), the rate drops to 3%). The very low level of wh-in-situ questions may at least partly be the consequence of the method of elicitation in which an indirect question with a clause initial wh-phrase and no inversion was used as a prompt (*Je veux savoir où il est allé* – "I want to know where he went" despite the fact that one possible answer is the direct question *Où il est allé?* "where did he go?", which is likely to have inflated the proportion of fronted wh without inversion (89%).

[9] The elicited production result (~20% wh-in-situ) is culled from several sources (Strik, 2007; 2008) and is an estimate based on limited text descriptions of the results. A pilot study (reported in Strik, 2008) is not reported because the elicitation method was similar to that used by Zuckerman & Hulk (2001).

[10] I thank the Language Acquisition lab members, Géraldine Legendre and Renee Cong, for their contribution to this corpus analysis.

01;10 and 04;01. For each of these utterances, we studied the context in which the question was used and a native speaker coded them as either discourse-given or discourse-new. In general, there is a significantly higher percentage of discourse-given contexts for wh-questions in both child production (145 utterances have discourse-given contexts out of 170 wh-utterances, or 85.3%) and child-directed speech (131 out of 192 utterances, or 68.2%). The differences between discourse contexts in natural child-adult interaction and in elicited tasks are likely one of the reasons behind the differences in results reported across studies.

3.1.2. Brazilian Portuguese

The acquisition studies on wh-variation in Brazilian Portuguese share a similar problem with the French studies: there is inconsistency across studies. Lessa-de-Oliveira (2003) reports a case study of one child whose most frequent question type in her input was wh-in-situ (53,5%); yet, she produced mostly fronted wh (70.2%), suggesting that frequency is not the sole factor motivating a child's preference for a variant. However, Sikansi (1999) reports only 3.75% of wh-in-situ in child input in another case study, making it more difficult to form an overall impression about the distribution of wh-variants in Brazilian Portuguese. It is not yet clear which variant is the higher-frequency one.

In an experiment controlling for discourse factors, Vieira & Grolla (2020) find that children display a strong preference for fronted questions compared to adults, even in contexts allowing both variants. In particular, in non-prominent Common Ground contexts, adults prefer fronted questions (66.5% production rate) over wh-in-situ (33.5% production rate). When given prominent Common Ground contexts, adults used fronted and wh-in-situ equally (50.5% - 49.5%). However, children consistently produced more fronted

questions in both non-prominent and prominent Common Ground contexts (84.3% and 74%, respectively). This tendency to overuse one variant over another seems to be in line with other cases of children's regularization when faced with multiple variants.

3.1.3. Interim Summary

From Chapter 1, we know that children have a tendency to regularize to a dominant variant during the acquisition of variation. This result is somewhat replicated in the acquisition of wh-variation in French and Brazilian Portuguese. In general, at least in elicited tasks, children demonstrate a preference for fronted questions compared to in-situ questions, although fronted questions may not always be more dominant in terms of frequency. I will now turn to the acquisition of English wh-questions, starting with corpus analyses of wh-questions in children's input (3.2.) and in their own production (3.3.).

3.2. A corpus analysis of wh-question variation in child-directed speech

The presence of wh-in-situ questions in English child-directed speech is occasionally noted in the child language acquisition literature: for example, Becker & Gotowski (2015) report that 16% of all wh-questions produced by adults in Eve's data (Brown, 1973) are wh-in-situ, and Gotowski (2017) similarly reports 22% of wh-in-situ questions in Adam's input data. However, to the best of our knowledge, there has been no attempt to differentiate between in-situ EQs and in-situ PQs. The in-situ structures are typically all assumed to be EQs. Given that PQs and EQs target different information, it is crucial to evaluate their respective presence in child-directed speech. I conducted my own corpus analysis to offer a more accurate picture of wh-in-situ English questions in this context.

To get an estimation of the percentage of wh-in-situ questions in the input, one data file for each month between 2;0 and 4;0 years of age of 10 children was randomly selected, including Sarah and Adam: Brown corpus (Brown, 1973), Shem: Clark corpus (Clark, 1978); Trevor: Demetras corpus (Demetras, 1989); Abe: Kuczaj corpus (Kuczaj, 1977); Lily and William: Providence corpus (Demuth, Culbertson, & Alter, 2006); Naomi: Sachs corpus (Sachs, 1983); Roman: Weist corpus (Weist & Zevenbergen, 2008), Laura: Braunwald corpus (Braunwald, 1985).

Using the CLAN tool (MacWhinney, 2000), all wh-questions from child-directed speech in the selected files were extracted. Since we are only interested in cases in which an in-situ utterance is possible, I further excluded subject *who/what* questions (whose analysis is ambiguous between in-situ and vacuous fronting), embedded wh-questions, "what-if", "how come", and "what about" questions, and limited the search to only main clause questions, resulting in a total of 9039 questions. Questions that do not carry any other piece of information besides the wh-phrase (e.g., "*now what?*", "*because why?*", "*so what?*", "*for what?*") or expressions that are not actually used as genuine sentential questions such as "*you know what?*" were also excluded. From the extracted data, a total of 1361 in-situ questions were recovered, which take up 15.1% of all main clause questions in this sample of child-directed speech. The percentage of in-situ questions in the input varied among children. Out of 10 children, 2 received more than 20% of in-situ questions, 5 received between 10% and 20%, and 3 received less than 10%.

To classify whether each in-situ question is a probing or an echo question, the surrounding utterances were examined in close detail. For example, although it is not clear enough to tell if the in-situ question in (*51*a) "*It's a what*?" is an EQ based on the prior

utterances alone (as the structure of the question does not match exactly with what the child said previously), we can rely on the mother's response to classify this wh-question as an EQ. When there was not enough information to uniquely infer whether a question should be an instance of EQ or PQ, it was regarded as ambiguous. (51b) is an instance of an ambiguous in-situ question. The adult and the child were looking at a book. The adult's in-situ question could be an EQ given that its structure matches with the previous utterance of the child. However, the child's answer was not a repetition or clarification; moreover, the fact that the adult was asking the child about details in a book also suggests that the adult knew the answer, and this could be a follow-up question to lead the child to describe the scenario in the book.

(51) a. (excerpt from Adam, Brown corpus)
Cousin[11]: is it a square?
Child: no square, is clown.
Cousin: *it's a WHAT?*
Child: Mommy
Mother: it's a clown, he said.

b. (excerpt from Roman, Weist corpus)
Adult: Look at these bunnies. What do you think is happening?
Child: They are gonna catch […].
Adult: They're gonna catch what?
Child: They're gonna hide in creek.

EQs appear more frequently than PQs in phrasal questions (e.g., "*did what?"* or "*little what?"*). However, when considering only full sentential questions (e.g., "*it is a what?"*), PQs appear more frequently (54.6%) than EQs (33.2%). The results are summarized in Table 2. In general, children gave appropriate answers to these questions, indicating that they understood these structures. In particular, children responded

[11] The "cousin" is an older child only producing adult-like utterances throughout the corpus.

appropriately to EQs about 90% of the time, though occasionally (about 10% of the time) ignored the question and gave no answer.

Table 2. The distribution of in-situ wh-questions in child-directed speech.

	Full sentential	**Phrasal**	**AVERAGE**
EQ	227 (33.1%)	290 (43%)	37.98%
PQ	369 (53.7%)	242 (35.9%)	44.89%
Ambiguous	91 (13.2%)	142 (21.1%)	17.13%
Total	687	674	100%

Finally, note that caregivers frequently alternate between PQs and fronted questions and use them in child-directed speech as if they are interchangeable, as shown in the example from the Weist corpus below. This further supports the claim that PQs and fronted information-seeking questions are closely related, pragmatically speaking, at least in child-directed speech.

(52) Father: hey Roman, if the dinosaur roars what's the baby gonna do?
Child: it gonna roar and it's gonna say like this (roar).
Father: yeah but if the dinosaur roars the baby is gonna be what?
Child: scared.
Father: no the baby's scared what's it gonna do?
Child: it gonna eat the thing.
Father: no no the baby, the baby's gonna what?

3.3. A corpus analysis of wh-question variation in child production

The search process for wh-in-situ questions in spontaneous child production was similar. Data files for each month between 2;0 and 4;0 years of age of the same 10 children were randomly selected, using the same inclusion and exclusion criteria. Our search returned a total of 10241 wh-questions, 407 (3.9 %) of which were in-situ wh-questions.

However, most of these in-situ utterances were not genuine questions but corresponded to either a fixed expression such as "*for what?*" (31 counts, 7.6%) or an expression that was not actually intended as an information-seeking or echo question such as "*you know what?*" or *"guess what?"* (266 counts, 65.4%). Children sometimes also asked questions and answered them themselves (e.g., "*they buy some more scrambled eggs. Three what? Three scrambled eggs*"). Such utterances were excluded from the analysis, leaving 75 in-situ questions (i.e., less than 1% of all wh-question production). The final result is summarized in Table 3. It is worth noting that the majority (52 counts, 70%) of these wh-in-situ utterances came from a single child, Adam, possibly due to his adopting a unique variant. The rest of the children produced rather few in-situ questions.

Table 3. The distribution of in-situ questions in child production.

	Full sentential	**Phrasal**	**AVERAGE**
EQ	2 (16.7%)	14 (22.2%)	21.3%
PQ	6 (50%)	21 (33.3%)	36%
Ambiguous	4 (33.3%)	28 (44.5%)	42.67%
Total	12	63	100%

Many of the in-situ questions were ambiguous (e.g., Mom: "*that's ocean*" – Child: "*ocean what?"),* as it is not clear whether the child was simply repeating the last word his mom said (despite adding a wh-phrase) or s/he was genuinely asking for clarification (the mother usually did not give a clarifying answer). Some of the utterances were clear instances of EQs:

(53) Mother: He was talking about President Kennedy.
Child: Talking about WHAT?

There were also clear instances of PQs[12]. For example, in the occurrence below, in order to ask his mother about the new object that wasn't mentioned in the text previously, the child used an in-situ question:

(54) Child: Mommy, this is a what?
Child: It's a what?
Mother: Paper punch.

Summing up, out of all genuine in-situ questions, children produced more PQs than EQs. However, the number of such utterances is very small – less than 1% of all wh-question utterances, and 70% of them were produced by a single child. Overall, our results confirm the claim that children rarely produce in-situ wh-questions (Valian & Casey, 2003; Becker & Gotowski, 2015).

3.4. Experiments on children's comprehension and production of wh-question variation

Our corpus analyses show that children rarely produce PQs spontaneously, however, they are able to respond appropriately when adults use PQs. A comprehension experiment was conducted to confirm that children accept and understand PQs as information-seeking questions (comprehension study). I then followed with a production experiment to see if there is a preference in children's production of wh-variants, given

[12] Adam frequently produced a PQ immediately after a fronted wh-question, most likely mirroring the adult behavior mentioned in (6). However, adults typically rephrase the original question into an in-situ PQ only if the child fails to answer the fronted one. Adam, on the other hand, did not wait for a response. This suggests that Adam may have used PQs in a different way compared to adults, as he asked these questions without knowing the answer.

(i) Child: What is that?
Child: It's a what?
Researcher: I don't know what it is, do you?

their understanding of PQs (from the previous experiment) and an appropriate pragmatic setting.

3.4.1. Experiment 1: Comprehension of wh-in-situ

Given the assumption in the language acquisition literature that English-speaking children only understand in-situ wh-questions as EQs (e.g., Takahashi, 1991; Becker & Gotowski, 2015), the comprehension task sought to investigate whether children would be able to differentiate these two types of wh-in-situ (i.e., give repetition to EQs and new information to PQs). I leave out fronted wh-questions in this task because their inclusion would lead to a mismatch in conditions (three types of wh-questions but only two types of answers, repetition versus new information), which can induce a response bias. Note that young children as young as 20 months of age can already demonstrate above chance accuracy in simple wh-questions comprehension tasks (Seidl, Hollich, & Jusczyk, 2013).

Method

Participant

Twenty children were recruited in Baltimore for the study. All of them were native English speakers and were tested in person. The mean age of the children was 4;01 (range: 3;06 – 5;06; 7 boys, 13 girls). Of them, one child was excluded due to an unusually high number of irrelevant answers, and two children were excluded due to failure to follow instructions.

Fourteen adults were additionally recruited to serve as a control group. Of them, two were excluded because they were outliers, i.e., their scores were three standard deviations away from the mean score of the sample. This left us with twelve adult

participants (age range: 19 – 24; 4 males). All of them were students at Johns Hopkins University and were tested in person.

Material

Task design. The context of twelve scenarios making up the experiment was explicitly specified as a classroom-like setting involving a participant, a storyteller (research assistant), and an alien classmate (the experimenter). The role of the alien was to comment on the stories as the storyteller told them. Each scenario led to a target question. In total, there were twelve wh-in-situ questions (six PQs and six EQs). Of them, four were object "what"-questions, another four were object "who"-questions, and four were "where"-questions; each wh-word appeared twice in PQs and EQs. The questions are reported in Appendix A. No subject wh-question was included in the experiment, given the ambiguity of analysis (in-situ/vacuous fronting).

Each question had three possible answers, including a target, a non-target, and a wrong/irrelevant answer. In each scenario, the two characters Bill and Jill would pass by an event but did not get to observe its full development. They talked to each other about the event, with one character saying: "I wonder what happened". The alien classmate then turned to the participant and whispered what he thought had happened. The alien's opinions, however, always violated Grice's Maxim of Quantity by being under-informative. The alien would give a description that matched both the target answer and a non-target answer, essentially narrowing down the choices from three to two, but not enough to uniquely identify the target answer. Half of the time, the storyteller was able to hear the alien and acknowledged his answer, and he would turn to the participant to ask for his/her own answer using a wh-in-situ structure (PQ condition). Half of the time, the

storyteller noticed the alien was saying something but could not hear it clearly, and he asked for the participant's help for clarification using a wh-in-situ echo question (EQ condition).

A sample scenario with illustrations is provided below. In an echo trial, the target answer would be "*the white building*", the non-target answer would be "*the hospital*", and the irrelevant answer would be "*the apartment/the library*". In a probe trial, the target answer would be "*the hospital*", the (under-informative) non-target answer would be "*the white building*", and the irrelevant answer would be "*the apartment/the library*". Assuming that 4 year-olds cannot read yet, a "bookstore" drawing was chosen to represent the "library" to maximize illustration as "library" drawings are typically a generic building with no books shown.

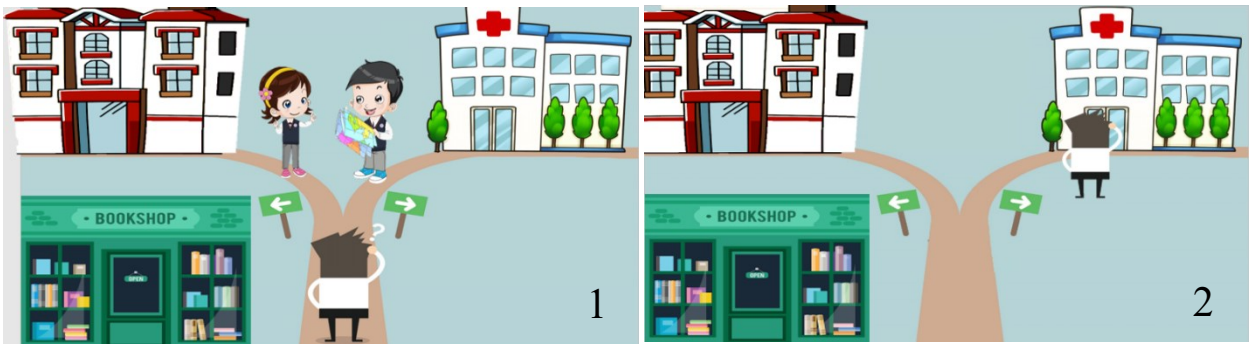

Figure 3. Trial example in the comprehension task

Billy and Jilly are standing in the middle of the road. To their right is a hospital, to their left is an apartment, and going down is a library[13]. While waiting for Billy to read the map, Jilly sees a man appear. He looks at all the buildings carefully, as if he is trying to decide which one to go to. However, by now Billy has figured out

[13] In the experiment, animation effects were used to display the building one by one so that it was clear to the child which building was which.

the map: “Let’s go up that way”, he says, and the two kids walk away. After they have gone, the man finally walks into one of the buildings. But Billy and Jilly do not see this. On their way, they talk to each other about the man. Jilly says: “I wonder where the man went”. Alien puppet: “I think the man went to the white building”. Storyteller: “the man went where/WHERE?”

Two question fillers were included to keep children engaged. In each filler, three possible choices were also introduced, however, there was no right answer and no visual or auditory cues about which option should be chosen. For example, in a filler, three types of drinks were shown on the screen as the storyteller said to the participant: “Bill is very thirsty on this hot summer day. Can you choose a drink for Bill?”.

Challenges in design One challenge with this design is its potential lack of a control condition (i.e., fronted questions). This is because both PQs and fronted questions are information-seeking questions. Thus, an experiment with equal trials of each type of question (EQs, PQs, and fronted questions) would result in twice as many new-information answers as repetition answers and can induce bias in the response, as participants may respond more or entirely with new-information answers. In order to keep an equal rate of the two answer types, the number of PQs and fronted questions would need to be half of EQs. This can also lead to detrimental consequences of having mismatched rates of question types, not to mention that it will lengthen the experiment and tire out the child. A 12-trial task in an equal-answer setup (6 EQs, 3 PQs, 3 fronted questions) has too few PQs (our target question) to be useful, yet simply increasing the number of PQs to 5 would result in 20 trials (10 EQs, 5 PQs, 5 fronted questions). A between-subjects study in which

each group is exposed to a type of question could potentially address this issue, but it would not tap into the ability to differentiate between different in-situ questions as well as a within-subjects design.

Pragmatic considerations It is expected that participants will generally prefer to be informative and precise (Grice, 1975) and thus would give the target answer over the alien's under-informative answer in a PQ trial. At the same time, in an EQ trial, participants should know that the storyteller is interested in what the alien has said instead of their own opinion, thus they can only repeat the alien's answer even though they may perceive it as being under-informative.

Typically, the original speaker (i.e., the alien) should be the one to respond to the storyteller's EQ instead of a third party (the participant). Therefore, I opted to maximize pragmatic plausibility by adding more details to the setup. In the first practice echo trial, the storyteller reminded the participants that the alien liked to voice his (unsolicited) opinions, but only to the participant. The participant was told that the alien was afraid of the storyteller and only talked to the participants, refusing to answer the storyteller's questions directly. Sometimes the storyteller could hear the alien and sometimes not, but the storyteller was interested in everyone's opinion and specifically wanted to know what the alien said. The alien would hide behind the participant when he heard the storyteller ask the EQ. The storyteller then asked the participants to help the alien out whenever he got shy. This EQ scenario was designed to reflect a typical experience in a classroom setting – the teacher tells a story and asks a question, and someone in the back responds but not loud enough for the teacher to fully hear the answer. The teacher then asks an EQ and someone else in the front who heard the answer can repeat it for the teacher. Most child

participants responded correctly in their first attempt, with some responding correctly to the EQ even before the storyteller started explaining further about the alien, suggesting that they understood the objective of an EQ.

Pragmatic cues, including hand gestures (gesturing towards the participant in a probe trial, or putting a hand to ear in an echo trial) and cue words appropriate to a classroom setting ("*Class,* [PQ]" or "*Hmm*, [EQ]"), were included in certain fixed trials to increase the pragmatic plausibility of the task and make the questions more natural-sounding. The cues for each type of question were controlled so that they matched in number (e.g., a hand gesture was used for one PQ and one EQ) and properties (e.g., the cue words had the same length). I will return to the use of such pragmatic cues in the Discussion section.

Prosody of the questions. The questions were not recorded and instead were asked directly by the research assistant playing the storyteller role for pragmatic reasons: it would be pragmatically implausible if the whole story is told by the storyteller in his natural voice, but whenever he asks a question, the question is in a recorded voice played through the computer. Recording the whole story would solve the problem of potential inconsistency, but it would have made the story less engaging and less interesting to young children. Thus, both the story and the questions were set in the storyteller's natural voice. Since prosody served as an important factor that participants could rely on to distinguish the two types of questions (an EQ has an exaggerated pitch accent on the wh-phrase while a PQ does not), I examined the possibility that the storyteller might not reliably produce exactly the same prosody for the same question in every trial and experiment. A post-hoc acoustical analysis was conducted using the Praat software (Boersma & Weenink, 2019), excluding from data

analysis any trials in which the prosody of the question is significantly different from the rest of the sample. Two data points (out of 170) of EQs in the child experiment were thus excluded from the final analysis due to inconsistent prosody. Since each participant provided 10 data points, we ended up with a total of 168 data points for children and 120 data points for adults.

Procedure

The comprehension task took approximately 20 minutes. Participants were explicitly told that they were in a classroom where they would be listening to a story along with an alien "classmate" named Terry. The storyteller was played by a research assistant and the alien classmate puppet was played by the main experimenter. Participants were told at the beginning that the alien was scared of the storyteller, and he would only talk to the participants but not the storyteller. Illustrations of the story were shown on a big TV screen in the testing room. Participants were directed to pay attention to the illustrations.

At the end of each scenario, one cartoon character would raise a problem. The alien would whisper to the participants his thoughts about the problem. Following that, either a wh-in-situ EQ or a wh-in-situ PQ would be asked by the storyteller. When the participants had finished answering the question, the story continued to the next scenario. To keep the child participants engaged in the task, the storyteller rewarded them with a sticker after every four questions.

The first two scenarios of the story were used as practice trials (one with a PQ and one with an EQ). In both practice trials, if the participants responded incorrectly, the storyteller would try to guide them to the target answer by providing hints or suggestions, though never explicitly corrected them by giving out the target answer. Such feedback was

only given in the practice trials. The data from the two practice trials were not included in the analysis, resulting in 10 answers per participant.

Results and discussion

Participants' responses were divided into three categories: target, non-target, and irrelevant/wrong answers. A target answer means that the participants gave the right information to the right type of question (e.g., repeating the alien's answer in an EQ trial). A non-target answer means that the participants gave the right information to the wrong type of questions (e.g., repeating the alien's answer in a PQ trial). An irrelevant/wrong answer means that the participants gave wrong information (e.g., saying the man went to the library when he actually went to the hospital). The distribution of the answers by category is shown in Table 4. Children only gave irrelevant answers 4.8% of the time, which indicates that they were able to understand and pay attention to the story. They also provided almost twice as many target answers (61.4%) as non-target answers (33.7%).

Table 4. Distribution of the answers by category

	Target answers	**Non-target answers**	**Wrong answers**
Children (3;06 – 5;06)	61.4%	33.7%	4.8%
Adults	91.7%	8.3%	0%

However, children were not as good as adult controls in interpreting the intention of the two types of questions. Although the overall adult performance was not perfect (91.7%), more than half (7 out of 12) of the adult participants achieved perfect accuracy. The other 5 adult participants made errors but only towards the end of the task, which

could be a result of a loss of attention due to the task being childish and overly easy for them. In contrast, the performance of the child participants ranged from 50% to 77.78% (or 50% to 80% *target* answers when excluding wrong answers), with none of them ever achieving perfect accuracy. The data is represented in Figure 4.

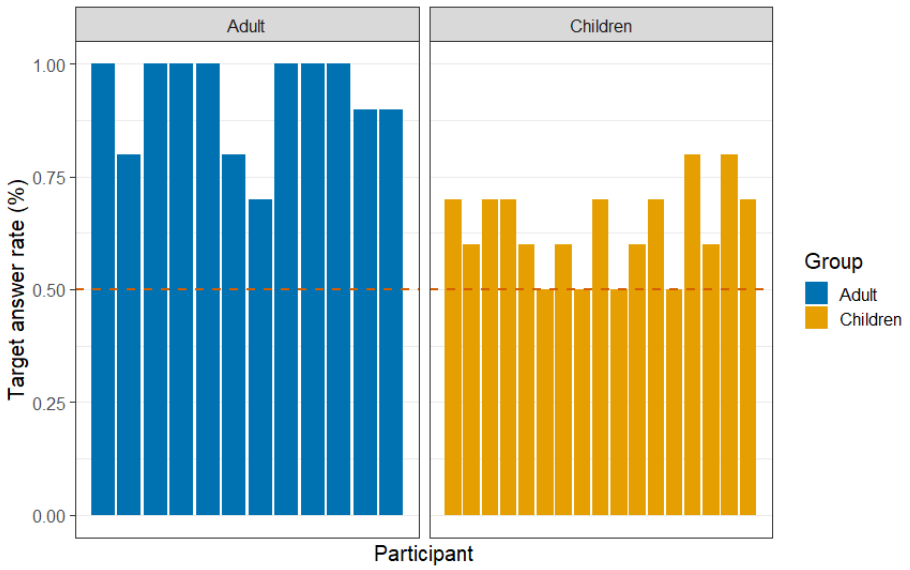

Figure 4. Target answer rate of children and adults

Excluding irrelevant answers, children otherwise correctly produced target answers over non-target ones 64.5% of the time. None of the child participants consistently produced only one type of answer to all 10 questions throughout the experiment. In other words, every child used both types of answers (echo-appropriate and probe-appropriate) at least once. Given the relatively small sample size, non-parametric Wilcoxon tests were conducted. The test showed that children correctly produced the target answer significantly above 50% chance level ($p = 0.001$, effect size = 0.80). However, the results varied within each type of question. Children performed significantly better with PQs than with EQs (M_{PQ} = 76.6% and M_{EQ} = 51.3%, $p < 0.001$, effect size = 0.97) (Table 5). While their accuracy was significantly above 50% chance for PQs ($p < 0.0001$, effect size = 1.2), it was only at chance level for EQs ($p = 0.67$, effect size = 0.1) This disparity was not

observed in the adults' performance. Adults were equally good at inferring the intention of PQs and EQs (Ms = 91.7%). Adults also outperformed children both overall ($p < 0.0001$, effect size = 1.08) and within each type of questions (ps < 0.01).

Table 5. Percentage of choosing target over non-target answers

	PQs	**EQs**	**Overall**
Children (3;06-5;06)	76.6%*	51.3%	64.5%*
Adults	91.7%*	91.7%*	91.7%*

* indicates significance above chance level.

While children seemed to struggle slightly more with object "*who*"-questions, a Kruskal-Wallis test suggests that there was no significant difference in the performance within each subtype of wh-questions of both adults ($H(2) = 2.56$, $p = 0.28$) and children ($H(2) = 1.47$, $p = 0.48$). Figure 5 illustrates this.

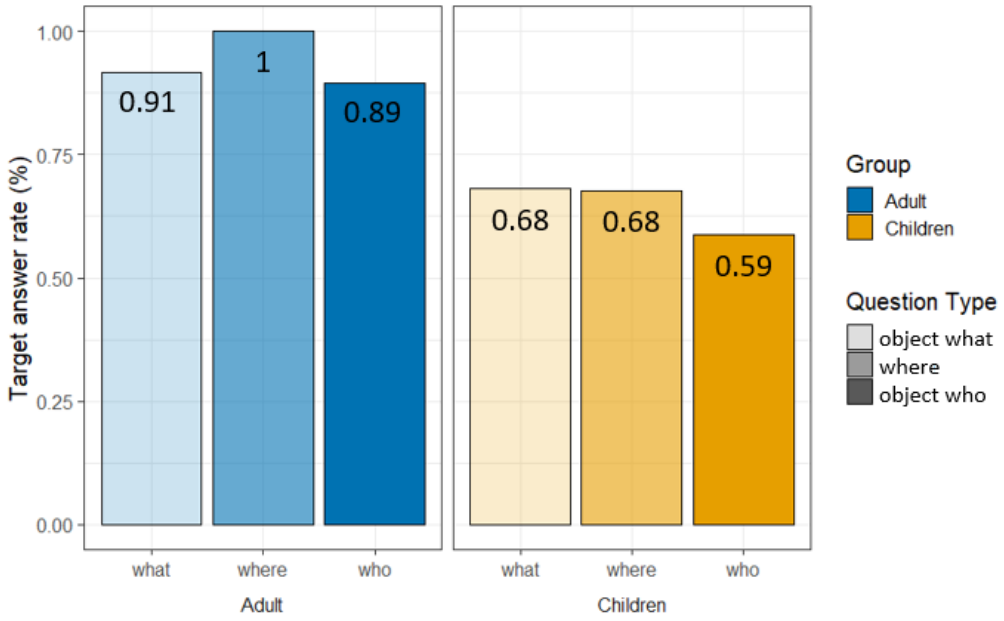

Figure 5. Target answer rate by sub-type of wh-question produced by adults (left) and children (right)

Finally, the child data was submitted to a logistic mixed-effect model using the lme4 package in R (Bates et al., 2015). The dependent variable was the Accuracy of each question. Age, Question Type (probe vs. echo), and Pragmatic Cues (hand gesture, cue words, or none) were included as fixed factors. Participant and Question Item were included as random factors.

There was a significant effect of Question Type ($\beta = 1.29$, SE = 0.37, $p < 0.001$) but no effect for Age (β -0.06, SE = 0.23, $p = 0.80$), hand gesture ($\beta = 0.27$, SE = 0.6, $p = 0.66$), or cue words ($\beta = 0.46$, SE = 0.68, $p = 0.5$). This suggests that the random inclusion of pragmatic cues did not make any trials significantly easier than others.

The results suggest that children are able to make use of prosodic information in language comprehension to differentiate between the two types of in-situ wh-questions. While the extra pragmatic cues that were included to increase the naturalness of the questions may have aided the differentiation process, the prosodic difference between PQs and EQs was the only factor that was consistently present in every trial. The extra pragmatic cues, on the other hand, were not as reliable: some trials consisted of only hand gestures, some consisted of only cue words, some consisted of both, and some consisted of none. Results from the mixed-effect analysis show that there was no effect of pragmatic cues; in other words, questions with extra pragmatic cues were as challenging as those without any such cues. If children couldn't reliably use pragmatic cues to differentiate the two types of questions, it must be that they employed prosodic cues. However, the fact that children overall performed worse than adults suggests that they may not be as sensitive to the prosody of questions as adults are. This is in line with previous studies, which claim that although children are able to use prosodic information in sentence processing, they use

such information less effectively than adults do to infer the intended meaning (e.g., Snedeker, 2008; Ito et al., 2012; Sekerina and Trueswell, 2012, Hupp and Jungers, 2013).

Summary of Experiment 1

The results show that in the comprehension task, children performed above chance level. They were able to provide almost twice as many target as non-target answers. If children (wrongly) assumed that PQs and EQs have a similar intention, we would expect the percentage of target answers to be the same as non-target answers. The significant difference in percentage shows that children, at the very least, recognized that two different types of in-situ wh-questions were asked, and the fact that there was a strong preference for target answers over non-target answers shows that they were able to assign the right intention to the right type of question with moderate accuracy.

3.4.2. Experiment 2: Production of wh-in-situ

While the corpus analysis shows that children do not produce wh-in-situ, this could be because children are rarely in a pragmatically appropriate context for PQs. In this task, children's production of wh-questions was tested in contexts where PQs and fronted information-seeking questions are both acceptable and can be used interchangeably.

Method

Participants

After completing Experiment 1, both children and adult participants were given a 10-minute break in an adjacent room before getting invited back to the testing room for Experiment 2. Note that the comprehension task (Experiment 1) was always completed prior to the production task to ensure uniform prior exposure to wh-in-situ.

Materials

Participants were introduced to an alien character, Beeple. Beeple came to planet Earth to learn about the Earth and its culture. Before Beeple left to return to his planet of origin, it is important to ensure that he had learned enough about Earth. Thus, the participants' task was to ask Beeple multiple questions to quiz his knowledge.

There were twelve trials in total, which included four object "*what*"-, four "*where*", and four object "*who*"- questions. In each trial, participants were shown an illustration of Beeple standing next to certain objects or characters. The experimenter would prompt the participants by saying: "*Let's ask Beeple if he knows about [general description of the object]*". Beeple's responses were pre-recorded. After the participants asked the question, the experimenter would play the recorded audio file. The recorded answers were acoustically modified to sound alien-like. An example of a practice trial with an in-situ question is presented in Figure 6.

Figure 6. A practice trial example in the production task.
Experimenter: "Let's ask Beeple if he knows the word for the food the boy is eating. Let me show you how to do that: Hey Beeple, the boy is eating what?".

The scenarios were designed to match the context in which PQs are used in child-directed speech. Typically, PQs are used when (1) the addresser already has an answer in mind and (2) the addresser is more interested in assessing the addressee's knowledge than

the answer itself. These conditions are satisfied in the production task. First of all, I ensured that children knew the answer to the wh-question by only using simple target objects or characters (e.g. *apple, pizza,* or *mom* etc.). Secondly, the task was set up so that participants were interested in the alien's ability to answer, as they needed to decide whether the alien had learned enough to return to his planet. Thirdly, the participants were placed into an "authority" role: they knew more about English than an alien who was learning human language, and they were encouraged to give the alien feedback ("Good job, Beeple" if the response was correct, or "That's wrong" followed by a correction if it was incorrect). It is important to note that these are contexts in which PQs are often used, but are not meant to strongly favor PQs. Given the interest in testing economy-based accounts, such contexts in which PQs and fronted questions are both acceptable and can be used interchangeably are ideal, as we can tease apart whether a PQ production is primarily motivated by economy preference or by pragmatic constraints.

Procedure

In each scenario, the experimenter instructed the participants to ask Beeple about an object or a person in his surrounding environment. The first two scenarios were used as practice trials. To avoid a strong priming effect on PQs and demonstrate that participants had the freedom to choose the type of question they wanted to use, I included both an in-situ and a fronted wh-question in the practice trials (randomly introduced as questions 1 and 2). After the practice trials, participants were encouraged to produce the questions by themselves, with no feedback or correction given.

Results and discussion

Practice trial performance

The practice trials were similar to a repetition task. The experimenter demonstrated first how to ask a question from Beeple and asked the participants to repeat after him. While adults had no trouble following the instruction and repeating the question the experimenter asked (either in-situ or fronted), several interesting behaviors in children were observed. First, while children had no trouble understanding the repetition request for a fronted question, it took them longer to repeat an in-situ question. Even though the experimenter asked them to simply repeat the question, six children (35%) immediately gave an answer. Second, seven of them "auto-corrected" the in-situ question into a fronted one (e.g., "*where are the kids going?*" even though the experimenter said "*the kids are going where?*") or made a "failure to delete" error when fronting the wh-phrase, resulting in two copies ("*where are the kids going where?*") (see Crain and Nakayama (1987) and Roeper and de Villiers (2011) for more discussion on double marking errors in child production). These observations point to children's preference for fronted wh-questions but also their awareness of in-situ strategies.

Main trials performance

One 3;08 year-old child refused to produce questions by herself and thus her data was not included in the analysis. Not every child participant successfully completed all ten main trials, in particular, one child completed nine and one child only completed eight, resulting in a total of 157 data points. Among child participants, only one child produced PQs (with no mistakes), and only for two trials (2/157 = 1.27%). The remaining child participants consistently used fronted wh-questions. They did make some grammatical

errors in their fronted wh-questions. Common mistakes included auxiliary omission and absence of subject-verb agreement. In addition, one child produced only "*what*"-questions regardless of the scenario (e.g., "*what is the boys playing at?*" instead of "*where*", and "*what is the fairy talking to?*" instead of "*who*"). Overall, the percentage of well-formed fronted wh-questions was 67%, while 33% of the utterances included at least one type of grammatical error. Since the goal of the production task was to test whether children were willing to produce PQs given an appropriate context, I will not discuss further the grammatical errors found in fronted wh-questions.

One adult participant produced PQs for eight out of ten trials and spontaneously commented that he found PQs "easier to produce". However, the remaining eleven adults also consistently preferred fronted wh-questions throughout the production task. In total, there were ten in-situ questions produced by each adult participant. While this number is small, a Chi-squared test with Yates' correction suggests there is a difference in performance between adults and children (X-squared = 6.56, df = 1, p = 0.01). The data is summarized in Table 6.

Table 6. Adults' and children's elicited production of wh-questions

	PQs	**FQs**
Children	2 / 157 (1.27%)	155 / 157 (98.73%)
Adults	10 / 120 (8.3%)	110 / 120 (91.7%)

3.4.3. Discussion

General performance

Besides EQs, in-situ wh-questions asking for new information (PQs) are also present in child-directed speech. This finding challenges previous studies that rely on the assumption that in-situ wh-questions in English can only function as EQs, or that children never hear non-echo wh-in-situ questions in English (Yip and Matthews, 2000, 2007). Our comprehension task further demonstrates that even children as young as 4 years of age are able to differentiate between the two types of questions. Their moderately high accuracy with PQs in the comprehension task (76.6%) suggests that children understand and accept in-situ PQs as information-seeking questions. This directly contradicts a claim commonly found in the child language acquisition literature that children only recognize wh-in-situ as EQs (e.g., Takahashi, 1991; Becker & Gotowski, 2015; Park-Johnson, 2017).

Why did children perform significantly worse with EQs than with PQs? One potential answer suggested by a reviewer [of the published version of this experimental study, Nguyen & Legendre, 2022] is that children in our study were biased to answer according to their beliefs instead of the alien's, given the use of the attitude verb *think* and the infelicity in the alien's response. Previous studies have shown that children tend to evaluate *think* sentences based on reality or their own beliefs instead of others' belief, leading to their poor performance on false belief tasks (e.g., de Villiers, 1995; Papafragou, Cassidy, and Gleitman, 2007). However, I rule out this hypothesis because our task is not a belief evaluation task but a repetition task in the case of EQs. De Villiers and Pyers's longitudinal study (2002) shows that children within the same age range as those in our study do not have difficulty repeating *think* sentences that report false beliefs. By the third

round of the study, when the participants' age range was between 3;07 – 4;05, their performance on such a task was above 90%. It is thus unlikely that the use of the verb *think* in the task is the main reason behind the lower accuracy with EQs. Instead, given the dominance of information-seeking questions in the input, I hypothesize that either children have a bias for more precise answers, or their default interpretation of questions is information-seeking. In the first case, since answers to PQs are more informative, children's failure to respond correctly to EQs may be due to cognitive factors tied to executive control, e.g., children's inability to suppress the (more obvious) answer that they already had in mind (Gualmini et al., 2008). In the latter case, since the majority of questions children are exposed to are questions asking for new information, it is possible that they have a default or bias toward an information-seeking interpretation. To get the non-default interpretation, children would need to rely on additional cues such as prosody. EQs would be more challenging because EQs require children to notice and interpret the prosody correctly, and studies have shown that they do so less effectively than adults (e.g., Snedeker, 2008). Note that children have no issue responding to EQs in spontaneous settings when there is no competing option, as found in our corpus work. Hence the low performance on EQs in our task is not an indication of children's inability to process EQs, but more likely a problem with accessing the right repetition answer when there is a new-information competitor.

Regularizing behavior

In an elicitation task where it is appropriate to use either PQs or fronted questions, which type will children choose? Our results indicate that English-speaking children strongly prefer fronted wh-questions over wh-in-situ despite their good comprehension of

the latter question type. A Chi-squared test suggests a difference in adult regularizing performance versus children.

What factors motivate this preference? Recall that the pragmatic contexts used in our production task were not supposed to strongly favor PQs over fronted questions. Instead, they were contexts in which PQs and fronted questions can be used interchangeably. Adults, who routinely use structures derived by syntactic movement, are not expected to be motivated to go for PQs. However, for children, structural economy-based accounts (e.g., Jakubowicz's DCH) predict that they would have a bias towards simpler constructions as it would alleviate the amount of cognitive resources required to form an utterance. In other words, a structural economy account would expect children to resort to the structurally simpler but pragmatically equivalent in-situ PQs at a stage when they struggle with forming fronted wh-questions (shown by their inversion errors in English and their strong preference for the grammatical no inversion option in French). This prediction was not borne out in either the French literature or our English elicitation tasks, suggesting that structural economy alone is not sufficient to account for our result. I however do not claim that economy plays no role at all in the acquisition of multiple variants, as this result may simply suggest that there is a trade-off between economy and other factors.

Frequency is a potential factor: in English, fronted wh questions appear much more frequently in the input (>70%) compared to information-seeking wh-in-situ. Even though wh-in-situ is structurally simpler, fronted wh-questions are easier to learn as children hear them more. This could serve as a simple explanation for English. However, recall that the cross-linguistic wh-question frequency reports are messy and vary across individuals

(Chapter 2), and generally frequency alone is not sufficient as an explanation. If frequency is the only factor that conditions what variant is preferred by children, we would have trouble explaining the case study reported by Lessa de Oliveira (2003), in which the child showed a preference for the less frequent variant. Moreover, frequency cannot explain the differences between results obtained from elicited studies versus corpus studies in French: a good number of elicited studies uniformly report that children prefer producing fronted wh-questions over wh-in-situ, however, results from corpus studies range from no preference (i.e., equal usage of the two types) to a preference for wh-in-situ. Assuming the 16.6% wh-in-situ rate reported by Becker and Gotowski (2015) for French is correct, we would then need to explain the asymmetry in wh-in-situ production rate between French- and English-speaking children given that the input frequency is roughly the same.

When comparing wh-fronted and wh-in-situ variants, structure (syntactic movement vs no movement) and frequency (high vs low frequency) are not the only factors that differ between them. Important, but much less discussed in the literature, are their respective discourse-pragmatic properties. In English (as well as in French and Brazilian Portuguese), wh-in-situ is more *discourse-marked* than wh-fronting, and can only appear in some restricted contexts while fronted questions can appear in a wider range of contexts. At the base minimum, wh-in-situ needs to satisfy Common Ground requirements: the non-wh portion of the question needs to be discourse-given (though English has additional restrictions for wh-in-situ besides Common Ground, in particular, the addresser usually has authority over the addressee, see Biezma, 2020 for a more detailed discussion of English questions; Hamlaoui, 2011 for French questions; and Vieira & Grolla, 2020 for Brazilian Portuguese questions). Assuming that learners are sensitive to and obey the discourse

restrictions of these wh-variants, learning to use wh-in-situ would still require learning to use wh-fronting for contexts in which the constraints are not satisfied. On the other hand, learning wh-fronting does not necessarily require learning to use wh-in-situ. This is because producing fronted questions when the context is appropriate for wh-in-situ may not be the most economical option but is still acceptable, yet producing wh-in-situ when the context is appropriate for fronted wh-questions would result in comprehending/processing issues. Thus, fronted wh-questions seem to be the more parsimonious option to learn. A discourse-based explanation would thus suggest that learners prefer the unrestricted variant (i.e., the variant that can be used in most contexts) when learning multiple variants, predicting the preference for fronted wh-questions. This hypothesis may also avoid the problems with pure frequency accounts, such as the case study of the Brazilian Portuguese child who prefers fronted questions even though her input is predominantly wh-in-situ. The asymmetry between the production rate of French- and English-speaking children would then be due to French wh-in-situ being less marked than English wh-in-situ. Note that our corpus study of French finds that French-speaking children start out with fronted questions before gradually shifting to wh-in-situ. Hence, the initial acquisition pattern of French-speaking children is similar to English-speaking children: early on when they have not fully mastered the constraints on wh-in-situ, they go with the 'safer choice' or fronted questions; the fact that discourse restrictions of French wh-in-situ can be relaxed (wh-in-situ can occur in out-of-the-blue contexts, Adli, 2006) makes it easier for children to learn and 'transition' to this type, compared to English. A discourse-based account thus could work in tandem with a frequency-based account to potentially resolve problems of inconsistency in frequency reports mentioned above.

In sum, structural economy or frequency alone is not sufficient to account for the acquisition patterns of multiple wh-variants cross-linguistically; nor is their combination. I propose that a crucial factor (that may work in tandem with frequency) is discourse markedness. More generally, investigating the role of discourse restrictions in the acquisition of multiple variants can shed new insights into factors that facilitate acquisition, as this factor still remains understudied in the domain of language acquisition.

3.5. Chapter 3 summary

This chapter has presented results from corpus analyses and behavioral experiments on the acquisition of English wh-questions in children between 3;06 and 5;06 years old. Corpus analyses suggest that children are exposed to both fronted wh-questions and in-situ wh-questions (PQs), though the frequency distribution of these variants can vary across households. A comprehension experiment further confirms that children correctly understand PQs as information-seeking questions and can separate them from EQs. However, spontaneous production of children does not contain many instances of wh-in-situ, and even when provided with pragmatically appropriate contexts for PQs in the production experiment, children still show a strong preference for fronted wh-questions. Such preference for fronted wh-question is also found cross-linguistically in French and Brazilian Portuguese acquisition studies.

The preference for fronted questions is unexpected under the traditional economy account, which predicts that children would prefer the structurally simpler wh-in-situ variant. Frequency accounts can explain the English data. However, the inconsistency in cross-linguistic frequency reports and the asymmetry between French-speaking children's

production in spontaneous speech and in elicited speech suggest that frequency is not the sole factor motivating this preference. I proposed that other linguistic factors, such as discourse markedness, may also play a role in conditioning children's preferences when being exposed to multiple variants. In the next chapter (Chapter 4), I present additional evidence from behavioral experiments using artificial language learning to demonstrate the role of discourse factors in language learning.

Chapter 4: Regularization and its conditioning factors

In the brief review of previous studies on regularization in Section 1.3., we have seen that when children regularize, they tend to regularize to the more frequent variant. This happens with both linguistic (e.g., Hudson Kam & Newport, 2009; Schwab et al., 2018) and non-linguistic stimuli (e.g., Derks & Paclisanu, 1967). In the wh-production experiment reported in Chapter 3, we again see a similar pattern: 3-to-5 year-old children produce the more frequent English fronted wh-questions exclusively, even though they demonstrate a good understanding of wh-in-situ.

However, cross-linguistic data suggests that there is more to regularization than just a frequency effect. Firstly, there are cases when a fronted wh-question is not the more frequent variant in Brazilian Portuguese child-directed speech, yet the child still shows a preference for it in production. Secondly, frequency is not an independent property, but is typically paired with another linguistic feature. For example, in English, French, and Brazilian Portuguese, the more frequent fronted wh-question is also the syntactically more complex but discourse-unmarked variant, while the less frequent wh-in-situ variant is syntactically simpler yet more pragmatically/contextually constrained. Intuitively, syntactic complexity does not seem to be the right explanation. Children should not go for the syntactically more complex structure, which arguably requires more cognitive resources. Thus, I turn to the next common characteristic: discourse-unmarkedness.

In all these three languages, a fronted question is the unmarked or neutral variant (compared to wh-in-situ), in the sense that it can be used in almost any context. Wh-in-situ, on the other hand, tends to have pragmatic or discourse constraints such as givenness. Thus, a speaker regularizing to fronted questions would minimize the chance of violating

discourse constraints and maximize the chance of efficient communication. In that sense, discourse markedness can be a potential factor that conditions regularization.

There are, however, two main problems with a discourse-based hypothesis. The first problem is that this hypothesis relies on a strong assumption about children's sensitivity to discourse constraints. There are a few studies proposing that children are indeed sensitive to discourse constraints. For example, Aravind et al. (2018) show that when clefts (e.g. "It is the cat that the dog chased") – a structure that typically requires referential givenness – are used in contexts that do not satisfy such constraints, children struggle more with comprehension. However, generally speaking, children's discourse-pragmatic abilities are relatively understudied. The second problem is that in natural languages, discourse restrictions tend to go together with frequency. A variant with stricter discourse restrictions entails that it can be used in fewer contexts than a variant with fewer or no restrictions. Thus, a discourse-marked variant like wh-in-situ is typically less frequent than the discourse-unmarked fronted wh-question. This seems to be true generally across the three languages[14]. Given that frequency itself can play a role in conditioning the preferred variant, it is unclear whether the preference for fronted wh-questions is due to higher frequency, discourse-unmarkedness, or both.

In this chapter, I explore the role of discourse markedness in regularization. Besides informing us more about the acquisition of wh-questions, this work can also further our understanding of regularization. While there are a good number of studies on the

[14] Although there are certain contexts, including but not limited to child-directed speech, in which the Common Ground-based discourse requirement for wh-in-situ is typically satisfied, allowing wh-in-situ to be used more frequently.

regularization of linguistic stimuli (e.g., Hudson Kam & Newport, 2005, 2009; Austin, 2010; Schwab et al., 2018), regularization has often been discussed in terms of general factors like frequency or consistency and less so in terms of language-specific factors. Our results thus contribute to the discussion about whether regularization is domain-general or domain-specific.

In the next section (4.1.), I present an artificial language learning (ALL) experiment that aims to study the role of discourse markedness. The results show that when frequency is held constant, learners prefer producing the discourse-unmarked variant over the discourse-marked variant. Section 4.2. presents a follow-up study that investigates whether a similar pattern is found between a strictly-constrained marked variant and a loosely-constrained marked variant. Finally, the context design of the ALL experiments is validated in section 4.3.

4.1. Experiment 3: Discourse markedness in learning variation in artificial grammar

4.1.1. Why Artificial Language Learning (ALL)?

Among the studies on variation in the input and regularization, many of the most influential works rely on ALL experiments, (e.g., Hudson Kam & Newport (2005, 2009; Fedzechkina et al., 2012, 2017; Culbertson et al., 2012). ALL experiments are a favorable choice as they allow us to tap into participants' language learning abilities. More importantly, ALL experiments enable researchers to control and manipulate different factors when designing the language. Researchers can also ensure that participants are exposed to the same amount of variation, which is rather tricky in natural language: given

that speakers have a choice of variants to use, there will certainly be cross-speaker differences in how frequently a variant is used. For example, in the case of wh-questions, while parents uniformly use fronted questions more frequently, the rate of wh-in-situ PQs varies from 5% to 20% in our corpus analyses. While this does not affect our general characterization of PQs as being the less frequent variant, however, it does raise the question about the minimum threshold before something is learned or used if frequency is the deciding factor. Finally, to the best of my knowledge, most of the ALL experimental work done on the acquisition of variation has focused on lexical and/or morphological acquisition. There are rather few ALL studies that look into syntactic variation, with a few notable exceptions, e.g., Culbertson et al., 2012. Thus, our work contributes not only to the literature on variation acquisition and regularization but also to the ALL literature.

Still, ALL studies are not free of issues. Given the limited time course of an experiment, learners are typically only exposed to a small aspect of the language (e.g., morphological markings, transitive versus intransitive verbs, gender system, etc.). This renders the learning experience less realistic because learners do not experience potential interference or facilitation from other non-targeted structures in the language. Another issue frequently discussed is the participants in ALL experiments. While a number of studies (e.g., Hudson Kam & Newport, 2005, 2009; Schwab et al., 2018) have suggested that adults and children may employ different learning strategies, whereby adults probability match while children regularize, Fedzechkina et al., (2012, 2017) propose that adult behavior actually depends on the design of the experiment. Most of the studies that include both adults and children as participants may be too simple for adults, which allow them to easily learn everything and reproduce the input. Fedzechkina et al. show that if the

input languages are complex enough, adult learners also deviate from the input and show regularizing behavior. This results in a dilemma: an ALL experiment that is learnable for children may fail to induce regularization behavior from adults due to its simplicity, yet an ALL experiment that is complex enough for adults may not be suitable for children and hence does not allow us to directly compare the learning outcome between the two populations.

Nevertheless, I follow Fedzechkina et al. here in proposing that adults and children may both resort to regularizing as a learning strategy when faced with variation, if the input is cognitively taxing enough. The ALL experiment described in this section is intended for adult learners, however, a modified version in the future may be suitable for older children (8 years old and above).

4.1.2. The experiment

Methods

Participants

Forty-two adult participants (age range from 19 to 36, mean age = 28.2, female = 21) were recruited on the Prolific platform. All of the participants reported being raised as an English monolingual only and knowing no second language.

Besides task-based measurements, participants' browser activities were also recorded. The experiment automatically recorded the number of times as well as the inactive duration when participants moved away from the experiment (by opening a new window or a browser tab). Eight participants with more than 5 minutes of total inactive time throughout the experiment were excluded, leaving us with 34 data points.

Materials

The artificial language contains 4 verbs, 8 nouns, and 2 wh-words (Table 7). Participants were first introduced to the "native" speakers of the language. They were instructed to learn the language by observing how the native speakers used the language in their daily life. Each sentence was shown as part of a conversation between 2 native speakers.

Table 7. Lexicon of the artificial language

Nouns	Verbs	Wh-words
Szet	gomey	gat
pham	kamo	gwo
thaang	zapit	
ghy	patta	
shom		
mun		
plet		
bhob		

Participants were exposed to declarative and interrogative sentences in the language. The sentences described or asked about simple transitive events in which a human character performed actions on another human character or an object. The events included 4 possible actions (kicking, hitting, carrying, and driving), 6 possible characters (police officer, bride, student, nurse, farmer, and chef), and 2 possible objects (ball and boxes). Each event was depicted as if it was a scene in a movie. Participants were first presented a close-up view of the event, followed by a full view that showed one of the native speakers watching the event on a big TV screen. There were two conditions: prominent Common Ground and non-prominent Common Ground. Common Ground (CG)

is defined here as having shared, equal access to the source of information (Clark & Brown, 2006).

For declaratives in the non-prominent CG context (Figure 7a), the speaker watches the event on the screen alone while describing the action to the listener via the phone. The listener does not see the event directly and relies on the speaker as the only source of information. For declaratives in the prominent CG context (Figure 7b), the speaker and the listener are watching the event on the big screen together while the speaker describes the action to the listener.

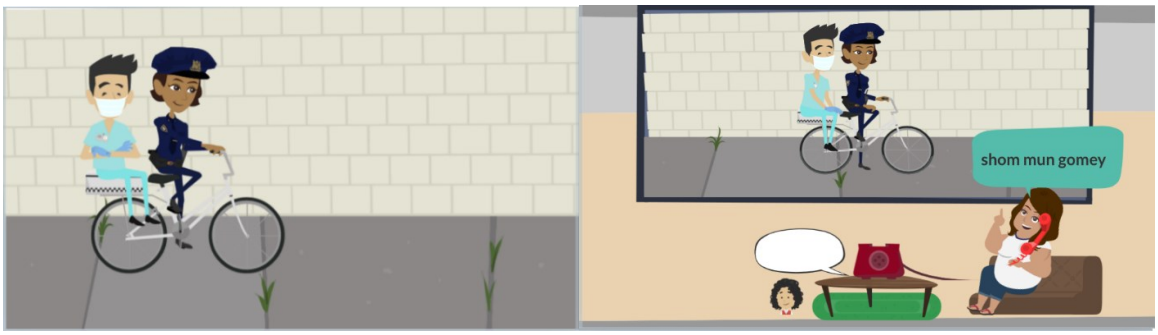

Figure 7a. An example of a declarative sentence in a non-prominent CG context. "The police officer is driving the nurse."

Figure 7b. An example of a declarative sentence in a prominent CG context. "*The farmer is kicking the nurse.*"

Similarly, for interrogatives, the non-prominent CG context (Figure 8a) involves a character (the addressee) watching the event alone while describing it to the other character (the addresser) via the phone. To make the context pragmatically plausible for a question,

the addressee would leave out the details about either the subject or object of the event at first. The addresser then asks a question about such detail, and the addressee responds. To simplify the matter and let the participants focus on the interrogative sentence, the initial description of the addressee is depicted using an illustration of the event instead of words.

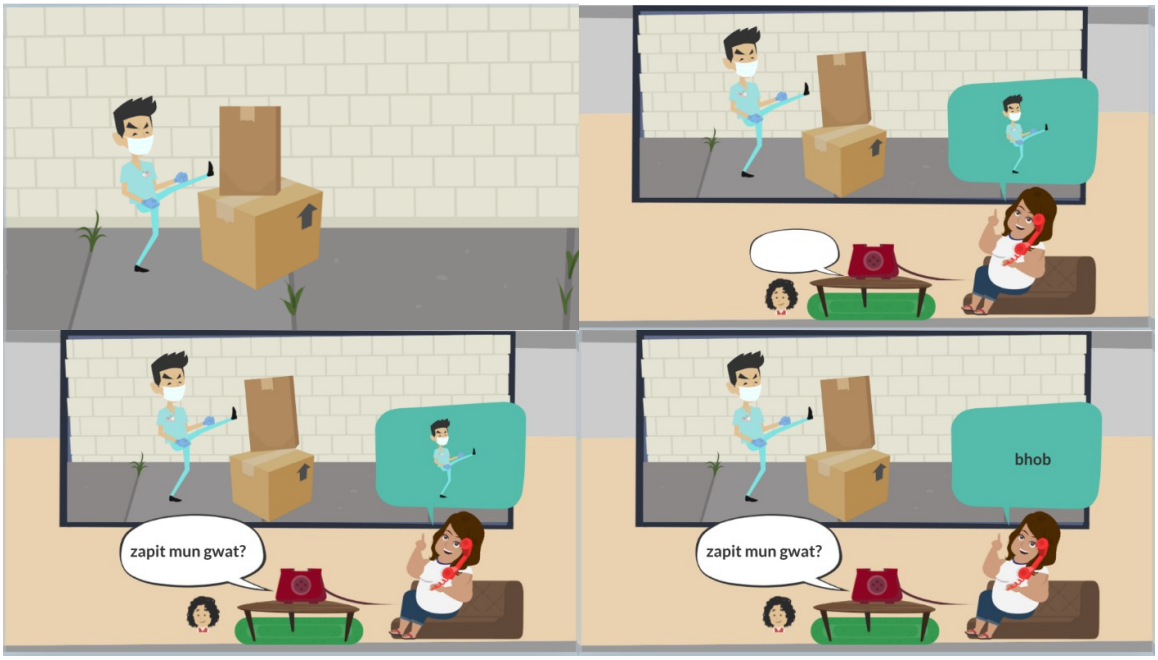

Figure 8a. An example of an interrogative in the non-prominent CG context. “What is the nurse kicking? –Boxes.”

For interrogatives in the prominent CG context (Figure 8b), the addresser and the addressee are watching the event on the big screen together when the addresser asks about a detail of the event. Since both parties have equal access to the information, it would seem infelicitous for either of them to ask about such information. Thus, the question here is portrayed as a probe question, which is used to assess the addressee’s knowledge about the information rather than to elicit the information.

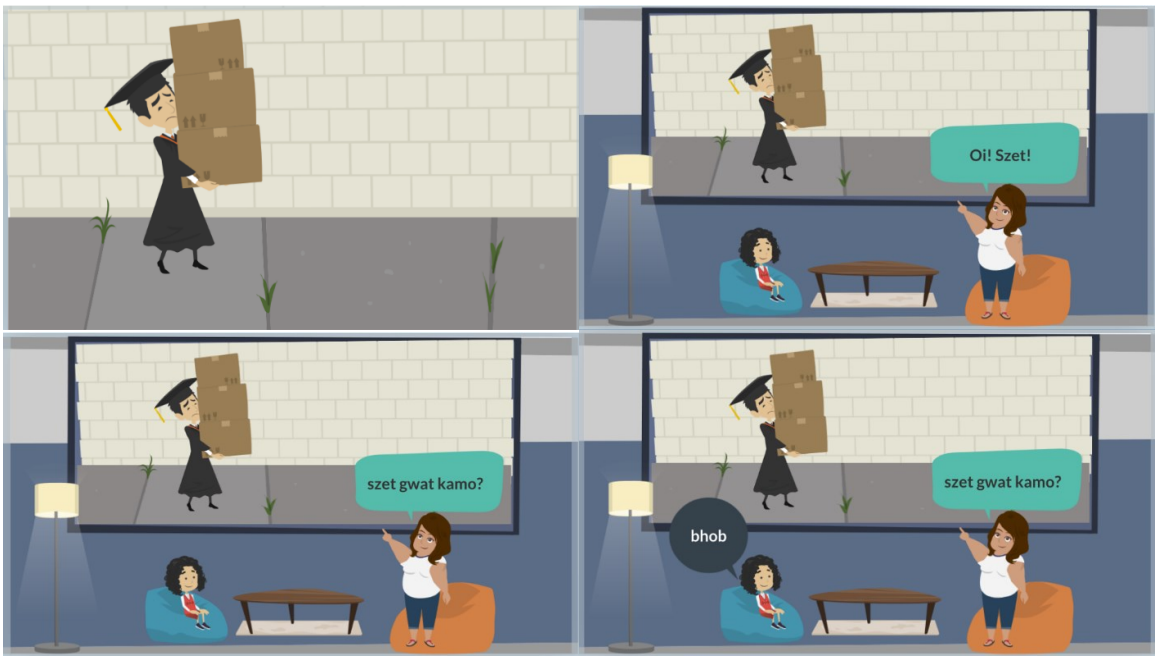

Figure 8b. An example of an interrogative in the prominent CG context. "*Look! A student! What is the student carrying? – Boxes.*"

Word order**.** Declarative sentences have canonical SOV word order, while wh-interrogatives have flexible word order: SOV and VSO. SOV interrogatives have an identical structure to declaratives, with the wh-word replacing either the subject in a subject question (hence WhOV) or the object in an object question (hence SWhV). However, SOV interrogatives are more restricted than VSO interrogatives and are used in the prominent CG contexts only. On the other hand, VSO interrogatives can be used in both the non-prominent CG and prominent CG contexts.

The SOV interrogatives in this artificial language share some similarities with the wh-in-situ variant in natural languages, specifically in English: there is no movement involved and the question is more discourse-restricted. Likewise, the VSO interrogatives share some similarities with the fronted wh-question variant: they involve transformational V to T movement and the question is discourse-unmarked, i.e., unrestricted. However, there is a difference between the artificial constructions and English questions: the moved

element in the artificial VSO interrogative is the verb while it is the wh-word in a fronted English wh-question. In a sense, one might argue that the VSO interrogatives are also "wh-in-situ" questions, as the moved element is not the wh-word. However, VSO interrogatives were chosen to represent moved-wh questions for a number of reasons. First, given that our artificial language contains only simple 3-word structures, using actual fronted wh-questions would result in ambiguous questions. For example, "who Mary hit?" in the language can mean either an object question or a subject question. This would introduce a confounding factor (ambiguity) into the experiment, as participants may try to avoid ambiguous structures. Second, using actual fronted wh-questions would also mean that subject questions cannot be used in the test phase, as it is impossible to tell whether a subject question is used as the in-situ or fronted variant. Finally, and most importantly, given the claim from section 2.2.3. that the position of the wh-word is not the main factor driving its discourse marked/unmarked status, I believe that the difference between the VSO variant in the artificial language and the fronted wh-question variant in English would not negatively affect the main goal of the experiment, which is to test whether there is a trade-off between structural economy and discourse restriction when learning multiple variants. Ultimately, the variants in this artificial language share the same characteristics in terms of structural economy and discourse restriction with the wh-variants in English.

Procedure

The experiment consisted of an exposure phase and three tasks. The first task was always the Production task to avoid additional exposure to the language from the Comprehension and Forced-choice task. The two other tasks, Comprehension and Forced-choice, were randomized in order. The whole study took approximately 45 minutes.

Exposure Phase (25 minutes)

Participants were exposed to a total of 160 trials. The training trials were balanced between declaratives and interrogatives, SOV interrogatives and VSO interrogatives, as well as prominent CG and non-prominent CG (Table 8). There were more non-prominent CG declaratives than declaratives, suggesting that the non-prominent context is more common in the language. Each declarative trial lasted 13s and consisted of 2 images: a close-up view of the transitive action that lasted 3s, and a full view of the speaker describing the action that lasted 10s. Each interrogative trial lasted 20s and consisted of 4 images (Figure 8): a close-up view of the transitive action (3s), a speaker calling attention to the action (5s), a speaker asking the question (8s) and a listener responding (4s). A 2-second blank screen was shown between each trial.

In this phase, each of the verbs appeared 40 times, each noun appeared between 24-25 times, the "what" wh-word appeared between 20-22 times, and the "who" wh-word appeared between 58-60 times.

Table 8. Distribution of trials by sentence type and condition.

	Declarative	***SOV interrogative***	***VSO interrogative***	***Total***
Prominent CG	10	20	10	***40***
Non-prominent CG	30	0	10	***40***
Total	***40***	***20***	***20***	***80***

Comprehension Task (~6 minutes)

The task consisted of 12 trials, 6 of which included interrogatives and 6 included declaratives. On each trial, participants saw a sentence along with 2 images and were asked to choose the image that matched the sentence the best. Four trials were designed to test participants' knowledge of the words in the language. Each trial involved 2 images

depicting different actions (verb comprehension) between two characters (noun comprehension) or a character and an object (wh-word for subject/object comprehension). The remaining trials were designed to test participants' knowledge about word order, with the 2 images depicting the same action and same characters but in reversed roles. Overall, the comprehension task served the purpose of establishing whether the participants had learned the target language.

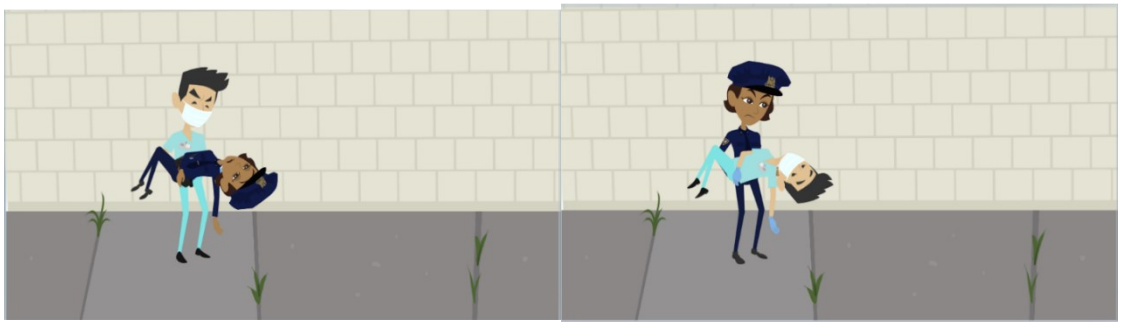

Figure 9. An example trial in the Comprehension task testing word order knowledge.

Production Task (~10 minutes)

This is the main task pertaining to regularization in the production of wh-questions in the artificial language. Participants were shown pictures similar to an interrogative trial in the exposure phase, with the actual question retracted. They were instructed to guess what the question might have been given the event and the answer. Since the focus of this task was to assess participants' knowledge about and preference in choosing the interrogative structure and not on the ability to memorize vocabulary, all the possible words in the language were provided. Participants had to choose the appropriate words and put them in the right order.

There were 10 trials in total. Five of the trials involved a prominent CG context and 5 involved a non-prominent CG context similar to what they had seen in the exposure phase. For each trial, accuracy, type of error, and the question variant used were recorded.

A response was counted as correct if it replaced the element to be asked with a wh-word, had the right subject-object word order, and had at most one vocabulary error. For incorrect responses, the type of error could be incorrect word order (e.g., SVO) or incorrect words. The type of variant used was recorded regardless of accuracy.

Figure 10. An example trial in the Production task. Given that the answer was "chef", the question the participant needed to come up with would be "who was hitting the student?'.

Forced-choice Task (~10 minutes)

This task is similar to the Production task, albeit at a less cognitively-demanding level. Instead of producing the questions themselves, participants were asked to choose between the SOV and VSO variants. Participants were shown pictures similar to an interrogative trial in the exposure phase, but with certain details obscured. They were instructed to choose an appropriate question to ask about the obscured detail. The goal of the task was to see if, without the cognitive burden of a typical production task, participants would still demonstrate a regularizing pattern or if they would try to match the distribution of the variants.

Similar to the Production task, there were 10 trials in total. Five of the trials involved a prominent CG context, and 5 trials involved a non-prominent CG context.

New context

In addition to the prominent and non-prominent CG contexts, there were 2 production trials and 2 2AFC trials involving a new context that participants had not seen, for example, having the addresser and addressee watch the screen separately before meeting each other to talk about it. These new-context trials were analyzed separately to assess participants' ability to generalize beyond what they have learned in the input.

Results

a. General learning of the language

Comprehension task accuracy

In general, participants demonstrated good accuracy in the comprehension task (Figure 11). The overall accuracy across all participants was 93%, significantly above chance level ($p < .0001$). The range of accuracy was between 58.3% and 100%, with only 1 participant scoring below 60%, 6 participants scoring between 60%-90%, and 27 participants scoring above 90%.

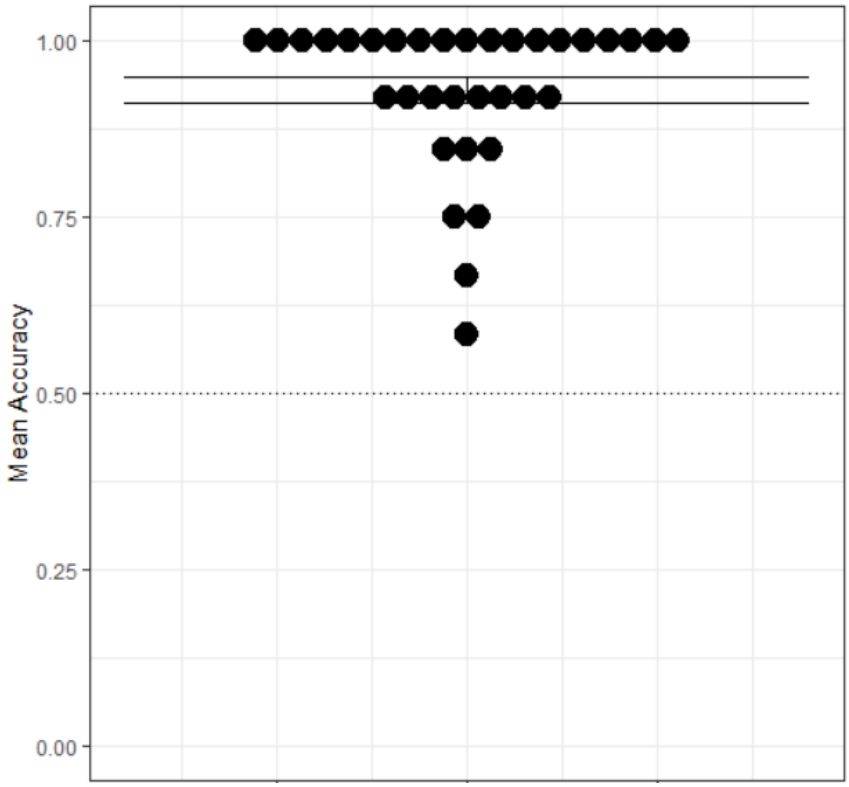

Figure 11. Individual participant means of comprehension accuracy, error bars show 95% confidence intervals on by-participant means.

A mixed-effect logistic regression model using the lme4 package in R (version 4.0.3) was used to examine the comprehension accuracy. The model included the binary response for each test trial (1 for correct response and 0 for incorrect response, dummy coded) as the dependent variable, with SentenceType (e.g., declarative versus interrogative, dummy coded with interrogative as baseline) and ComprehensionTarget (e.g., the question targeting word order or vocabulary knowledge, dummy coded) as the main effects, and Participant and Question Item as random effects. Table 9 reports the model. The results show that there was no effect of Sentence Type or Comprehension Target. Participants performed equally well with declarative comprehension (M = 92.1%) and interrogative comprehension (M = 94.3%), as well as with trials targeting vocabulary comprehension (M = 91.5%) and trials targeting word order comprehension (M = 93.4%)

Table 9. Summary of the regression model of participants' performance in the comprehension task.

	Estimate	*Std. Error*	*z-value*	*p-value*
Intercept	2.732	0.581	4.707	< .0001 ***
Sentence Type = Question	0.756	0.510	1.487	.138
Comp. Target = Word Order	0.792	0.458	1.731	.083

*Note: * p < .05; ** p < .01; *** p < .001*

Production task accuracy

A production trial was counted as correct if it replaced the element to be asked with a wh-word, had the right subject-object word order, and had at most one vocabulary error. Overall, participants reached an accuracy of 63% in their production.

There were four types of errors in the production task. The most frequent type was confusion about the position of the subject-object. This is surprising, given that most participants did rather well in the comprehension trials targeting word order. In line with Hendriks and Koster (2010) and Humphreys (2012), I hypothesize that the lower accuracy in production tasks was due to production requiring more intake (to learn from) and cognitive resources such as planning and memory. Another type of error was SVO interrogatives. This is most likely due to a grammatical transfer effect from English to the artificial language (e.g., a failure to inhibit). While there wasn't a high percentage of this error type, it raises a concern about the difference between ALL and first language acquisition, as the former, like second language acquisition, involves an additional task of suppressing previous linguistic knowledge (Ettlinger, Morgan-Short, Faretta-Stutenberg, & Wong, 2016). Finally, there was a small number of vocabulary errors in which participants used more than one word incorrectly, and a non-significant number of

incomplete errors in which participants did not provide a completed sentence. A full breakdown of these error types is reported in Table 10.

Table 10. Types of errors in the production task.

Type of error	Number of trials (Percentage)
Word order	112 (29.9%)
Vocabulary	10 (2.5%)
Incomplete	2 (0.1%)

b. Overall regularization

Since understanding participants' production choices when learning the two variants was the main goal of the study, incomplete productions and SVO productions were excluded from this analysis. SVO productions are likely the result of the transfer effect from English and do not reflect a variant preference, as they are as similar to SOV (only differ by the V-O placement) as to VSO (only differ by the V-S placement).

The results are illustrated in Figure 12a and Figure 12b. Figure 12a shows a categorization of participants into those who demonstrated a strong preference for one variant over the other by boosting the frequency of that variant to over 80% as "regularizers" (following Hudson Kam & Newport, 2009). In the Production task, there were 16 regularizers (47% of the participants): 14 of them regularized to VSO and 2 regularized to SOV. In the 2AFC task, there were 5 regularizers (out of 34 participants): 3 VSO regularizers and 2 SOV regularizers. A closer look at individual performance shows that 7 participants produced 100% of the VSO variant in the Production task, while no participant produced exclusively SOV. Interestingly, one participant (out of 2) who produced SOV dominantly apparently imposed their own rule on the language by

consistently fronting one particular verb; in other words, this participant always produced the VSO variant when the verb was 'kamo' and produced the SOV variant with all other verbs. It is unclear how the participant had arrived at this rule, as the distribution of the four verbs was balanced between the two variants. However, a number of previous studies in ALL have also reported cases in which participants come up with their own rules to systematize the language (e.g., Hudson Kam & Newport, 2009; Wonnacott, 2011). In the 2AFC task, there was only one participant who consistently chose the VSO variant in 100% of the trials, and no participant consistently chose the SOV variant. Figure 12b shows the mean proportion of VSO across the two tasks and two discourse contexts. The Non-prominent CG context induced more VSO utterances than the Prominent CG context, and the gap in VSO proportion between the two contexts was more pronounced in the Production task than in the 2AFC task.

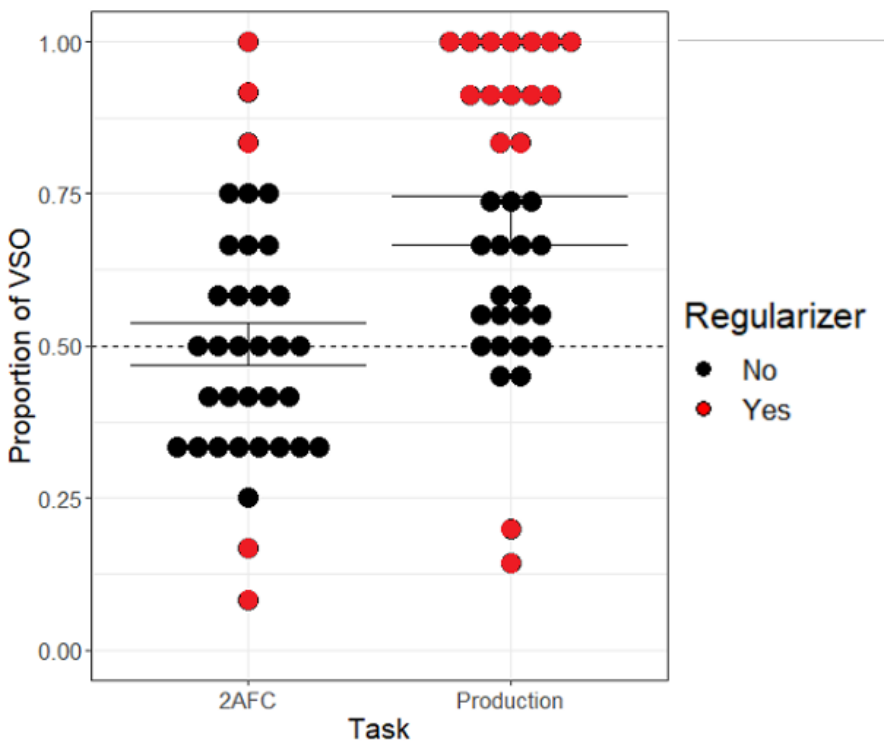

Figure 12a. Individual participant means of VSO proportion across two tasks, error bars show 95% confidence intervals on by-participant means. Dots are colored according to whether the participant is classified as a regularizer (red) or not (black) based on the proportion of VSO produced.

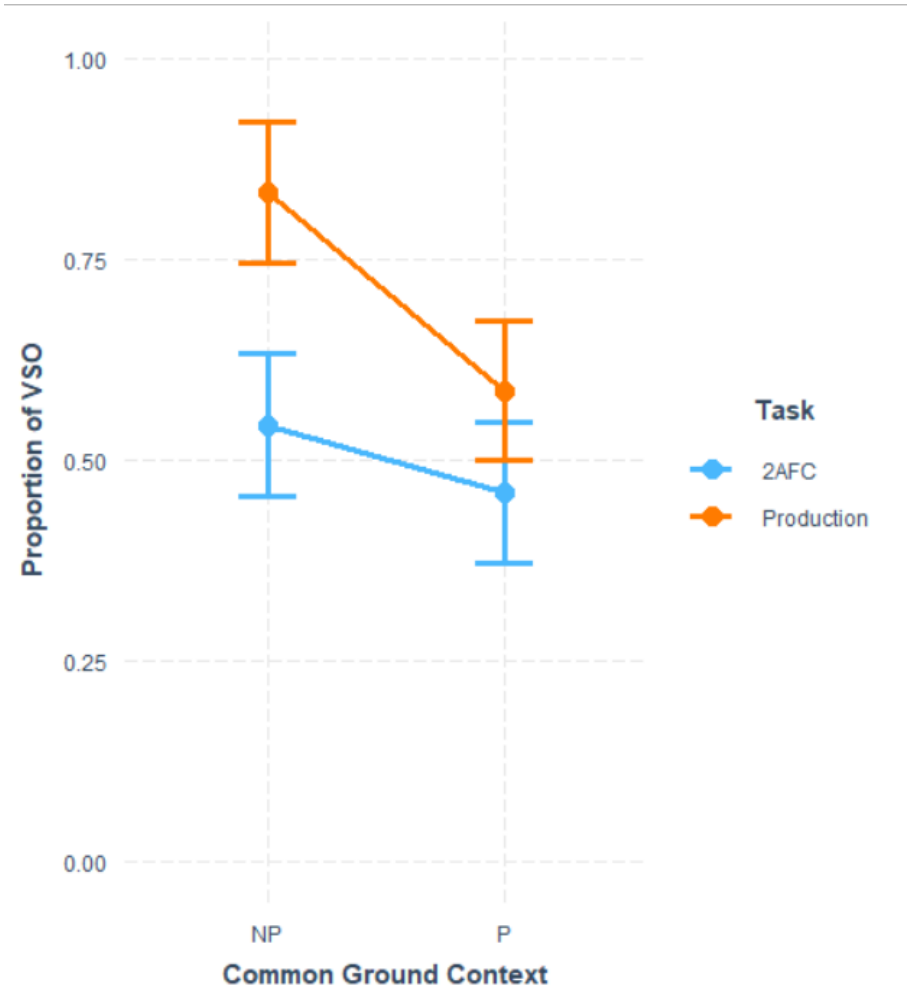

Figure 12b. Mean proportion of VSO utterances produced by participants across the two critical tasks and two discourse contexts.

The data was then submitted to a mixed-effect logistic model using the lme4 package in R. The model included the binary response for each test trial (either VSO (=1) or SOV (=-1)) as the dependent variable, with Task (Production versus 2AFC), and Discourse (prominent versus non-prominent CG context) as the main effects. The model also included random intercepts for Participant and Question Item and a by-participant

random slope for Discourse. P-values were obtained using the Satterthwaite approximation. The full model is reported in Table 11. The intercept shows that participants demonstrated a significant preference for the VSO variant. Participants produced a VSO-SOV distribution of 70.7%-29.3%, significantly different from the 50%-50% distribution in the input ($p < .001$). There is a significant effect of Task. As seen in Figure 12, participants produced more VSO in the production task than in the 2AFC task. When presented with a 2AFC preference task (which is associated with less cognitive demands than a production task), participants' preference for the unrestricted VSO variant lessened. Instead, participants matched the input distribution by overall choosing VSO 50.3% and SOV 49.7% of the time, which was not significantly different from the input distribution (t-test, $t = 0.068$, $df = 34$, $p = 0.945$). Furthermore, there was a significant effect for Task and Discourse. This suggests that participants regularized to VSO differently between the production and 2AFC task and their regularization was affected by the discourse context of the question.

Table 11. Summary of the regression model in Experiment 3.

	Estimate	*Std Error*	*z-value*	*p-value*
Intercept	0.546	0.042	13.047	<.001 ***
Task = Production	0.288	0.056	5.170	< .0001 ***
Discourse = P	-0.082	0.057	-1.431	0.155
Task*Discourse	-0.165	0.078	-2.097	0.03 *

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

Following Ferdinand et al. (2019), I also calculated the Shannon entropy as a measure of regularization (Shannon, 1948). The entropy of the artificial language with p(V)

= {0.5, 0.5} is 1 bit, in which V is the set of variants and p(V) is the probability distribution of the variants. Entropy is calculated by:

$$H(V) = - \sum p(v_i) log_2 p(v_i)$$

The entropy of the produced language is 0.88 bits, a reduction of 0.12 bits. More intuitively, this indicates that 12% of the variation in the language was due to regularization by the learners.

c. Regularization in the Prominent CG context

The preference for the VSO variant emerges more clearly in the Prominent CG context. Although the overall distribution of the two variants in the whole language was balanced at 50%-50%, the distribution of SOV – VSO in the Prominent CG context was 67%-33%, since this condition consisted of *all* SOV variants but only *half* of VSO variants. Thus, if participants were to match the frequency distribution, we should expect more SOV than VSO, at least in this condition. However, participants shifted the distribution to show a clear preference for VSO at 70%, as illustrated in Figure 13.

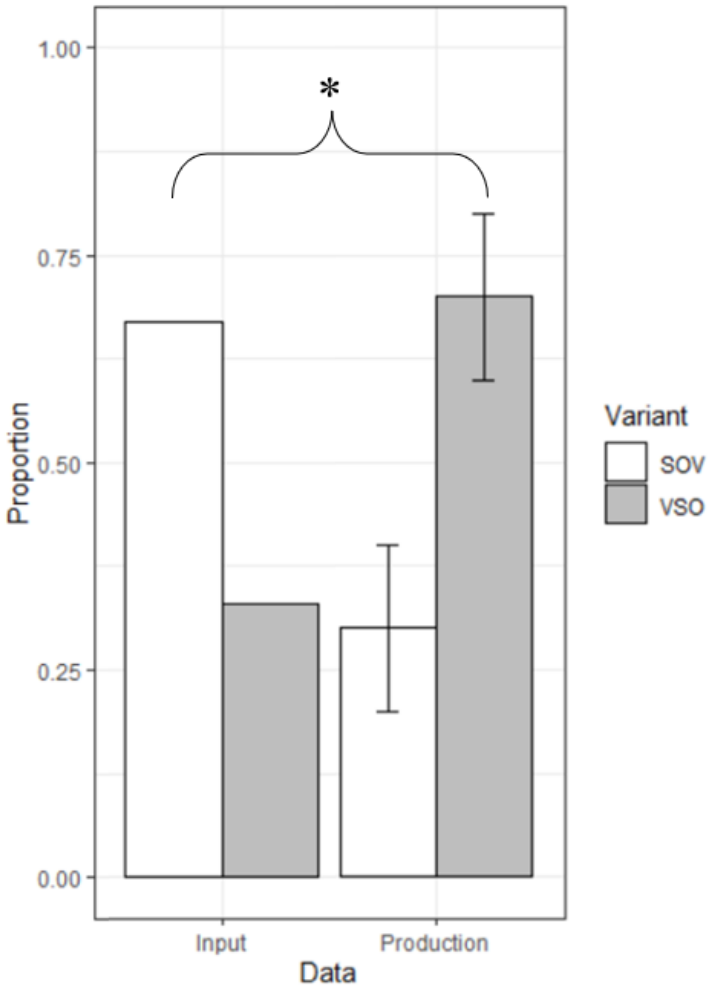

Figure 13. Distribution of the two variants in the Prominent CG context in the input versus tested production.

4.1.4. Discussion

Discourse-unmarkedness as a facilitating factor in learning

Overall, participants demonstrated high accuracy in the comprehension task, suggesting that they were able to learn the artificial language. Although not every participant displayed a preference for the VSO variant in the production task, the majority (~80%) of them did. The VSO variant involves an additional syntactic transformation (fronting the verb) compared to the canonical structure of the language (SOV) that participants were first exposed to with declaratives. The SOV variant is syntactically

simpler in the sense that it shares the same word order with the canonical (declarative) structure. However, the SOV variant is discourse-marked as it is typically used when there is prominent CG, while the VSO variant can be used in both prominent and non-prominent CG contexts. The overall preference for VSO in the production task suggests that when frequency is kept equal between the two variants, learners trade-off structural simplicity for discourse unmarkedness. This is not unreasonable: producing the discourse-unmarked variant minimizes the chance of violating discourse constraints (as VSO can be used in any context) and maximizes the chance of efficient communication.

It is possible, however, that the preference for VSO happens due to this variant being structurally different from declaratives. Learners may simply prefer that different sentence types (e.g., declarative, interrogative, imperative etc.) have different syntactic structures. Since SOV interrogatives and SOV declaratives share the same syntactic structure, learners may want to avoid the confusion by going for the VSO structure. By doing this, learners can also maximize efficient communication by presenting two cues to tell the interrogative goal of the sentence (i.e., structure and wh-word) instead of just the wh-word. Future studies can control and eliminate this possibility by having the in-situ variant being one-transformation more syntactically complex (e.g., wh-word movement) and the fronted variant being two-transformation more syntactically complex (e.g., wh-word movement and verb movement) than the canonical declarative. That way, both variants are different from the declaratives yet maintain their structural mismatch.

Note that while being a possibility in this ALL experiment, the possibility above does not extend well to natural languages. Recall that children rarely produce wh-in-situ while adults produce them frequently in child-directed speech. If producing a wh-in-situ

impairs processing (because listeners have to wait until the wh-word to interpret it as a question) and producing fronted wh-question helps with communicative goals (by signaling the illocutionary force early), then essentially we are saying that children are better communicator than adults, and adults tend to communicate less efficiently when talking to children than with other adults, which is counterintuitive.

Discourse unmarkedness as a factor in wh-question acquisition

In section 3.4., we saw that English-speaking children understand but do not produce in-situ probe questions in a behavioral experiment setting. Children's elicited production of wh-questions in languages that share some similarities with English, like French and Brazilian Portuguese, also shows the same general preference for fronted questions, though their distribution between the fronted and in-situ questions is not as extreme as English-speaking children's. Such preference is intriguing as it is not predicted by the influential structural economy-based accounts (e.g., in English: Brown (1968) on the number of transformations; in French: Jakubowicz's (2011) Derivational Complexity Hypothesis). Frequency-based accounts fare better, but frequency alone is not sufficient to explain the preference for fronted questions, as discussed in Chapter 4. Another factor is proposed here: discourse markedness. The adult ALL experiment described in this section is an attempt to study the role of discourse unmarkedness more carefully by controlling for frequency. The results show a similar trend to children's elicited production of wh-questions in English, French, and Brazilian Portuguese – learners prefer the discourse-unmarked variant, even when that variant requires additional syntactic operations. This is likely because the marked variant tends to be formally more complex (to process) and more prone to misinterpretation (Farkas & Roelofsen, 2017). Overall, our results highlight the

role of discourse unmarkedness in syntactic regularization and while they do not concern child regularization per se, they support the general hypothesis that children's production of fronted wh-questions may be motivated by a discourse factor. Future ALL adaptations to children of the experiments presented in the section will be needed to (dis)confirm the hypothesis.

Note that while the current ALL design is modeled after English wh-questions, it does not capture every aspect of the variation in English, in particular, the frequency distribution of the variants. To explore the role of discourse independently from frequency, the distribution of the two variants in the artificial language was balanced. In reality, it is very likely that the higher frequency of fronted questions also contributes to the regularization picture. How much of regularization is motivated by frequency and how much is motivated by discourse markedness is another question that I do not explore in the scope of this work. Since the two factors tend to go together in natural language, it does not make sense to try to claim that regularization happens solely due to one factor and not the other. The ALL experiment here only shows that discourse *can be* a factor conditioning regularization, and I propose that an account taking discourse into consideration can resolve some of the issues that a frequency-based account alone faces, such as the differences in the production rate of wh-variation across English and French or the asymmetry between wh-question comprehension and production in French studies.

Regularization in ALL

About half of the participants (16 out of 34) regularized their production. It was not surprising that not all adult learners regularized. For adults to show regularization, the artificial language has to be simple enough so that some learning can happen yet complex

enough to motivate them to deviate from reproducing the exact input (Hudson Kam & Newport, 2009; Fedzechkina et al., 2017). Adults differ drastically in their ability to learn (a second) language (Ehrman, Leaver, & Oxford, 2003; Dörnyei, 2014). Given the high variability among adult learners, it is not surprising that the experiment only induced regularizing behavior in some but not others.

In the 2-alternative forced-choice task, the number of regularizers dropped to only 5 out of 34. The difference in the number of regularizers between the two tasks is in line with previous studies (e.g., Schwab et al., 2018) and supports the idea that learners regularize in tasks that require more cognitive resources, as the production task arguably involves more retrieval effort (plus planning) than the 2AFC task. In general, the ALL experiment was able to show that participants regularize when there is more than one variant in the input, but such behavior is more likely to be found in more cognitively-taxing tasks, a point I will explore further in Chapter 5.

Finally, while 2 participants regularized to the structurally simpler variant (SOV), the majority of regularizers regularized to the discourse-unmarked (but more syntactically complex) variant (VSO). Regularization can be messy with learners being sensitive to different factors or even imposing their own rules (e.g., Hudson Kam & Newport, 2009; Wonnacott, 2011). Still, the overall pattern suggests that discourse is a factor in the regularization of this particular artificial language. This means that while regularization is not limited only to language learning, learners can be sensitive to certain linguistic factors when they regularize. If so, regularization cannot be construed as a strictly domain-general cognitive process. This finding is in line with Ferdinand et al. (2018), which claims that

there are at least two sources of regularization, a domain-general source based on cognitive load and a domain-specific source triggered by linguistic stimuli.

4.2. Experiment 4a[15]: The degree of discourse markedness in learning variation in artificial grammar

Experiment 3 has shown that when frequency is held constant, adult participants prefer producing the unmarked variant over the marked variant. Do participants only make a general distinction between marked and unmarked, or do they pay attention to the degree of markedness as well? Recall that in Section 3.4., I hypothesized that the difference in production rates of wh-in-situ between French-speaking and English-speaking children may be due to the difference in the strictness of discourse requirements of the two languages. In particular, English wh-in-situ is more strictly constrained than French wh-in-situ. While this hypothesis cannot be tested directly, results from an additional ALL experiment with a similar setup can lend support for, or show evidence against, the hypothesis.

In this section, I present a follow-up study to Experiment 3 in order to look into a finer distinction between "strictly constrained" and "loosely constrained" variants. Results from this experiment would provide a deeper understanding about how sensitive learners are to the degree of discourse markedness in language learning, as well as test if this explanation might shed light on the difference in wh-in-situ production between French-speaking (higher production rate of wh-in-situ) and English-speaking children (lower

[15] Experiment 4a and Experiment 4b (discussed in Chapter 5) are the same experiment but target different research questions. After completing the language learning task (the main target in Experiment 4a), participants moved on to complete a series of working memory tasks (the main target in Experiment 4b).

production rate) discussed in Chapter 4. To the best of my knowledge, no previous ALL study has explored this particular distinction.

I hypothesize that a variant that is less marked is more likely to be learned and used than a marked variant. Thus, learners exposed to an unmarked variant and a less-marked variant are expected to be less likely to regularize, compared to those exposed to an unmarked and a more-marked variant.

4.2.1. Methods

Participants

Ninety-five adult participants (age range from 20 to 40, mean age = 33.5, female = 49) were recruited on the Prolific platform. All participants reported being raised as an English monolingual only and knowing no second language.

The inclusion and exclusion criteria were identical to Experiment 3. Participants were recruited in several small "batches". The quality of data was checked after each batch, and the number of participants recruited for each condition was adjusted accordingly so that after exclusion, the total number of participants in each condition would be equal. Eleven participants with more than 5 minutes of total inactive time throughout the experiment were excluded, leaving us with 42 data points in each condition (84 in total).

Materials

The artificial language stimuli and procedure described in Experiment 3 were used in this follow-up study. The major difference between Experiment 4a and Experiment 3 lies in the frequency distribution of the variants. Experiment 4a consists of two conditions, the distribution of the stimuli in each condition is reported below in Table 12. The overall frequency of SOV-VSO was kept balanced, but the distribution of the marked variant SOV

was manipulated. In Condition 1 – the *Strict* discourse requirement ("English") condition, SOVs appeared mostly in a prominent CG context (90%) and rarely in a non-prominent CG context (10%). In Condition 2 – the *Loose* discourse requirement ("French") condition, SOVs appeared more frequently in a prominent CG context (70%) but could still appear in a non-prominent CG context (30%). The assumption is that the stricter the requirement is for a variant, the less likely it will appear in a non-conforming context. For example, both English and French wh-in-situ have to satisfy discourse givenness, however, French wh-in-situ may appear in an out-of-the-blue context (see Chapter 2, example (30)) and most likely does so more frequently than English wh-in-situ.

Based on the feedback from participants in Experiment 3 as well as the overall result that all participants successfully learned the artificial language, the number of Declarative trials was reduced by 20% to increase the difficulty of the task and shorten the length of the experiment.

Table 12a. Distribution of trials by sentence type and CG in Condition 1.

	Declarative	***SOV interrogative***	***VSO interrogative***	***Total***
Prominent CG	8	18	10	**36**
Non-prominent CG	24	2	10	**36**
Total	***32***	***20***	***20***	**72**

Table 12*2*b. Distribution of trials by sentence type and CG in Condition 2.

	Declarative	***SOV interrogative***	***VSO interrogative***	***Total***
Prominent CG	12	14	10	**36**
Non-prominent CG	20	6	10	**36**
Total	***32***	***20***	***20***	**72**

4.2.2. Results

Accuracy performance

The overall accuracy across all participants in Condition 1 was 82.36% and in Condition 2 was 82.94%, significantly above chance level (ps < .0001). The performances between the two groups were not significantly different. Figure 14 illustrates the results.

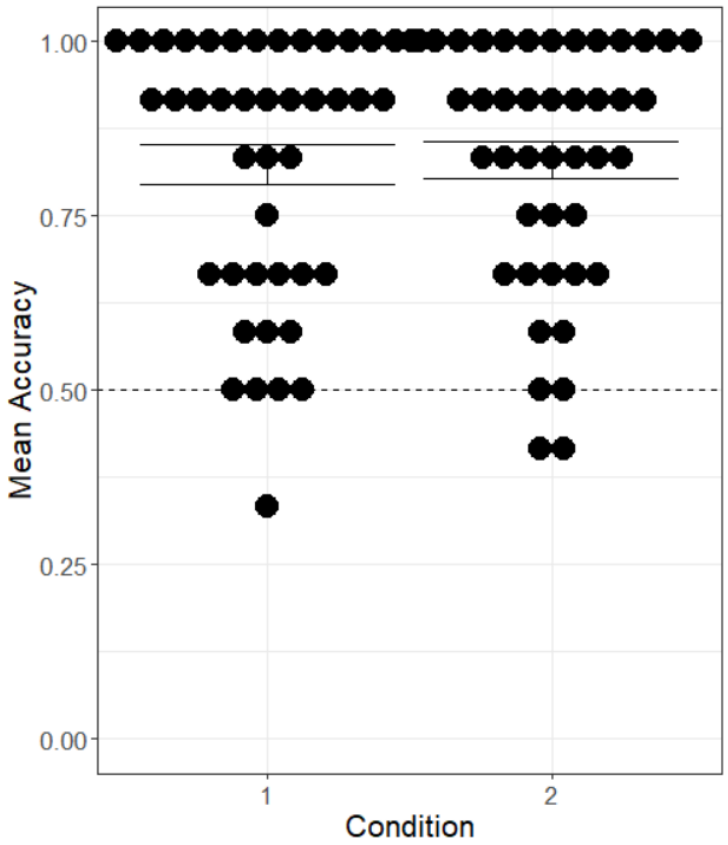

Figure 14. Individual participant means of comprehension accuracy in Condition 1 and Condition 2. Error bars show 95% confidence intervals on by-participant means.

A mixed-effect logistic regression model was used to examine the comprehension accuracy. The model included the binary response for each test trial (1 for correct response and 0 for incorrect response) as the dependent variable, Condition (1 versus 2) as the main effects, and Participant and Question Item as random factors (Table 13).

Table 13. Summary of the regression model of participants' performance in Experiment 4a Comprehension task.

	Estimate	*Std. Error*	*z-value*	*p-value*
Intercept	1.557	0.473	3.332	.0008***
Condition	0.280	0.300	0.933	.35

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

The overall comprehension accuracy is lower than that in Experiment 3, with a few participants falling below 50%. This is most likely due to the reduction in the total number of trials participants are exposed to, which makes the learning task here more difficult. Still, participants in general demonstrated good comprehension, suggesting that the majority of them had successfully learned the language.

Overall regularization in production and 2AFC

Production data that was either incomplete or contained serious structural errors (e.g., using an SVO structure) was excluded. Out of 1008 total utterances, approximately 74 utterances (7.4%) were excluded. Of these, 44 utterances were SVO and 30 utterances were incomplete (lacking either the main verb or wh-word). As in Experiment 3, SVO utterances were likely a result of language transfer from English, with such transfer tending to happen when speakers are not proficient in the target language (Amin, 2017).

Individual performance is displayed in Figure 15a. In Condition 1, 7 participants (~16%) produced more of the structurally simpler but discourse-marked SOV variant, and in Condition 2, 10 participants (23%) did so. While not every participant regularized to VSO, overall VSO was still the preferred variant in both conditions for the majority of the participants. Similarly to Experiment 3, participants who boosted the frequency of a variant to over 80% were categorized as "regularizers". In both conditions, there was no SOV regularizer, while Condition 1 had 10 (23.8%) and Condition 2 had 4 (9.5%) VSO regularizers. Figure 17b shows that participants in Condition 1 demonstrated a stronger preference for the unmarked variant VSO (66.8%) than those in Condition 2 (58.9%) in the Production task, but not in the 2AFC/forced choice task. Furthermore, the proportion of VSO in the 2AFC was lower than in the Production task for both conditions. These results

overall replicate the findings from Experiment 3: when being exposed to both marked and unmarked variants, participants tended to boost the proportion of unmarked variants to a higher number than was in the input, but such tendency is more pronounced in a (cognitively more demanding) production task than in a 2AFC task.

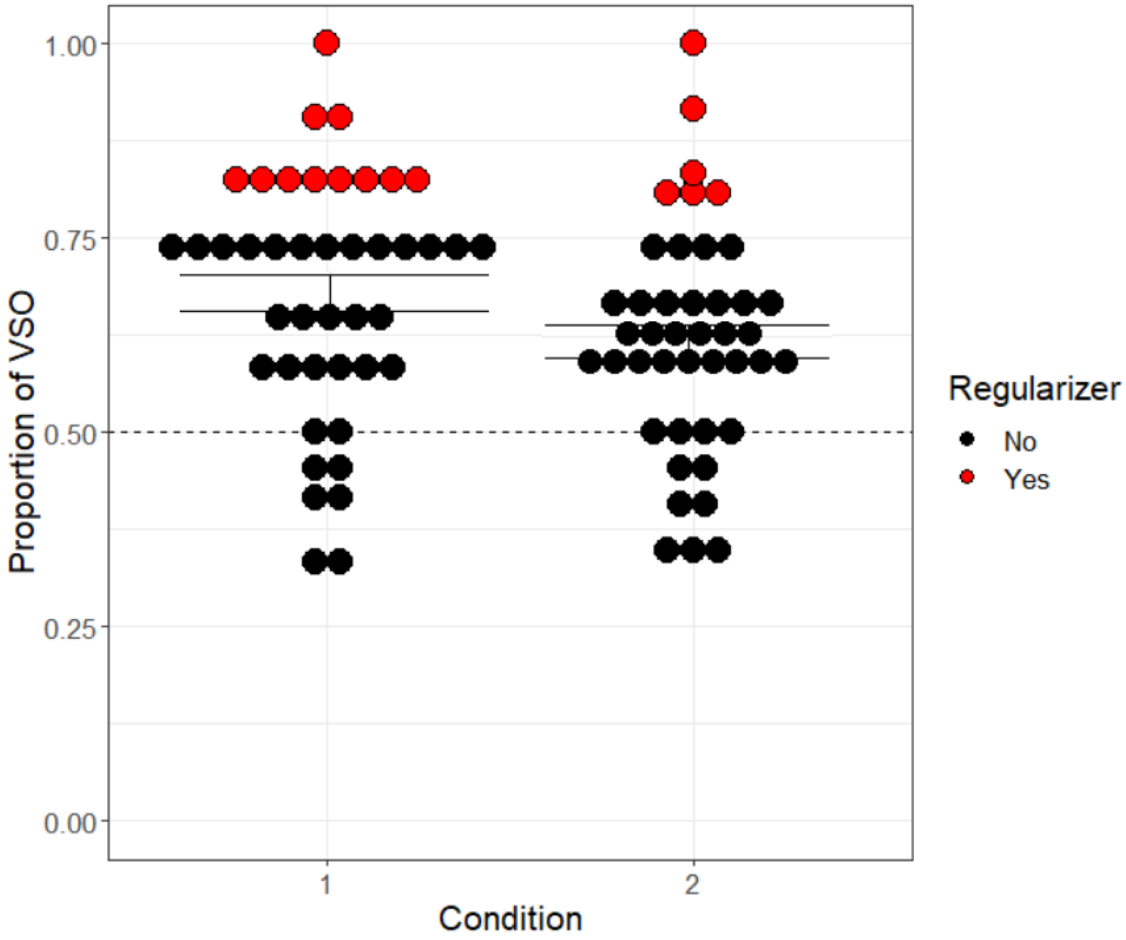

Figure 15a. Individual means of the proportion of VSO utterances produced by participants in Conditions 1 and 2, error bars show 95% confidence intervals on by-participant means. Dots are colored according to whether the participant is classified as a regularizer (red) or not (black), based on the proportion of VSO produced.

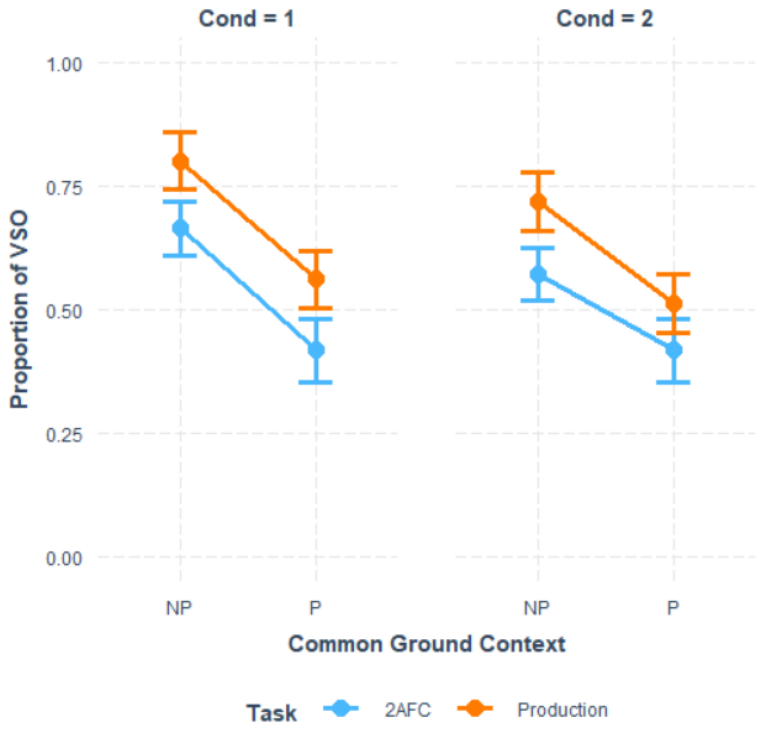

Figure 15b. Mean proportion of VSO utterances produced by participants in Condition 1 (left) and Condition 2 (right) across the two critical tasks and two discourse contexts.

The data was submitted to a mixed-effect logistic model using the lme4 package in R. The model included the binary response for each test trial (either VSO (=1) or SOV (=-1)) as the dependent variable, with Condition (1 versus 2), Task (Production versus 2AFC), and Discourse (Prominent versus Non-prominent CG context) as the main effects. The model also included random intercepts for Participant and Question Item and a by-participant random slope for Discourse. Table 14 reports the full model.

Table 14. Summary of the regression model in Experiment 4a.

	Estimate	*Std Error*	*z-value*	*p-value*
Intercept	0.295	0.078	3.788	.0004 ***
Condition = 2	-0.185	0.080	-2.297	.02 *
Task = Production	0.230	0.081	2.821	.004 **
Discourse = P	-0.413	0.099	-4.144	< .0001 ***
Condition*Task	0.019	0.112	0.175	.861
Condition*Discourse	0.190	0.118	1.609	.107
Task*Discourse	0.099	0.126	0.787	.431
Task*Discourse*Condition	-0.136	0.165	-0.819	.412

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

The significant intercept suggests that participants produced the VSO variant above chance level. Furthermore, there was a significant effect for Condition, Task, and Discourse. This suggests that participants regularized to VSO differently between the two conditions and the two tasks. Moreover, they did not regularize to VSO across the board but their regularization was affected by the discourse context of the question. This suggests that even though both groups showed sensitivity to the constraining rules, there was a difference between the level of sensitivity. The learning of the discourse constraints is discussed in detail in the Discussion section (4.2.3.). No interaction effect was found.

Regularization in the Prominent CG context

As shown in Table 14, participants regularized to VSO differently depending on discourse contexts. This section provides a deeper look at the preference for VSO in the Prominent CG context.

In the Prominent CG context, either variant can be used, making it the ideal context to look for a preference. The distribution of VSO-SOV in the Prominent CG context in Condition 1 was 35%-65%, since this context consisted of half of VSO variants and 90% of SOV variants. Thus, if participants were to match the frequency distribution, we should expect more SOV than VSO, at least in this condition. However, participants shifted the distribution to produce more VSO at 56%, significantly different from the original distribution (t-test, $p < .001$). Similarly, the distribution of VSO-SOV in the Prominent CG context in Condition 2 was 42%-58%, since it consisted of half of VSO and 70% of SOV. Participants in this condition also shifted the distribution to produce more VSO at 51%, significantly different from the original distribution ($p = .001$). Figure 16 illustrates these results.

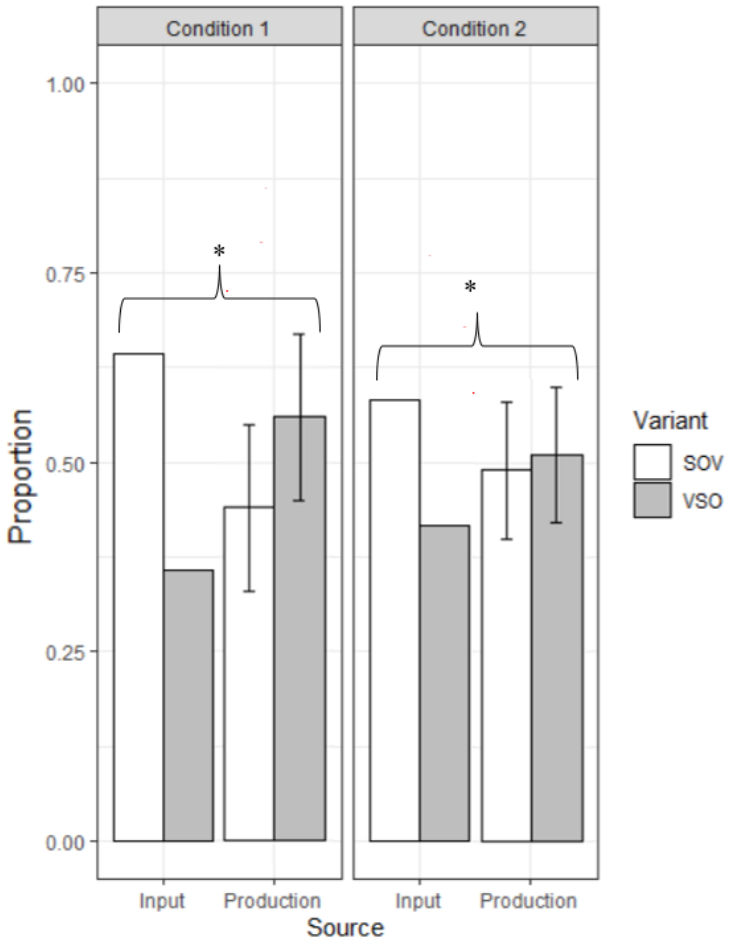

Figure 16. Distribution of VSO-SOV in the Prominent CG contexts in Condition 1 (left) and Condition 2 (right).

4.2.3. Discussion

Overall, the results from Experiment 4a confirm the earlier claim about learners' preference for discourse-unmarked variants. In both the Strict and Loose conditions, learners prefer the unmarked VSO, though learners in the Strict condition have a stronger preference compared to learners in the Loose condition.

What implication does this result have for the acquisition of natural languages? Languages can differ in the wh-strategies they allow as well as in the constraints operating on the marked variant. In particular, English wh-in-situ is more marked than French and

Brazilian Portuguese wh-in-situ, as discussed in Chapter 2. Taking English as the Strict condition and French and Brazilian Portuguese as the Loose condition, a generalization from the results of Experiment 4a would predict that English-speaking children have a stronger preference for fronted wh-questions than French- and Brazilian Portuguese-speaking children. Let us take a look again at the production rate of fronted questions by children across these languages. Note that the elicited tasks reported in Table 15 are all designed so that the production contexts allow for both fronted and wh-in-situ questions, similar to the Prominent Common Ground context in the ALL experiment.

Table 15. Children's production of fronted questions in an elicited task that allows for both question variants.

Language	**% of fronted questions**	**Discourse constraints on wh-in-situ**
English (Nguyen & Legendre, 2022)	98.73%	Strict
Brazilian Portuguese (Vieira & Grolla, 2020)	74%	Loose
French (Prévost et al, 2017)	74.1%	Loose

The results from Experiment 4a align with the acquisition data: children acquiring a language that has stricter constraints on wh-in-situ (English) show a stronger preference for fronted questions than children acquiring French and Brazilian Portuguese. While there are many other factors that can potentially contribute to the acquisition results, including language-internal ones like the frequency distribution of fronted questions versus wh-in-situ as well as language-external ones like how pragmatically well-designed the behavioral experiments are, the parallel trends between results from Experiment 4a (ALL) and results from natural language studies in Table 15 highlight how the degree of discourse markedness may influence learners' preferences.

In Experiment 4a, discourse markedness was expressed through frequency distribution. A strictly constrained variant is less likely to appear in a non-conforming context than a loosely constrained variant. Can discourse markedness be expressed in other ways that are different and independent from frequency distribution? This would be an interesting question to explore in future research.

Learning of the discourse requirement

Did participants actually learn the constraints underlying variation? Given the distribution of the variants, there are 2 possible rules participants could have inferred. Participants who noticed first that there were 2 different structures and tried to infer what constraints their use might arrive at *Rule a* (see below). Participants who noticed first the 2 different contexts and tried to infer what entails such difference might arrive at *Rule b.* Note that the rules are not mutually exclusive and only differ based on what basis participants relied on to infer them.

(ii) Rule a: VSO variant can be used in both contexts, but it's more appropriate for the SOV variant to be used in Prominent (P) contexts.

Rule b: Both variants are acceptable in P contexts, but in Non-prominent (NP) contexts it's more appropriate to use the VSO variant.

Rule a can be tested by calculating how many SOV utterances were used in P contexts out of the total number of SOV utterances. If participants obeyed *Rule a*, then the percentage of SOV used in P contexts should be higher than the percentage of SOV used in NP contexts. Given that the total number of NP and P contexts was fixed (each participant was given 5 P and 5 NP in the Production task), I compared the percentage of P in SOV with chance level (at 50%). In Condition 1, out of all SOVs, 73.3% of them were in P contexts, significantly higher than chance level ($p < .001$). With regard to individual performance,

78.6% (33 out of 42) of participants had a higher proportion of SOV in P than in NP. In Condition 2, out of all SOVs, 66.8% of them were in P contexts, significantly higher than chance level ($p < .001$). However, not as many participants (64.3%, or 27 out of 42) had a higher proportion of SOV in P than in NP contexts.

Rule b can be tested by calculating how many VSO utterances were used out of all NP contexts. If participants obeyed *Rule b,* they should use more VSO than SOV in the NP contexts. Since participants produced the variants themselves (i.e., the number was not fixed), I compared the distribution of VSO-SOV in NP versus P contexts. Once again, in both conditions, participants used significantly more VSO in NP than in P (M_{VSO_Cond1} = 81.1%, M_{VSO_Cond2} = 72.8%, ps < .001). *Rule b* was obeyed by the majority of participants (85.7% and 80.95% in Conditions 1 and 2).

The distributions of the variants across contexts are recapped in Table 16. Overall, participants in Condition 2 were not as sensitive to the constraints as participants in Condition 1, which is not unexpected given that the constraint in condition 2 was weaker.

Table 16. A recap of participants' performance with regard to *Rule a* and *Rule b*.

	Cond. 1	**Cond. 2**
(*Rule a*) SOV in P contexts out of all SOV	73.3%	66.8%
(*Rule a)* % of participants producing more SOV in P than in NP contexts	78.6%	64.3%
(*Rule b*) VSO structure out of all utterances in NP contexts	81.1%	72.8%
(*Rule b*) % of participants producing more VSO than SOV in P than in NP contexts	85.7%	80.95%

Between the 2 rules, participants seemed to be more sensitive to *Rule b* than *Rule a.* This is perhaps because it was easier to notice the different contexts based on illustrations not requiring language processing. On the other hand, in order to register the different word order structures, participants had to first figure out the meaning of each word in the language.

Finally, at the end of the experiment, there was an optional open-ended question asking the participants if they noticed anything about the different structures or different contexts in the experiment. While not every participant responded to this question, a few stated that they noticed the differences between face-to-face versus over-the-phone conversations, with one participant successfully capturing the distinction between the two conditions in terms of whether the addressee had the same information as the addresser. This data is reported in Appendix D.

Comparing Experiment 3 and Experiment 4a

In general, the results from Experiment 4a mostly replicate the results from Experiment 3. Participants in both experiments demonstrated successful learning of the artificial language and an overall preference for the unmarked VSO variant. The discourse

constraint is strictest in Experiment 3 (SOV: 100% Prominent CG, 0% Non-prominent CG), followed by Experiment 4a Condition 1 (SOV: 90% Prominent CG, 10% Non-prominent CG), and finally Experiment 4a Condition 2 (SOV: 70% Prominent CG, 30% Non-prominent CG). The results show that the overall produced VSO percentages parallel constraint levels: highest in Experiment 3 (70.7%), followed by Experiment 4a Condition 1 (66.8%) and Condition 2 (58.9%), confirming that the preference for the unmarked variant is influenced by the strictness of the constraints on the marked variant.

However, despite showing a weaker preference for VSO, curiously there was no SOV regularizer in either Condition 1 or Condition 2 of Experiment 4a. On the other hand, Experiment 3 induced the strongest preference for VSO, yet there were 2 participants who regularized to the SOV variant. Why were participants in Experiment 4a less likely to regularize to the SOV variant? I hypothesize that this was due to the total number of Declaratives in the exposure phase being reduced by 20%. As participants were exposed to fewer Declaratives, they might have gained a weaker sense of SOV being the canonical word order of the language, and in turn, gained a weaker sense of the SOV variant being the syntactically simpler structure.

4.3. Question variants in ALL versus in natural language

In the ALL experiments above, the contexts for questions were designed to closely match the actual contexts in which English wh-in-situ and fronted questions can be used. For wh-in-situ, I targeted one specific context among potential ones discussed in Chapter 2, Section 2.1., namely common ground requirements. Since having appropriate discourse-pragmatic contexts is important for learning, this section presents an acceptability

judgment task that seeks to validate the contexts in the ALL tasks. The acceptability task can also provide a glimpse into whether participants treat the artificial variants similarly to English wh-question variants.

4.3.1. Methods

Participants

A total of 104 (age range: 18-24, 66 females) participants were recruited through SONA (a cloud-based participant management system for universities). All participants were students at Johns Hopkins University, who self-identified as English native speakers.

Materials and design

The task assessed people's perception of English wh-questions under different contexts. The task included fronted questions and wh-in-situ questions in prominent CG contexts (5 of each variant) and in non-prominent CG contexts (5 of each variant), resulting in a total of 20 target questions. In addition, there were 5 embedded wh-questions serving as fillers for the task. Illustrations from the ALL experiments were reused to create the different contexts, with the artificial language stimuli being replaced by English sentences displayed visually (Figure 17).

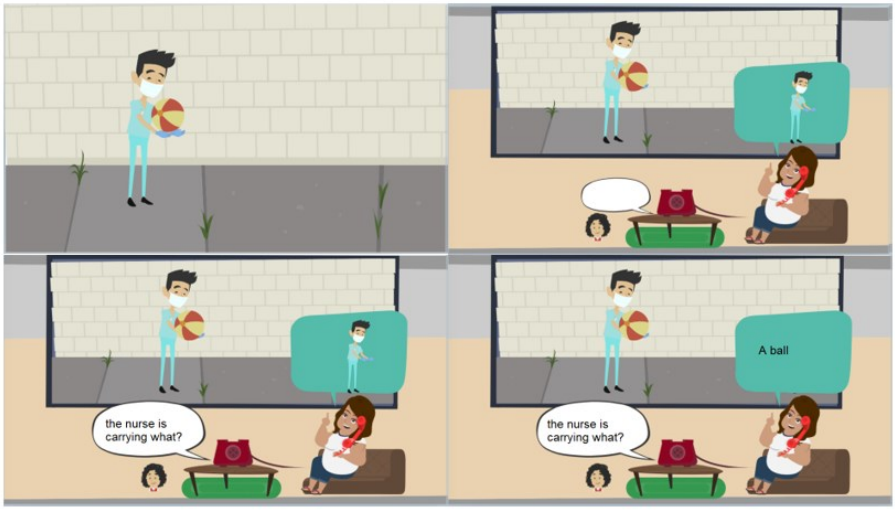

Figure 17. Example of a wh-in-situ in a Non-Prominent CG context.

For each question, participants were asked to rate the naturalness of it on a 7-point scale ranging from 0 to 6. A score between 0-1 indicates extreme unnaturalness while a score between 5-6 indicates that the sentence is perfectly natural.

Results

Overall, all target questions were perceived as natural, with fronted questions being rated as "very natural" (M = 5.97) and wh-in-situ questions being rated as "natural" (M = 4.63) in both contexts. Figure 18 illustrates the results.

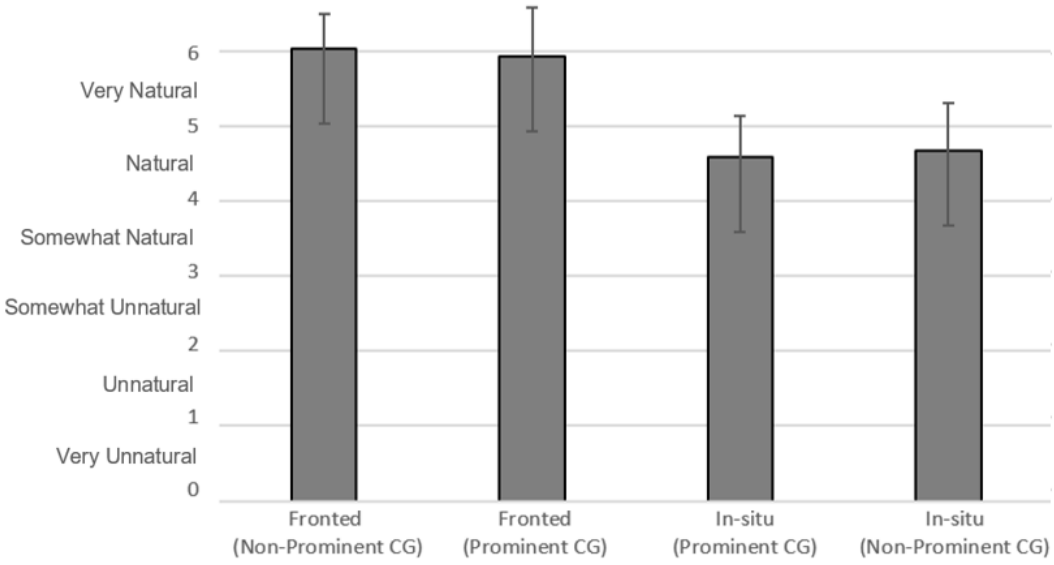

Figure 18. Perceived naturalness of fronted and in-situ wh-questions in Prominent and Non-prominent CG contexts

The data was analyzed using a linear regression mixed model with the rating being the dependent variable, Question Type (fronted versus in-situ) and Discourse being the fixed effects. The model also included Participant and QuestionNumber as the random effects and a by-participant random slope for Discourse. The results, summarized in Table 17, show that there is an effect of Question Type but no effect of Context. Participants rated fronted questions as being more natural than wh-in-situ regardless of contexts.

Table 17. Regression model of the acceptability task

	Estimate	*Std Error*	*t-value*	*P-value*
Intercept	6.03	0.12	49.9	< .0001 ***
Question Type = In-situ	-1.34	0.10	-12.3	< .0001 ***
Discourse = Prominent	-1.02	0.01	-0.93	.351

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

4.3.2. Discussion

Overall, the acceptability judgment task shows that participants perceived all question variants as natural. This implies that the contexts in the experiment were pragmatically plausible and appropriate for both types of wh-questions. The results indirectly validate the general context set-up for the ALL tasks.

While both variants were rated as natural, fronted questions were perceived as more natural than wh-in-situ questions. This is not surprising, given that wh-in-situ is the more marked variant, which can require more effort from the readers/listeners to accommodate. The similar judgment found for both prominent and non-prominent CG contexts for wh-in-situ questions, however, is more surprising. The prominent CG contexts in the ALL

experiments were designed based on contexts in which wh-in-situ is found in natural speech and included many characteristics of wh-in-situ, such as the power dynamics between the speaker and the listener (adult-child in the experiment). We may expect that participants would rate wh-in-situ in these contexts as more natural than wh-in-situ in non-prominent CG contexts. While surprising, this result is, however, not unreasonable. English wh-in-situ is complex, and while wh-in-situ questions frequently appear in prominent CG contexts similarly to those in the ALL experiments, there are also other situations where this variant can be found. For example, the wh-in-situ questions in the non-prominent CG contexts may have been interpreted as follow-up questions such as (55) (discussed in Pires & Taylor, 2007).

(55) A: I'm going to California.
B: And you're leaving when?

Follow-up questions do not require appropriate speaker-listener power dynamics or speaker knowledge/presupposition about the topic. The type of questions only requires that the question topic to be at issue, which participants may have inferred from the illustrations.

This result shows that the discourse requirements for English wh-in-situ are complicated, in contrast to the ALL experiments where participants were exposed to highly controlled and limited stimuli and as a result, could only learn from what they were exposed to. This does not invalidate the ALL methodology and our ALL results but points to the intrinsic limits of a single experiment.

To be able to learn the discourse requirements for the ALL task, participants could not simply have transferred their English knowledge to the artificial language. The fact that no effect for context was found in the English acceptability judgment task while the ALL participants did learn the discourse requirements (as discussed in Section 4.2.3) suggests

that they learned from the artificial language and did not go through the ALL tasks using their English knowledge.

4.4. Chapter 4 summary

Chapter 4 presented two ALL experiments and an acceptability judgment task to validate the ALL design. Experiment 3 exposed participants to a syntactically simple but marked (i.e., discourse-restricted) variant and a syntactically more complex but unmarked (i.e., not restricted) variant while controlling for the frequency of the two variants. The majority of participants showed a preference for the unmarked variant and many regularized to this variant in the production task. Experiment 4a further investigated how different degrees of markedness can influence the regularization tendency by comparing a strictly restricted condition against a loosely restricted condition. Participants who were exposed to the strictly restricted condition showed a stronger preference for the unmarked variant than participants who were exposed to the loosely restricted condition. Overall, results from Experiment 3 and Experiment 4a align with the child acquisition data discussed in Chapter 3. The results highlight the previously overlooked role of discourse markedness in learning multiple variants of a syntactic construction, suggesting that learners prefer the unmarked variant even when such variant is syntactically more complex. Moreover, this shows that regularization behavior is sensitive to domain-specific factors such as language discourse.

Chapter 5: Regularization & cognitive ability

As indicated in Chapter 1, there are two basic aspects of regularization that are in need of better understanding. The first focuses on factors conditioning the variants to be regularized, while the second focuses on the motivation and mechanism of regularization. The previous chapter (Chapter 4) was dedicated to exploring the former through behavioral ALL experiments. In this chapter, I pursue the second aspect of the problem, asking questions such as "why and how does regularization happen?". I hypothesize that learners – specifically children – regularize to reduce the cognitive burden when having to learn multiple variants (Section 5.1.). In particular, the relationship between regularization and cognitive ability (such as working memory) is explored through both an ALL experiment (Section 5.2.) and a computational model (Section 5.3.).

5.1. Regularization as a mechanism to reduce cognitive burden

So far, we have seen that children have a tendency to reduce variation through regularization. Adults do so too, under the right conditions. What motivates learners to regularize? This chapter explores this question and focuses on a different aspect of regularization: regularization as a mechanism to reduce cognitive burden.

Previous studies have shown that when being exposed to variation, children tend to regularize more than adults (Hudson Kam & Newport, 2009), and younger children regularize more than older children (Austin, 2010). Regularization thus seems to reflect developmental changes. The differences in regularizing behaviors between younger

children and older children, or children and adults, could be tied to the difference in available cognitive resources at various points in development.

Learning multiple variants of the same grammatical item is inherently cognitively taxing, and learners can reduce the load by regularizing to fewer variants. Specifically, Hudson Kam & Newport attribute regularization to limited memory capacity, which leads to difficulty retrieving certain forms while over-producing others. This explanation receives support from findings showing that a) high-load retrieval tasks (e.g., production) typically induce more regularization than low-load retrieval tasks (e.g., 2-alternative-forced-choice or grammaticality judgments) (Wonnacott & Newport, 2005; Schwab et al., 2018; Ferdinand, Kirby, & Smith, 2019; Saldana, Smith, Kirby, & Culbertson, 2021), and b) adults start to regularize more as the complexity of the learning task increases (Hudson Kam & Newport, 2019; Fedzechkina et al., 2012, 2017). Still, difficulty when retrieving certain variants does not necessarily mean that children learn all variants equally and only regularize at production; it is still possible that retrieval problems happen due to memory allocation during the learning stage, e.g., children may encode some variants differently, making it harder to retrieve them later, especially when the retrieval task is more cognitively taxing.

Perfors (2012), however, claims that memory limitations alone are not sufficient to motivate regularization; instead, both memory limitation and a prior bias favoring less variation need to work together to induce regularization. To demonstrate that memory alone does not lead to regularization, Perfors manipulates the cognitive load in an experiment. Participants were asked to learn an artificial language similar to the one in Hudson Kam & Newport (2009) while doing another task. There were 6 additional tasks

varied in terms of difficulty, including judging the sensibility of a sentence and/or the accuracy of an equation, memorizing 3 or 6 letters, and so on. Perfors found that participants under a more cognitive-taxing load did not regularize more. Perfors followed up with a computational model to investigate the potential effect of both memory limitation and a prior bias to favor regularization, concluding that both factors need to be present for regularization to happen. However, out of the three stages of memory (encoding, storage, retrieval), Perfors's experiment specifically targeted only the encoding stage, as the additional tasks only interfered with learning the language rather than producing it. Similarly, Perfors manipulated memory limitation by adding functions mimicking memory loss and memory decay to the model, which again only target memory encoding and storage, but not retrieval. Although there was no cognitive load manipulation of retrieval in the study, Perfors argued that retrieval could not be the explanation source, given that regularization does not cease as children's retrieval difficulties lessen (citing the case of Simon, who did not gradually become more inconsistent in his production). However, as discussed in Chapter 1, Section 1.2., Simon's situation constitutes a special case of variation -- the inconsistencies in his input are ungrammatical. It is possible that as he got older, he received more input from other signers besides his parents, which helped shape his production. Austin (2010) in fact finds that older children (7-to-8 year-olds) regularize less than 5-to-6-year-old children. Taking all these factors into consideration, it doesn't appear that Perfors' rejection of a memory account for regularization is (fully) justified.

Culbertson, Smolensky, & Legendre (2012) propose cognitive biases as a motivation for at least some cases of regularization in an ALL task. They suggest that learners prefer and regularize to variants that are *harmonic* (e.g., showing consistency in

ordering in the case of learning the respective word order of adjectives, nouns, and numerals) to reduce the computational cost of processing. It is, however, unclear in Culbertson et al.'s proposal whether regularization happens at the stage of encoding, in which case learners find it easier to learn harmonic variants, or regularization happens at the stage of retrieval, in which case learners prefer to produce harmonic variants despite learning all variants.

Compared to Hudson Kam & Newport's memory proposal, Culbertson et al.'s proposal is more specific in terms of predicting the direction of regularization (which variants learners are more likely to regularize to), but the two accounts are not necessarily mutually exclusive. In general, regularization seems to happen to reduce cognitive burden, though it is not clear yet at which stage (encoding, retrieval, or both) regularization happens.

In the next section, I will first investigate the relationship between regularization and working memory using a different method from previous attempts. I will then report on a simple computational model that attempts to simulate regularization behavior using the idea of an input filter, i.e., learners only make use of a subset of their input. Overall, the work reported in this chapter suggests that there is a relationship between regularization and working memory.

5.2. Experiment 4b: The relationship between working memory and regularization

In Chapter 4, we saw that the more cognitively taxing Production task induced more regularization than the 2AFC task. This supports the idea that regularization happens to

reduce cognitive load, and hints at the relationship between regularization and other general cognitive abilities like working memory. However, while working memory is often invoked (e.g., Hudson Kam &Newport, 2009; Perfors, 2012), to the best of my knowledge, there has not been a study that directly tests whether working memory predicts regularization behavior. A behavioral experiment designed to specifically test for the correlation between memory score and regularization rate attempts to fill that gap.

5.2.1. Method

Participants

All 84 participants from Experiment 4a subsequently participated in this experiment. Due to technical issues, data from 2 participants were not properly recorded, leaving us with a total of 82 data points.

Procedure

After completing the ALL task, participants were asked to complete a series of working memory tasks. The order of each working memory task was randomized.

Materials

The three working memory tasks were selected so that they covered both visual and auditory memory, as well as both linguistic and non-linguistic stimuli. Additionally, these are classic working memory tasks that are on the harder side and they can yield good sample variance.

a) Visual backward digit span (BDS)

The backward digit span is one of the most commonly used tasks to assess working memory capacity (Hilbert et al., 2015). Participants are asked to watch a string of digits appear on the screen and repeat them (through typing) in reverse order. The task starts with

4 digits and increases as participants proceed. The number of digits increases by one when participants correctly recall the string and decreases by one when participants fail two trials in a row. This task requires participants not only to store the digit string in their memory but also to manipulate it. Participants are scored based on the length of sequences that they can recall.

b) Audio 3-back task

The N-back task is also used extensively as a paradigm to measure working memory (Owen, McMillan, Laird, & Bullmore, 2005; Jaeggi, Buschkuehl, Perrig, & Meier, 2010). In this task, participants listen to a stream of letters, and they are asked to decide for each stimulus whether it is the same letter as the stimulus presented *N* times (in this case, N = 3) before. For example, in the string "H K K T Z Q H Z M", participants need to respond by pressing the instructed key when they hear the second Z, as Z was also presented 3 trials earlier. Participants are scored based on their hit rate (correctly respond when a stimulus is presented 3 trials earlier) and false alarm rate (incorrectly respond when a stimulus is a non-target).

Interestingly, researchers have found that results from the N-back task often do not correlate with other classic working memory tests like the digit span task (Miller et al., 2009), which has led to concerns about the validity of N-back. However, results from neuro-imaging studies have shown that engaging in the N-back task increases activation in areas commonly activated during working memory tasks, (Owen et al., 2005). While further studies have supported the claim that N-back is a valid measurement for working memory, it has been argued that N-back demands the ability to recognize and shift attention while digit span demands serial recall (Kane et al., 2007; Frost et al., 2021), which explains

the potential lack of correlation in results from the two tasks. N-back is claimed to be useful to predict individual differences in high cognitive functions, as it allows the cognitive load to be manipulated in a straightforward way (Jaeggi et al., 2010). Thus, the use of both N-back and backward digit spans together yields a more comprehensive picture of participants' cognitive and working memory abilities.

c) Sentence repetition task

The Sentence Repetition task (SRT, Baddeley, Hitch, & Allen, 2009) requires participants to type back speech tokens of increasing number of syllables. The task contains 15 trials and takes less than 10 minutes to complete. The first trial is a 7-syllable sentence (e.g. "the birds were singing all day"), and each following trial presents a new sentence that is one or two syllables longer than the previous trial. The task progresses until the final sentence, which has 26 syllables. A list of all the sentences used in this task is listed in Appendix E

Sentences used in SRT, adapted from Van Hedger. Participants are scored based on the maximum syllable length they are able to recall, as well as the proportion of words they correctly identify in each trial.

To perform the task, participants need to process the phonological representation of the sentence and extract its meaning, then store and retrieve the sentence from memory to reproduce it. In general, the SRT is designed to tap into linguistic ability and working memory (von Eckardt & Potter,1985; Marinis & Armon-Lotem, 2015). Studies have reported that both linguistic interferences (such as the use of non-words or badly-formed sentences) and short-term memory impairments can lead to poor performance in SRT (Hanten & Martin, 2000; Polišenská, 2011).

5.2.2. Results

Working memory performance

From the three working memory (WM) tasks, a composite score was calculated for each participant using the *multicon* package in R. A composite score can capture a more complete picture of memory and processing abilities (Wilde, Strauss, & Tulsey, 2004) than individual score from each WM task. The distribution of WM scores is reported in Figure 19, showing that there is some between-subject variability.

A regression model suggests that Gender is a significant predictor for the Composite score, while Age is not. In particular, Male participants had higher scores than Female participants (M_{Male} = 12.52, M_{Female} = 11.06, t = 2.43, df = 80.8, p = .01). Researchers have sometimes found gender differences in WM (e.g., Harness et al., 2008; Saylik, Raman, & Szameitat, 2018), however, this result is not a main point of the present investigation; it is only used to ensure that all the appropriate effects are included in the main mixed-effect model.

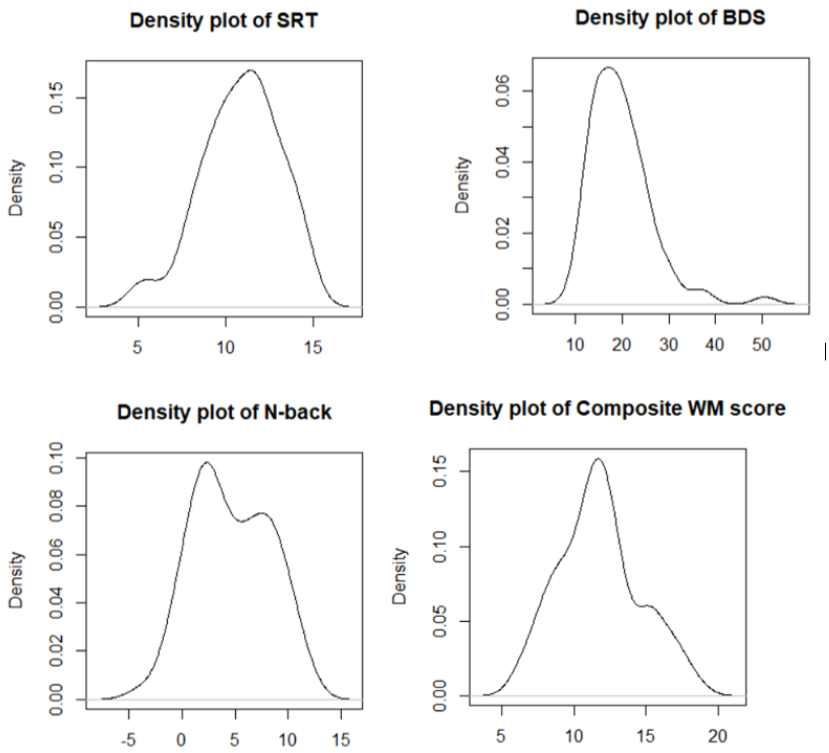

Figure 19. Distribution of working memory score across tasks.

Correlation between working memory score and regularization

The data was submitted to a linear mixed-effect model using the lme4 package in R (Bates et al., 2015). Our model included the Regularization rate from the ALL study reported in Chapter 4 (i.e., how much a participant deviated from the original 50:50 VSO-SOV distribution) as the dependent variable, with Condition (1/Strict versus 2/Loose), Composite WM score, and Gender being the main effects, and interaction and Participant as the random effect (Table 18). There is a significant effect for WM score, Condition, as well as the interaction between WM score and Condition.

Table 18. Results of the linear regression model to test the effect of Condition and Composite WM score in predicting Regularization rate.

	Estimate	*Std. Error*	*t-value*	*P*
Intercept	0.59	0.18	3.22	.001 **
Condition	-0.29	0.11	-2.72	.008 **
CompositeWM	-0.03	0.01	-2.34	.02 *
Condition * CompositeWM	0.018	0.009	2.09	.03 *
Gender = Male	0.08	0.05	1.57	.12

*Note: * $p < .05$; ** $p < .01$; *** $p < .001$*

Further investigation suggests that the Composite score only significantly correlates with the Regularization Rate in Condition 1 ($r_{pearson} = -0.32$, $p = .02$) but not in Condition 2 ($r_{pearson} = -0.06$, $p = .66$). The negative correlation in Condition 1 suggests that participants who have a greater rate of regularization tend to have a lower composite working memory score. The lack of correlation in Condition 2 is likely because there is not enough regularization in this condition.

To avoid multicollinearity, scores from the three individual WM tasks were not included in the regression model reported above. Interestingly, a further look at the individual task results reveals that none of them has a significant correlation with the regularization rate in Condition 1, as shown in the correlation matrix in Figure 20. Specifically, the N-back, SRT, and BDS tasks all have non-significant correlation with the regularization rate ($p > .05$). Only the composite score has a significant correlation relationship at $r = -0.32$ and $p = .02$. Recall that the three WM tasks were selected to cover stimuli from different domains and modalities (linguistic versus non-linguistic, visual versus auditory). The tasks also tapped into different aspects of WM (e.g., storing and

manipulating information, serial recall, recognition, etc.). This suggests that a single type of WM measure is not sufficient to evaluate the relationship between WM and regularization.

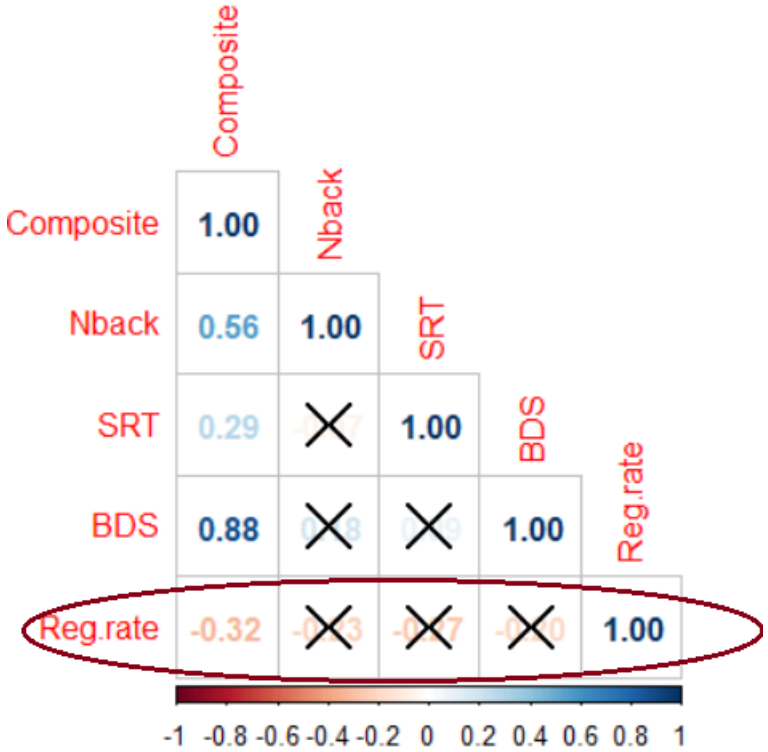

Figure 20. Correlation matrix between the composite individual WM score and regularization rate in Condition 1. Any correlation coefficient that has a p-value higher than .05 (non-significant) is crossed out.

Correlation between N-back and constraint violation

While there was no correlation between individual WM tasks and the regularization rate, there was a correlational relationship between the N-back task and the rate of violating discourse requirements. Recall that the SOV interrogative was the marked variant and was restricted to the Prominent CG contexts. Any instance of SOV in the Non-prominent CG context thus was marked as a violation. The N-back score significantly correlated with the rate of violation for participants in both Condition 1 and 2, so I congregated the data of all participants together. Overall, the correlation between the N-back score and the rate of

violation was significant ($p = .03$), albeit the magnitude of the correlation was weak ($r_{pearson} = -.27$). The negative correlation suggests that participants who violate discourse constraints more also tend to have lower performance in the N-back task, as illustrated by Figure 21.

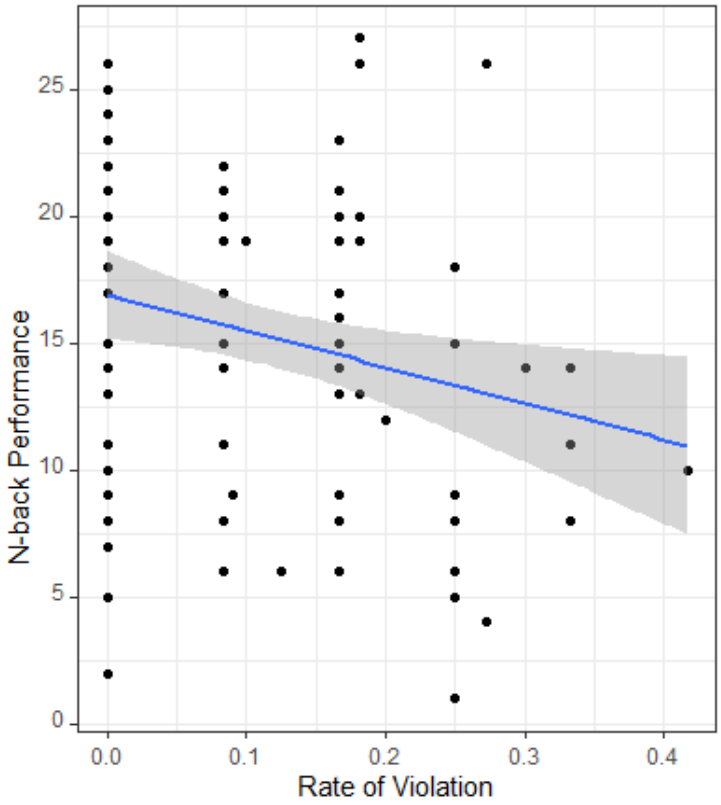

Figure 21. Negative correlation between Violation rate and N-back performance. The blue line represents the linear correlational relationship between the Violation rate and N-back score at r = -0.27. The shaded area represents the 95% confidence level interval.

5.2.3. Discussion

Working memory and regularization

Many studies have asked whether working memory plays a role in regularization, but the reported results have been mixed. Overall, support for the role of working memory in regularization has come from experimental studies showing that regularization happens more when the memory retrieval demand is higher (Hudson Kam & Chang, 2009; Hudson

Kam & Newport, 2009). At the same time, there have been studies suggesting that increasing cognitive load does not change learners' regularizing behavior (Perfors & Burns, 2010; Perfors, 2012). However, Perfors & Burns (2010) and Perfors (2012) tested for the role of memory and cognitive abilities by putting participants under interference tasks, for instance, participants had to learn the language while solving equations or reading sentences aloud. This approach introduces potential confounding factors, such as divided attention, the ability to manage interference, or the ability to suppress irrelevant information.

Experiment 4b, in contrast, has explored the relationship between working memory and regularization through a simpler approach. I hypothesized that if working memory contributes to regularization, there would be a correlation between working memory performance and the rate of regularization. Moreover, if regularization happens as a mechanism to reduce cognitive burden, as has been proposed in the literature (e.g., Hudson Kam & Newport, 2009; Schwab et al., 2018), then participants with lower working memory performance would regularize more. The results from Experiment 4b confirm these hypotheses. Importantly, the regularization rate correlates with the composite score but not with any individual working memory score. This suggests that the cognitive processes involved in regularization are complex and a single working memory task, which usually only captures one aspect of working memory (e.g., recall versus recognition) in one domain (e.g., linguistic versus non-linguistic) and modality (e.g., auditory versus visual), is not sufficient to reflect them.

Besides this theoretical contribution, results from Experiment 4b also have implications for the methodologies used in research on regularization. The correlation

between working memory performance and regularization rate was only found in Condition 1 – the Strict condition, in which participants regularized significantly more than participants in Condition 2 – the Loose condition. The weaker regularization tendency in Condition 2 made the regularization rate of participants more condensed around the mid-range, resulting in a smaller range and lower variance. I propose that the lack of a significant correlation between working memory and regularization rate in Condition 2 is likely due to insufficient regularization. This result suggests that before concluding that an insignificant correlation indicates a lack of relationship, regularization researchers should check to see if the insignificant correlation may have stemmed from a weak or no regularization. For example, in Perfors (2012), the *regularization index* was defined as the proportion of trials in which the participant produced their most frequent variant. While participants who regularized more would indeed have a higher regularization index, this measurement wasn't sensitive to whether a participant regularized or simply reproduced the input distribution. For instance, if a variant was presented 60% of the time in the input and a participant produced that variant 60%, the *regularization index* as defined by Perfors would still be 60%, while in reality the participant did not regularize. Thus, the lack of correlation between working memory capacity and regularization index in Perfors (2012) could happen due to the so-called regularization index being an inappropriate measurement for regularization and/or a lack of actual regularization behavior induced by the task. Therefore, when investigating the relationship between regularization and another factor, such as working memory, it is important to ensure first that the task yields (strong) evidence of regularization.

N-back as a measure of pattern recognition

The negative correlation between N-back performance and discourse constraint violation in the ALL experiment was unpredicted but nevertheless constitutes an interesting finding. The N-back is the only working memory measurement that had a significant relationship with the discourse constraint violation, while the other two working memory tasks (BDS and SRT), as well as the composite score, all returned a non-significant correlation. What is unique about the N-back task that separated it from the other two tasks?

The N-back task focuses more on recognition than recall. In the BDS task, participants had to store the digit string in WM, manipulate the string to produce the reverse order, and recall the reversed string by typing out the numbers. In the SRT task, similarly, participants had to recall the sentence they heard by typing it out. On the other hand, the N-back task is a continuous recognition task in which participants need to determine whether a stimulus was previously presented N (in our study, N = 3) steps back (Pelegrina et al., 2015). Participants need to continuously update, maintain, and replace items in their working memory (Soveri et al., 2017; Mencarelli 2019). For example, suppose that participants are given the string "H K K T". When they hear "T" and notice that it does not match the letter presented 3 trials earlier, they can drop the letter "H" from their working memory to reduce the load, but still need to maintain "K K". Participants need to always maintain at least 3 items, however, these 3 items are constantly updated as new stimuli are added to the string. The total score for the N-back task is the number of hits minus the number of false alarms. Thus, participants face two goals: correctly respond when a stimulus is presented 3 trials earlier (increasing the number of hits) and avoid responding when a stimulus is a non-target (reducing the number of false alarms).

Given the characteristics of N-back and the correlation result obtained in the experiment, I hypothesize that there might be some similarities between the task and the rule recognition process in the ALL task. In the N-back task, participants were exposed to a continuous stream of stimuli and needed to evaluate whether a new letter matched with the letter presented 3 steps earlier. In the ALL task, participants were exposed to a continuous stream of trials (sentences in their context), and had to infer the constraints they needed to evaluate the new sentence against previous sentences. Learners of the artificial language (and of natural languages more generally) also need to maintain linguistic information from the input while formulating, updating, and replacing hypotheses about the language. The ALL task, however, is much more complex and involves other processes such as pattern recognition and frequency tracking, which is likely the reason why the correlation was weak.

There have not been a lot of studies on the relationship between N-back and rule-learning in languages, and many studies that look into the relationship between N-back and language learning only use N-back as a measure of working memory (e.g., Hansen et al., 2016; Lukasik et al., 2018). The correlation between the discourse constraint violation rate and N-back along with the lack of correlation between the violation rate and other working memory tasks together suggest that the N-back task may share some similarities with the rule-learning process in language. Further studies are needed to confirm and explore this hypothesis in more detail.

5.2.4. Experiment 4b Summary

Experiment 4b supports the hypothesis that regularization happens to reduce cognitive burden(s). Participants with a lower composite working memory score also

displayed a higher regularization rate. This is in line with previous studies which find that younger children regularize more than older ones (Austin, 2010), and children regularize more than adults (Hudson Kam & Newport, 2009; Schwab et al., 2018). The experiment also finds a weak correlation between the N-back task and the rate of discourse constraint violation; in particular, participants with lower N-back performance violate the discourse constraints in the language more. This result suggests that the N-back task can potentially tap into some aspects of rule learning, and future studies can explore this relationship further.

In the next section, I continue to explore the relationship between cognitive resources and regularization through a computational model. The use of a computational model allows me to test many factors that are impossible to manipulate in an experimental setting, such as the amount of intake data and the initial bias of the learners.

5.3. Testing the input-filter hypothesis: a computational model

It is generally acknowledged that children's cognitive skills are not yet fully developed, and thus are more limited than adults', and adults outperform children in many cognitive areas, such as working memory and executive functions. At the same time, it is well-accepted that children acquire language much more easily and effortlessly than adults, which seems to be a paradox considering that many theories on language acquisition also highlight the importance of cognitive skills.

Newport (1990) attempts to resolve this problem through the Less is More hypothesis. The hypothesis initially arose from studies comparing language abilities between first- and second-language learners, as well as between early and late second-

language learners. These studies found that late learners' language performance is typically worse than early learners', suggesting a critical period for language acquisition. In particular, the Less is More hypothesis aims to explain the underlying mechanism of the critical period by proposing that children's limitations actually give them an advantage in learning a language. Specifically, children's limited cognitive skills[16] force them to "start small", i.e., only considering a small amount of input, which in turn enable them to isolate and analyze the components of the linguistic input more effectively.

Elman (1993) offers a connectionist computational model to support the Less is More hypothesis. The model was trained on a set of simple noun-verb (e.g., *cat* runs) and noun-verb-noun sentences (e.g., *cats chase dogs*), as well as sentences containing multiple embeddings (e.g., *girl who raises birds sees boy who feeds dog who chases cats).* Elman claims that, compared to neural networks that are fully formed and 'adultlike', neural networks that were constrained to have limited working memory at first before "maturing"

[16] The Less is More hypothesis does not specify the source of the limitations, whether they come from limited cognitive capacity or working memory, a smaller amount of input, or a lack of prior knowledge that may interfere with the learning process. While a number of studies from both computational (e.g., Elman, 1993; Goldowsky & Newport, 1993) and experimental (e.g., Cochran, McDonald, & Parault, 1999; Chin & Kersten, 2010) claim to support the Less is More hypothesis, these results in fact support different variants of the hypothesis as they target different sources of limitations. For example, Gopnik, Griffiths, & Lucas (2015) and McDonough, Choi, and Mandler (2003) look into the advantage of having limited prior knowledge, Goldowsky & Newport (1993) and Elman (1993) look into limited cognitive capacity, while Chrysikou, Novick, Trueswell, & Thompson-Schill (2011) investigate limited executive control in language learning. Since the discussion of regularization mainly focuses on cognitive capacity and working memory, I also limit the discussion of the sources of limitations in the Less is More hypothesis to these.

performed better in processing complex sentences with multiple embeddings. Specifically, when the model was trained on the whole corpus at full capacity, it was unable to learn the grammar. When the input or the memory capacity was limited at first and improved gradually, the model successfully learned the grammar. This finding is taken to support the importance of "starting small" in language acquisition (though note that Rohde & Plaut (1999) could not replicate Elman's results).

Wirth regard to the original Less is More hypothesis, in a follow-up study, Goldowsky & Newport (1993) introduce the idea of an input filter by using a simple computational model. The model's main task is to map a linguistic form with its corresponding meaning. In this model, both the form and meaning are made up of *features* – for a form, this could be similar to syllables or morphemes, and for a meaning this can be thought of as semantic features. To do the mapping, the model needs to pay attention to the co-occurrences of form features and meaning features. However, the model ends up producing a lot of noise, i.e., chance occurrences that do not reflect a real meaning relationship. Goldowsky & Newport propose that this reflects a problem in language acquisition that "the data available to a language learner (can) support many different hypotheses about the underlying system", and conclude that more data is needed to successfully narrow down the right hypothesis. To address this issue, the authors thus impose an input filter on the model that randomly removes half of the features of each word. Goldowsky & Newport reason that a random filter would be the most conservative assumption to implement, and acknowledge that in reality, there are likely biases about which part of the data can be ignored and which part requires attention. With the input filter, the signal-to-noise ratio improves. Goldowsky & Newport claim that the input filter

forces the model to concentrate on small units, and in doing so, successfully reduces the amount of noisy signal that occurs by chance. As the model "matures", the filter can become less restrictive over time and allow the model to learn fine-grain features that are lost during the filtering process. Goldowsky & Newport make a connection between the original model with adult learners learning a second language, as well as between the model with an input filter and child learners. The input filter provides child learners with "cleaner" data, while adult learners will encounter many competing analyses of the language, leading to relatively poor learning. In sum, the general idea of an input filter is that, when cognitive resources are limited, children may not be able to learn everything presented to them in the input but may only focus on a small subset of it.

This idea of an input filter is also captured in the distinction between *input* and *intake* that other researchers have proposed (Gass, 1997; Gagliardi & Lidz, 2014; Omaki & Lidz, 2015): *Input* is the data available in the environment, while *intake* is the input data that is utilized by the language acquisition mechanism to make inferences about the grammar of the target language. The intake of children is smaller early on and gradually expands as they age. In other words, input filtering does not mean that the child "filters out" or actively ignores everything outside of their intake, but instead it suggests that the child can only make use of a subset of their input. This is not an unreasonable assumption, given that children's working memory is limited and the amount of linguistic input they are exposed to may be large; children are also known to be capable of selective learning (Kinzler, Corriveau, & Harris, 2010; Sobel & Kushnir, 2013).

In Chapter 3, I briefly discussed a study by Derks & Paclisanu (1967) in which children and adults were presented with two unpredictable stimuli, one appearing 70% of

the time and one appearing 30% of the time. In a task where they had to predict which stimulus was appearing next, children, by consistently regularizing to the more-frequent stimulus, ended up outperforming adults who tried to probability-match the input. Perhaps what helped children outperform adults in this simple task is also what helped children outperform adults learning a second language, as reported in Newport's (1990) study. This leads me to propose that the tendency to regularize may be a result of the input filter, which is motivated by limited working memory.

Experiment 4b has already established that there is a relationship between working memory and regularization. In the follow-up study below, I explore the idea of an input filter as a mechanism that may contribute to making children's production more systematic and regularized than their input. I begin the section by reviewing two previous and similar computational models (5.3.1) before presenting an alternative and novel one (5.3.2.).

5.3.1. An input-filter model and its relationship with regularization

Besides Goldowsky & Newport's computational model discussed above, a number of computationally-oriented studies have applied the idea of an input filter to solve learning problems when the input data is noisy or contains unpredictable variation. Perkins, Feldman, & Lidz (2017) and Schneider, Perkins, & Feldman (2019) model the acquisition of verb transitivity and English determiner agreement, respectively. Both studies find that by treating the variation in the input as "noise" and filtering them out, the model arrives at a more regularized and consistent grammar.

An input-filter model to learn verb transitivity

Perkins et al. (2017) observe that learning verb transitivity is not an easy task due to the presence of non-canonical clause types like wh-questions next to declaratives in the

input. For example, data from (56a) and (56b) may lead a learner to the conclusion that "read" is transitive and intransitive, while "review" can only be transitive. However, for learners without adequate knowledge about wh-dependencies or syntactic transformations, (56c) and (56d) may misguide them to conclude that "read" and "review" can both be intransitive. Interestingly, experimental evidence suggests that verb transitivity knowledge develops at 15-16 months, before wh-dependency knowledge which develops at 18-20 months (Seidl et al., 2003, Gagliardi et al., 2016).

(56) a. John reads a book. Mary reviews the exam.

b. John reads. *May reviews.

c. What did John read? What did Mary review?

d. John likes the book that Mary reviewed.

A potential strategy for young learners to avoid being misguided by these complex sentences is to perhaps put them aside until the learners reach a more mature stage of linguistic knowledge -- which raises the question: how do children know which input to ignore and which input to learn from? Perkins et al. draw from a Bayesian learning model to show that no prior knowledge about filtering criteria is needed. The model uses a set of data taken from four CHILDES corpora that contain sentences with the 50 most frequent action verbs. It simply assumes that some of its input are not reliable sources of information. The model starts with randomly initialized values for the probability of having noise (e.g., an erroneous parse due to the learners' limited knowledge) and the probability of such noise generating a direct object. Using these values, the model then calculates the posterior probability of each transitivity category (transitive, intransitive, or alternative) for each verb given the data, then uses the sampled transitivity values to sample new values for the initial noise probabilities. This process is repeated over and over until the model

converges. Note that the model makes a few assumptions. For example, the model only infers a single value for the probability of noise, which assumes that each verb has the same probability of being parsed wrongly by the learner. This simplifies the learning problem but may be unrealistic in real life, as the probability of noise is likely to vary from verb to verb as well as in different contexts.

Overall, this model assumes that its input contains noise and its goals are to infer both the probability of transitivity categories and the rate of noise from the distribution of the data. Perkins et al. report that their model performs substantially better than a model that lacks an input filter and twice better than a random chance model. The model shows that by simply having an input filter that assumes a fixed noise rate (inferred from the data distribution), with the learning performance improved as there are fewer erroneously parsed data. In reality, children may rely on many factors besides the verbs' distribution (such as prosody, the presence of a wh-word, or the context) to make a judgment about the reliability of the data.

An input-filter model to learn determiner agreement

In Chapter 1, Section 1.3., I briefly summarized the case of Simon, a child learning ASL from non-native parent signers. Despite being exposed to an input that contained many inconsistencies and grammatical errors, Simon's production was much more regularized and almost indistinguishable from children learning from native signers. Inspired by this study, Schneider et al. (2019) attempt to model the acquisition of English determiner agreement in children exposed to input from non-native parents. They used data from late learners of English, who often make errors in determiner agreement. Schneider

et al. considered number and countability in the determiner system -- for example, "*this"* can only be used with singular but not plural nouns.

Similar to the model in Perkins et al. (2017), this model assumes that the input consists of both signal and noise. The model also assumes that there is a noise rate, i.e., the proportion of input that is generated by noise instead of the grammar, though it does not know *a priori* if an utterance is a signal or noise. For each utterance, the model needs to infer a binary value about the type of nouns that is allowed, for example, α_1 represents singular nouns, α_2 represents plural nouns, α_3 represents mass nouns, and $\alpha_1 = 1$ signals that the determiner can be used with singular nouns, $\alpha_2 = 0$ signals that it cannot be used with plural nouns. The model also infers a binary value *e* where *e* = 1 if the utterance is noise and e = 0 if the utterance is grammatical. *e* is drawn from a Bernoulli distribution with the rate of noise ε, which is initialized randomly. Given the rate of noise, the model samples the α value for each class of determiners, then uses the sampled α value to sample new values for the initial rate of noise. The sampling process is repeated for 1000 iterations of Gibbs sampling.

Overall, the model performs significantly better at 59% accuracy, compared to a non-filtering model which has a 29% accuracy, suggesting that the model is able to separate signal from noise. However, similarly to the model in Perkins et al. (2017), this model assumes that there is a single constant value for the rate of noise that applies uniformly for each class of determiner.

In the next section, I will turn to describe an alternative, non-parametric Bayesian model for learning wh-question variants. Instead of setting a fixed noise rate, the model

explores how the learning outcome may be different given different filtering rates and levels of bias.

5.3.2. A non-parametric Bayesian model of wh-question learning

The model here is based on the experimental results of wh-question variants reported in Chapter 3. The goal of the model is to examine a *possible* mechanism (1) by which children regularize their production and (2) captures many of the proposals other researchers have put forward -- I do not claim that this is how children actually acquire multiple variants.

Importantly, I only consider here the basic learning problem, i.e., how a model may learn multiple wh-variants given their distribution in the input, without considering other pragmatic or discourse factors. This is done so that the model can be easily modified to work with other regularizing cases of different types of variants as well. While it is not a realistic representation of how children actually acquire language, the model can document possible mechanisms through which children arrive at a more regularized production from an input containing variation. I first describe the model before turning to a discussion of its theoretical contribution.

General property

The defining property of non-parametric Bayesian models is not an absence of parameters (as the name might suggest), but rather the ability to infer both the number of parameters as well as their values from data. In the case of clustering, non-parametric models can make inferences about how many clusters underlie a given data set and the internal properties of each cluster. Due to this flexibility, non-parametric models have been successfully applied to many clustering problems in linguistics and cognitive science,

including syllable, morpheme, and word segmentation (Goldwater, 2007; Johnson, 2008; O'Donnell, 2015; Seshadri, Remes, & Räsänen, 2017), phonetic category learning (Lee, O'Donnell, & Glass, 2015), syntactic/semantic rule learning (Abend et al., 2017), and psychological category induction (Sanborn, Griffiths, & Navarro, 2010).

While language-internal variation in wh-questions does exist, the number of types within a given language is likely to be quite small and certainly not unbounded (compare the number of lexical items that must be identified in word segmentation). I adopt the non-parametric approach because, unlike classical parametric finite mixture models, it does not force the learner to commit to the existence of a particular number of clusters (here, question types) in advance of analyzing the input data. For purposes of implementation, an upper bound of $K = 10$ is placed as the maximum number of wh-question types that the model can learn. This bound is much larger than the subset of questions analyzed here; though it can easily be raised to work with languages with more than ten wh-question types, such as French.

Data

Main clause English wh-questions were collected from four CHILDES audio corpora: HSLLD (Dickinson & Tabors, 2001), Snow (MacWhinney & Snow, 1990), Van Houten (Van Houten, 1986), and Weist (Weist & Zevenbergen, 2008). Questions were annotated by hand for wh-fronting and subject-auxiliary inversion. To extract the prosodic properties, the questions were forced-aligned with the Montreal Forced Aligner (McAuliffe et al., 2017) and subsequently analyzed using the PRAAT software (Boersma & Weenink, 2019) to extract the duration (measured in ms) and the f0 contour (final Hz - initial Hz) of the vowel in the wh-word.

Because the audio data available on CHILDES was limited, the 88 utterances coded as just described were reserved for testing only. The model was trained on 2000 simulated instances randomly generated according to the values in Table 19a. As further spoken examples of wh-questions in child-directed speech become available, I anticipate being able to train and test the model entirely on natural utterances in a future study.

Each question type inferred by the model consists of a probability distribution over several properties. Instead of examining an exhaustive list of linguistic properties, the present study is limited to the two morphosyntactic properties and two prosodic properties that are most relevant to our English case studies. A more comprehensive typology of wh-questions across languages would require more properties, and the model can be straightforwardly modified to include additional parameters. The morphosyntactic properties are discrete variables that can take on two values (1 corresponding to the presence of a property and 0 corresponding to its absence). They include the position of the wh-word and the inversion status of the auxiliary. The prosodic properties consist of two continuous variables: the duration (milliseconds) and F0 contour (ΔHz) of the wh-word. All fronted wh-questions require auxiliary inversion while in-situ questions do not, thus the values for wh-fronting and inversion are identical. As for the continuous variables, echo questions typically have longer durations of the wh-word and a rising intonation which is expressed as positive values of ΔF0. Probe questions and fronted questions have a shorter duration on the wh-word, and negative or close to zero values of ΔF0 to indicate falling or flat intonation.

Finally, the frequency distributions of question types in the simulated data are: 84% fronted questions, 9% probe questions, and 7% echo questions, based on the distribution

found in child-directed speech. The probability distributions of the morphosyntactic and prosodic properties within each question type were fit to the child-directed utterances.

Model specification

The non-parametric model proposed here is technically a Dirichlet Process Mixture Model (e.g., Gershman & Blei, 2012), as specified below.

Cluster probabilities

$$\alpha \sim Gamma(1,1)$$
$$v_\ell \mid \alpha \sim Beta(1,\alpha) \qquad \text{for } \ell = 1,\ldots,K-1$$
$$\alpha \sim Gamma(1,1)$$
$$w_1 = v_1$$
$$w_k = v_k \prod_{\ell=2}^{k-1} 1 - v_\ell \qquad \text{for } k = 2,\ldots,K-1$$
$$w_K = \prod_{\ell=1}^{K-1} 1 - v_\ell$$

Parameters of each cluster

$$p_{kj} \sim Beta(1,1) \qquad \text{for } j \in \{WhFront, Inver\}$$
$$\mu_{k\ell} \sim Normal(M_\ell, S_\ell) \text{for } l \in \{WhDur, Wh_{\Delta F0}\}$$
$$\log \sigma_{k\ell} \sim Normal(M_\sigma, S_\sigma)$$

Distribution of observations

$$p(\mathbf{y}_i|\mathbf{w},\mathbf{p},\mu,\sigma) \qquad \text{for } n = 1,\ldots,N$$
$$= \sum_{k=1}^{K} w_k \prod_j Bernoulli(y_{ij}|p_{kj}) \cdot \prod_\ell Normal(y_{i\ell}|\mu_{k\ell},\sigma_{k\ell})$$

The cluster mixture weights w_k are given by a stick-breaking procedure (Sethuraman, 1994). Starting with a unit-length stick, in each step a portion of the stick is broken off according to v_l and assigned to w_k. The independent random variables v_l have the distribution *Beta(1, α)*. Higher values of α will yield less concentrated distributions, allowing the weights to decay more gradually. As *α* decreases, less of the unit-length stick will be left for subsequent values, yielding a smaller number of clusters. The concentrated

parameter α can be regarded as the learner's belief about whether there are many or few types of wh-questions in a given language.

Each morphosyntactic property of the i^{th} question utterance is represented as a binary value $y_{ij} \in \{0, 1\}$. In the data sets, there are two morphosyntactic properties (wh-fronting and subject-auxiliary inversion), hence $j \in \{1, 2\}$. Each question type k assigns a probability $p_{kj} \in [0,1]$ that the j^{th} property will be present (=1) in a question of that type. The prior probability distribution over p_{kj} is a *Beta(1,1)* distribution, which assigns an equal prior probability to all values in $[0, 1]$.

Similarly, each prosodic property of the i^{th} question utterance is a continuous variable y_{il}. There are two prosodic properties in our data (duration and F0 contour on the wh-word), therefore $l \in \{1, 2\}$. Each question type k places a $Normal(\mu_{kl}), \sigma_{kl}$ distribution on the l^{th} prosodic property. The prior on the mean μ_{kl} is a *Normal(100, 50)* distribution for duration (which is necessarily positive) and a *Normal(0, 50)* distribution for F0 contour (which can be rising or falling). The prior distribution on σ_{kl} was a broad log-normal distribution, allowing for substantial variation within each question type.

The model is implemented in the probabilistic programming language Stan (Carpenter et al., 2017) and assessed for its ability to infer accurate English wh-question types and to correctly categorize question utterances drawn from child-directed corpora. The probability that the i^{th} question utterance, represented as two binary morpho-syntactic variables and two continuous prosodic variables, belongs to question type k is given by Bayes' Rule:

$$p(k|\mathbf{y}_i) = \frac{p(\mathbf{y}_i|k)\ w_k}{\sum_{k=1}^{K} p(\mathbf{y}_i|k')\ w_{k'}}$$

where each $p(y_i \mid k)$ is a product of two Bernoulli probabilities and two Normal densities.

Regularizing behavior

I attempt to model the regularizing pattern by manipulating two factors: a filtering rate and a parsimony bias α value.

The idea behind the filtering rate is that learners cannot utilize everything presented to them in the input when cognitive resources are limited. This ties back to the distinction between *input* vs *intake* introduced in Chapter 4. Following Gass, 1997, Gagliardi & Lidz, 2014, and Omaki & Lidz, 2015, *input* refers to the data available in the environment, while *intake* refers to the data from the input that learners actually utilize to make inferences about the target grammar. Early on, with limited cognitive resources, children's intake is smaller, in other words, the filtering rate to get to the intake from the input is higher. A weighted random sampling is used for the intake instead of pure random sampling, in which utterances that are consistent and frequent are more likely to be selected. The distribution of the three question types (84% fronted questions, 9% PQs, and 7% EQs) is kept constant in each sampled intake group, regardless of the intake size. To reflect a developmental trajectory, the filtering rate is slowly reduced over time until the intake matches the input. The values of data (N) and parsimonious bias α are reported in Appendix C.

The parsimony bias α value represents the learner's initial bias about the number of wh-question types in their language. A larger value of α would allow the learner to be more flexible in learning more categories of wh-questions. I hypothesize that early on, learners would have a stronger preference to learn as few variants as possible to reduce cognitive burden. Such bias is weakened over time and can eventually be overridden after sufficient exposure to the variable pattern. There are three conditions:

1. Data-Alpha: both the amount of intake data and α value were manipulated to increase over time, in other words, the filtering rate and the parsimony bias decreased over time.
2. Data-only: the amount of intake data increases over time, but the parsimony bias stays constant at 1 - the highest value that was tested in this project.
3. Alpha-only: only the α value increases over time. The amount of intake data stays constant at 2000, which matches the full input dataset.

Performance

Inference proceeded by Markov-Chain Monte Carlo (MCMC) sampling, as implemented in Stan, for 5000 iterations with the initial 2500 samples discarded as burn-in. Trace plots indicated that all parameters settled on stable values within the burn-in period, therefore without loss of detail, I present only average values over the remaining 2500 samples. The model accurately classified 97.7% of the simulated question utterances on which it was trained, and 86.0% of the natural child-directed speech test utterances. The main confusion in the test utterances was the misclassification of echo questions as probe questions, which is also a mistake that children made in the comprehension experiment reported in Chapter 4 (see also Nguyen & Legendre, 2022).

The sampling run shown in Table 19b converged on three clusters, ordered in descending probability that closely approximated the actual wh-question types in the training data in a (the other clusters inferred by the model had a total probability of 0.05 and are ignored here as noise).

Table 19. English questions based on natural child-directed speech (a) and inferred by the model (b)

a)

Type	%	WhFront	Inversion	WhDur (sd)	WhΔF0 (sd)
Fronted	.84	1	1	150 (49)	-6 (36)
Probe	.09	0	0	208 (85)	-21 (61)
Echo	.07	0	0	254 (60)	108 (64)

b)

Cluster	w	pWhFront	pInversion	μWhDur (σ)	μWh ΔF0 (σ)
1 ≈ Fronted	.827	1.0	1.0	119 (48)	4 (35)
2 ≈ Probe	.097	0.03	0.03	203 (78)	-8 (68)
3 ≈ Echo	.076	0.01	0.01	270 (49)	120 (58)

Figure 22 reports the changes in learning over time. In general, by slowly increasing the intake data and α to reflect developmental changes (Data:Alpha condition), the model displayed a regularizing pattern: initially the model only learned one cluster that showed the characteristics of fronted questions, then eventually expanded to two (fronted and in-situ). In the last two runs, with almost the full dataset, it was able to separate the two in-situ question types and learned all three clusters (fronted, in-situ echo, and in-situ information-seeking). The learned probability of fronted questions was also initially boosted to almost 100% before stabilizing around the intake rate. The same pattern was observed in the two smaller in-situ clusters: when the model first learned to separate the two in-situ questions into clusters 2 and 3, initially the value of the more dominant PQs was boosted to ~0.12 before getting closer to the intake value at 0.097. This confirms a

frequency-boosting pattern reported in previous studies (e.g., Hudson Kam & Newport, 2009; Schwab et al., 2018). In the Data-only condition, the same trend emerged. However, compared to the Data:Alpha condition, fewer data points were required for cluster 1 to reach its target value and for all three clusters to be learned. Finally, when only manipulating α in the Alpha-only condition, no effect was observed: all three clusters were learned at the same rate.

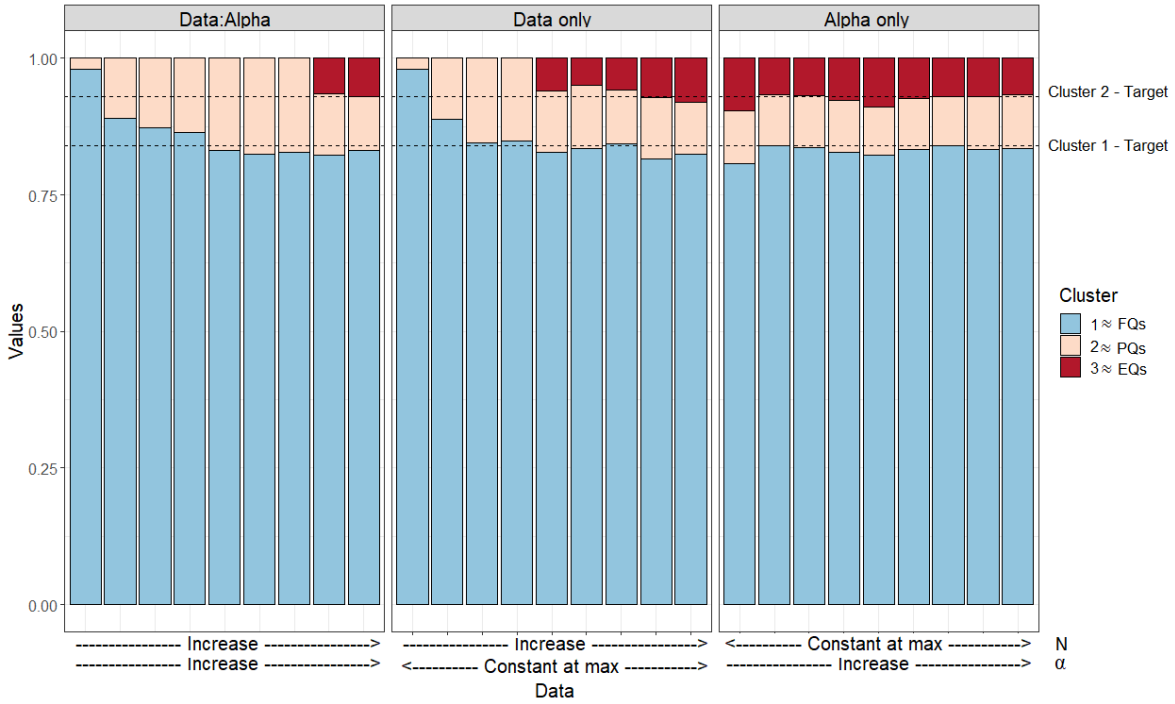

Figure 22. Learning over time in the three conditions. The dashed line below indicates the target value of Cluster 1 (which corresponds to FQs) and the dashed line above indicates the target of Cluster 2 (which corresponds to PQs).

Note that the values of α as well as the amount of training data in this simulation were only used to represent abstract developmental changes and were not meant to be interpreted as precise values. The result in the Alpha-only condition certainly does not mean that a learner just needs to hear 2000 data points to learn different wh-question types. Instead, the model is showing that different α values would require different (minimum) amounts of training data for the model to achieve learning. While a parsimony bias alone

may not be the primary motivation for regularizing, its interaction with intake quantity can capture the regularizing pattern in children.

Discussion

Overall, the Bayesian model described here successfully captures the regularizing patterns often found when there is variation in the input, in line with previous ALL studies (e.g., Hudson Kam & Newport, 2005, 2009; Schwab et al., 2018). To achieve this, two assumptions were made: (a) children make inferences about the target grammar based on a subset of the input, i.e., the intake (Goldowsky & Newport, 1993; Omaki & Lidz, 2015) and (b) children have a parsimony bias for learning initially a smaller number of variant types, though such bias is weakened over time as memory capacity and other cognitive abilities develop (Perfors, 2012). The interaction between the amount of intake data (N) and the learner's bias (α) shows the importance of intake quantity in regularizing. This is in line with Hendricks et al. (2018)'s empirical work on regularizing in Fering, in which bilingual children who are exposed to more Fering are shown to display an adult-like pattern that preserves the inconsistency in their production, while children who are exposed to less Fering end up regularizing the inconsistent feature (see Section 1.2 for details). In other words, the amount of intake can determine whether the learner would regularize their production or not. While the learner's parsimony bias alone may not induce a regularizing pattern, it can play a role during the learning process, as shown in the differences between the Data-only condition versus the Data-Alpha condition. The learner's bias can be used to account for cross-learner and cross-linguistic variation. A learner with a stronger parsimony bias may take longer to learn multiple variants of the same grammatical item compared to a learner with a weaker parsimony bias.

In natural and ALL experiments discussed in previous chapters, we have seen that regularization happens more with production (or cognitively-taxing tasks) than with comprehension (or cognitively even less taxing tasks such as two-alternative-forced-choice). However, the distinction between comprehension and production is not well-captured in this model. One might propose that the classification task is similar to comprehension and the sampling task to production, but that is perhaps not a good characterization as these two computational tasks are not very different. Instead, I provisionally assume that both comprehension and production utilize the same underlying learning mechanism. However, production may require more data to master and involve more cognitive load in planning (Hendriks and Koster, 2010; Humphreys, 2012). The asymmetry between the production-comprehension processes could simply be because production requires more intake and learning time than comprehension. Production thus 'lags' behind – when comprehension is in the three-cluster stage, production may still be in the one- or two-cluster stage, where regularization still happens.

However successful the non-parametric model is in capturing regularization behavior, it still faces several limitations. For instance, the morphosyntactic properties considered were manually coded and the prosodic properties were manually extracted. This makes the idealization that, first, these properties are always available to the learner in every instance of wh-question, and, second, they are always perceived correctly. In reality, both of these assumptions are likely to be violated, at least part of the time. The availability and perception of these properties depend on many factors, including but not limited to the environment (e.g., noisy versus not noisy) and/or the attention of the listener. Some properties, such as duration, may be more prone to errors than others (Gussenhoven and

Zhou, 2013). Moreover, while the properties used in this model are representative, they are not comprehensive. The model only looked at the prosody of the wh-word itself, but the intonation of the question at a sentence level can also bear important information (Déprez, Syrett, & Kawahara, 2013). Most importantly, no pragmatic or discourse properties were considered and all three variants of wh-questions were treated as unconditioned variants, i.e., they can be used interchangeably. As we saw in Chapter 2, this is untrue for English fronted and probe questions, given their respective discourse constraints and discourse-markedness properties.

The decision to overlook these differences in this initial model partially comes from the limitation in quantifying and implementing pragmatic properties in a computational model. The model was solely concerned with inducing regularization using an input-filter mechanism, as well as exploring the role of certain factors like the quantity of intake data and learner's parsimony bias. The question the model strives to answer is simple: what is a possible process to make the output more regularized and contain less variation than the input? Thus, the model focuses on variation itself rather than the characteristics of variants. Given that our goal was to explore a possible mechanism of regularization in general, the model is still valid in demonstrating how the input filter hypothesis can drive regularization.

5.4. Chapter 5 summary

Chapter 5 tested the hypothesis that there is a relationship between cognitive ability and regularization. Using an ALL paradigm, Experiment 4b shows that there is a negative correlation between learners' regularization rate and their composite working memory

score from 3 tasks: backward digit spans, N-back, and sentence repetition. In particular, participants who regularize more tend to have lower working memory scores. This finding is in line with previous studies showing that younger children (who have lower working memory capacity) regularize more than older children, and children, in general, regularize more than adults (who have higher working memory capacity). A Bayesian computational model followed which looks into a possible mechanism of regularization. The model captures the regularizing pattern in wh-question acquisition using an input filter inspired by the Less is More hypothesis. The main idea of an input filter is that early on, when cognitive capacity is limited, children would not be able to learn from everything presented to them in the input. Instead, children only make use of a smaller subset of the input, which is called the *intake.* The model simulation shows that when the intake is smaller and there is a stronger bias for parsimony, the output is more regularized. This reflects developmental changes and suggests that younger children (with smaller intake) would regularize more than older children. Results from the ALL experiment and the Bayesian computational model both (directly and indirectly) support the hypothesis that regularization happens to reduce cognitive load during the learning process.

Chapter 6: Conclusion

This dissertation has examined the test case acquisition of multiple wh-variants and made contributions to the field of language acquisition by providing novel theoretical insights into experimentally-collected data on a) the acquisition of probe questions in English, b) factors that condition the acquisition of variation, and c) regularization behaviors in the acquisition of variation. In this last chapter, I provide a summary of the insights yielded by the research (Section 6.1. to Section 6.3.) before discussing the future picture (Section 6.5.)

6.1. The acquisition of probe questions

Chapter 2 contributes to the general understanding of wh-questions by presenting new data from child-directed speech on probe questions – an understudied type of wh-in-situ question. Probe questions are information-seeking questions used under certain discourse contexts, such as when the addresser has some information about the answer or is interested in the addressee's ability to elicit an answer more than the answer itself. In English, probe questions share the same function and intonation pattern with fronted wh-questions, but are more discourse-marked: they typically cannot be used out of the blue and need to satisfy Common Ground requirements.

In child-directed speech, parents frequently alternate between probe questions and fronted questions as if they can be used interchangeably – which is documented on the basis of an extensive corpus study. Children being exposed to such variable input are likely to infer that probe questions are a variant of fronted wh-questions. Following Chapter 2,

Chapter 3 further investigates the acquisition of probe questions experimentally. While corpus analyses in Chapter 2 revealed that 3-to-5-year-old children produce rather few instances of probe questions spontaneously, follow-up behavioral experiments confirmed that children strongly prefer fronted wh-questions even in contexts that satisfy Common Ground requirements, despite understanding probe questions. These results contribute to the general understanding of children's acquisition of English wh-questions, which has previously been limited to the study of fronted questions alone. In particular, they challenge previous studies that assume English-speaking children are never exposed to non-echo in-situ wh-questions and show instead that children's input is more complicated than previously reported. These complications, in turn, raise important theoretical questions about learning and the relationship between the input to children and their own production.

The results show that children do not imitate their input. When being exposed to multiple variants, they do not try to match the distribution found in the input but instead regularize to one variant. By comparing the properties of probe questions with the properties of other wh-question variants in English (e.g., fronted questions and echo questions), we can further our understanding of factors conditioning the acquisition of variation, as detailed in the next section.

6.2. Factors that condition the acquisition of variation

The behavioral experiments in Chapter 3 show that children do not produce probe questions and fronted questions in a way that matches the input distribution, but instead regularize to one variant. The regularization direction is not random (e.g., some children

regularize to probe questions and some to fronted questions), but almost all children consistently prefer fronted questions. Cross-linguistically, such preference for fronted questions is reported in some studies in French and Brazilian Portuguese (e.g., Prévost et al, 2017; Vieira & Grolla, 2020), two languages that also allow for both in-situ and fronted wh-variants. This suggests that fronted questions have one or more characteristics that are more favorable to child learners.

In general, the wh-in-situ variant differs from the fronted wh-variant in the following aspects:

(1) Frequency: In English, fronted questions are dominantly more frequent than the wh-in-situ variant, with a distribution in child-directed speech at approximately 80% to 20%. In French and Brazilian Portuguese, however, the frequency distribution of the two variants is more inconclusive, as different studies report different numbers. Still, there is at least one study in each language claiming fronted wh-questions to be the more frequent variant (e.g., Beck & Gotowski, 2015; Sikansi, 1999).

(2) Discourse markedness: Across English, French, and Brazilian Portuguese, fronted wh-questions are the discourse-unmarked variant while wh-in-situ are discourse-marked. The discourse-markedness of wh-in-situ questions is expressed through their more restricted set of valid contexts. While fronted wh-questions can be used in almost any context, wh-in-situ can only appear in contexts that satisfy certain discourse requirements.

(3) Structural economy: Across all languages, (object and adjunct) fronted wh-questions are structurally more complex than their wh-in-situ counterparts, as the former involve at least one additional syntactic operation, i.e., the fronting of the wh-word.

Let us look at each proposal in more detail. First, it is not unreasonable to attribute the preference for fronted wh-questions to the higher frequency of this variant. A frequency-based theory is one of the dominant theories in language acquisition (Tomasello, 2005). In many artificial language learning studies that investigate regularization behavior during the acquisition of variation, learners have been reported to regularize to the more frequent variant (Hudson Kam & Newport, 2009; Schwab et al., 2018). However, frequency should not be the only factor motivating this regularization preference, as some cases have been documented, that cannot be explained by frequency. For instance, Lessa-de-Oliveira (2003) reports a case study in which the child showed a preference for fronted questions even when her input contained more wh-in-situ questions. To resolve this problem, a discourse-based account can work in tandem with a frequency-based account, since discourse-markedness and frequency tend to go together: stricter discourse restrictions entail fewer usable contexts and hence lower frequency. It is also not unreasonable to predict that learners prefer the less marked variant when learning multiple variants, as this is the most parsimonious option to learn, minimizing the chance of violating the restrictions, and maximizing efficient communication. Finally, structural economy, an alternative dominant theory in the acquisition of syntax (Jakubowicz, 2011), is not the right explanation in this case. Structural economy accounts predict that learners prefer the more economical (or syntactically simpler) wh-in-situ variant, which is not supported by the experimental results. Note that this does not mean economy plays no role in language acquisition at all. The current results only show that structural economy is not a factor conditioning the acquisition of multiple wh-variant, or, on a more cautious note, that there is a trade-off between economy and frequency/discourse markedness.

In brief, based on the characteristics of the variants, two factors have been identified as potentially playing a role in the acquisition of wh-variants: frequency and discourse markedness. Between the two, frequency has been studied extensively and is established as an important factor in language acquisition in general and acquisition of variation specifically. On the other hand, there are few studies on discourse markedness in acquisition, especially addressing its relevance to regularization -- to the best of my knowledge, there are none. Filling this gap, Chapter 4 is devoted to exploring whether (and how) discourse markedness conditions the regularization behavior found in the acquisition of variation.

Chapter 4 presents two ALL experiments. The ALL design was chosen so that the relative effect of discourse markedness and of frequency can be controlled and separated. In the first ALL experiment, participants were equally exposed to a structurally simple but discourse-marked and a structurally complex but discourse-unmarked variant before going through comprehension and production tasks. Results show that the majority of participants showed a preference for the unmarked variant in the production task. In the second ALL experiment, participants were randomly put into either a Strict condition or a Loose condition. In the Strict condition, they were exposed to a strictly constrained (i.e., highly discourse-marked) variant and a discourse-unmarked variant, while in the Loose condition, they were exposed to a loosely constrained (i.e., slightly discourse-marked) and a discourse-unmarked variant. An overall preference for the unmarked variant was found, similar to the results of the first ALL experiment. Such preference was significantly stronger in the Strict condition compared to the Loose condition. Overall, the two ALL experiments show that when frequency is controlled, participants are trading off structural

economy for discourse unmarkedness. The results highlight the role of discourse (un)markedness in the acquisition of wh-variants, and likely in the acquisition of multiple variants in general.

Besides identifying frequency and discourse (un)markedness as two factors that can condition the acquisition of variation, our results also show that regularization is not a strictly domain-general behavior. While regularization itself is not language-specific (as it can happen in visual learning as well (Derks & Paclisanu, 1967), the regularization behavior in (language) variation learning is still conditioned by linguistic factor(s) such as discourse (un)markedness. As we expand to other case studies of multiple variants in language besides wh-variation, we will most likely find more domain-specific factors that can condition the course of variation acquisition than the factors listed here.

6.3. Regularization behaviors in the acquisition of variation

The second aspect of the acquisition of variation that this dissertation focuses on is regularization. From both the experiments on multiple variants in natural language (Chapter 3) and on artificial language (Chapter 4), we see that learners tend to regularize to one variant in their production rather than faithfully reproducing their input. This finding is in line with a number of prior studies in which participants were exposed to multiple variants (e.g., in natural language: Pozzan & Valian, 2016; in ALL paradigms: Hudson Kam & Newport, 2005, 2009).

Regularization is not specific to language learning. Derks & Paclisanu, (1967) and Yurovsky et al. (2013) have reported studies in which young children regularize to the

simpler variant during picture learning. Thus, regularization seems to be a general response when learners are faced with variation in their input. But why does regularization happen? One hypothesis is that regularization serves as a means to reduce the cognitive burden associated with learning multiple variants at once. Instead of mastering all the variants, (which includes mastering the subtle differences between them), early on, learners simplify the learning process and minimize their chance of making an error by producing the dominant variant, which is the more frequent or more neutral (e.g., unmarked) variant. Support for this hypothesis previously comes from experimental results (e.g., Austin, 2010; Hudson Kam & Newport, 2009; Schwab et al., 2018) showing that the rate of regularization increases as learners have fewer cognitive resources, either due to age (e.g., children regularize more than adults) or due to the nature of the task they face (e.g., tasks that are more cognitively-taxing induce more regularization). This dissertation provides additional evidence for the hypothesis in three ways:

(1) It replicates the findings that more cognitively-taxing tasks induce more regularization: Results from two ALL experiments in Chapter 4 both show that the regularization rate is significantly higher in the production task than in the 2AFC task. This is likely because a production task requires more memory retrieval and planning than the 2AFC task.

(2) It demonstrates a relationship between working memory/WM ability and regularization rate: Using the ALL paradigm and a set of three WM tasks, Experiment 4b in Chapter 5 finds that participants with a lower composite WM score tend to have a higher regularization rate. To the best of my knowledge, this is the first study that directly tests

for a correlation between WM ability and regularization instead of manipulating WM through interferences.

(3) It provides a computational simulation of regularizing behavior: The non-parametric Bayesian model reported in Chapter 5 successfully captures the regularizing behavior in multiple wh-variants acquisition by making two assumptions about WM and cognitive resources. First, the model assumes that young learners start out with limited WM capacity and can only make use of a subset of their input (or intake), with the size of the intake growing over time to reflect an increase in WM capacity. Second, the model assumes that early on, learners have a stronger parsimony bias (in favor of a minimal number of variants) in order to reduce their cognitive burdens, and this bias weakens over time. When there is no parsimony bias or intake limit, the regularization behavior in the model disappears, suggesting that there is a relationship between regularization and limitations in WM and/or cognitive resources.

6.4. Novel contributions

This dissertation contributes new theoretical insights to both the acquisition of wh-questions and the acquisition of variation. With regard to wh-questions, this is the first study to look into the acquisition of English information-seeking wh-in-situ questions; it reveals that children's wh-question input is more diverse than previously claimed. It also demonstrates the role of discourse markedness, an understudied factor, in language acquisition, and it shows that regularization behavior is sensitive to both domain-general (e.g., frequency) and language-specific (discourse markedness) factors.

Methodologically, the ALL experiments in Chapter 4 focus on syntactic variation, extending to core syntax the current literature on regularization which has largely focused on morphological variation (e.g., Hudson Kam & Newport, 2005, 2009; Austin, 2010; Wonnacott, 2011; Schwab et al., 2018). Moreover, instead of manipulating working memory through interference tasks (e.g., Perfors & Burns, 2010; Perfors, 2012; Hudson Kam, 2019), our work directly tests for a correlational relationship between working memory composite score and regularization rate. This helps avoid potential confounding factors from the interference tasks, such as divided attention, the ability to manage interference, and/or the ability to suppress irrelevant information.

6.5. Future directions

The studies in this dissertation are initial steps in the study of a single test case of regularization in syntax from an interdisciplinary perspective, and there are a few directions I hope to pursue in the future. First, the current ALL studies have been run with adult participants only. This was done to ensure the learnability of the task, as this was my first time adopting an ALL design. Given that the results suggest that the ALL design successfully induced regularization behavior in adult learners, the next step is to run similar studies with child participants. In order to do so, a few adjustments need to be made. For example, all the text components in the study need to be converted to audio. Additionally, it is likely to take longer for children to learn the artificial language, so the study needs to be broken down into shorter sessions that run over the course of several days. These changes would also allow us to see if and how different learning modalities (i.e., over text versus over audio) may affect the learning outcome. Second, as briefly outlined in section

5.3.2., the non-parametric Bayesian model still faces a few limitations. The model could be extended to account for more linguistic properties besides just word order and prosody. With more resources and computational power, a future model could take in raw audio wh-question utterances as input in lieu of manually coded data.

Currently, this work has only focused on wh-variation as the main case study for the acquisition of variation. I hope to expand the work to other cases of syntactic variation in the future, such as polar questions or English dative constructions.

Finally, Experiment 4b has yielded an interesting result on the relationship between the N-back task and the rule learning process, showing that participants who violate the discourse constraints more also tend to have lower performance in the N-back task. This result wasn't explored further here as it was outside the scope of this dissertation; however, future work can explore this topic to gain more insights into either the nature of the N-back task or the rule recognition process. The first step would be to see if the result is replicable in an experiment that is specifically designed for rule learning. There also are several variations on the experiments that could be run. For example, one could see if any variation of the N-back (e.g., 1-back, 2-back, 3-back etc.) and any modalities and domains of the stimuli (e.g., text versus audio, linguistic versus non-linguistic) yields the same result as Experiment 4b.

In short, regularization in learning of syntactic variation is a rich area of research that promises to yield a lot more insights in this most fundamental of human abilities – learning natural language.

References

Abend, Omri, Tom Kwiatkowski, Nathaniel Smith, Sharon Goldwater, & Mark Steedman. 2017. Bootstrapping language acquisition. *Cognition*,164:116–143.

Adger, David. 2003. *Core syntax*. Oxford: Oxford University Press.

Adli, Aria. 2004. Y a-t-il des morphèmes intonatifs impliqués dans la syntaxe interrogative du Français? Le cas du qu-in-situ. In T. Meisenburg & M. Selig (Eds.), *Nouveaux départs en phonologie: les conceptions sub- et suprasegmentales.* Tübingen: Narr. 199-215.

Ambridge, B., Rowland, C. F., Theakston, A. L., & Tomasello, M. 2006. Comparing different accounts of inversion errors in children's non-subject wh-questions:'What experimental data can tell us?'. *Journal of Child Language*, *33*(3), 519-557.

Artstein, Ron. 2012. A focus semantics for echo questions. In ́Agnes Bende-Farkas and Arndt Riester, editors, *Workshop on Information Structure in Context*, 98–107. IMS, University of Stuttgart.

Austin, Alison. 2010. *When children learn more than what they are taught: regularization in child and adult learners.* University of Rochester doctoral thesis.

Banfield, Ann. 1982. *Unspeakable sentences: Narration and representation in the language of fiction.* Boston, MA: Routledge & Kegan Paul.

Barbu, Stéphanie, Aurélie Nardy, Jean-Pierre Chevrot, & Jacques Juhel. 2013. Language evaluation and use during early childhood: Adhesion to social norms or integration of environmental regularities?. *Linguistics,* 51(2), 381-411.

Bates, Douglas, Martin Maechler, Ben Bolker, & Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1-48.

Baunaz, Lena, & Cédric Patin. 2009. Prosody refers to semantic factors: Evidence from French wh-words. *Actes d'Interface Discours and Prosodie*, 93-107.

Becker, Misha, & Megan Gotowski. 2015. Explaining children's wh-in situ questions: Against economy. *Proceedings of BUCLD 39*, Cascadilla Press. 127-137.

Bellugi, Ursula. 1965. The development of interrogative structures in children's speech. In K. Riegel (Ed.). *The Development of Language Function*. Northumberland: Ann Arbor. 103-137.

Beyssade, Claire. 2007. The prosody of French interrogatives. *Nouveaux cahiers de linguistique française*, 28, 163-175.

Biberauer, Theresa, & Marc Richards. 2006. True optionality: When the grammar doesn't mind. *Minimalist essays*, 35-67.

Biezma, Maria. 2020. Non-informative assertions: The case of non-optional wh-in-situ. *Semantics and Pragmatics*, *13*, 18.

Blakemore, Diane. 1994. Echo Questions: A Pragmatic Account, *Lingua*, 94, 197–211.

Bobaljik, Jonathan, & Susi Wurmbrand. 2015. Questions with declarative syntax tell us what about selection. *MIT working papers in linguistics* 17, 13-32.

Boersma, Paul, & David Weenink. 2019. *Praat: doing phonetics by computer* [Computer program]. Version 6.0.52, retrieved 2 May 2019 from http://www.praat.org/.

Bolinger, Dwight. 1974. Meaning and form. *Transactions of the New York Academy of Sciences*, *36*(2 Series II), 218-233.

Boucher, Paul. 2010. Wh-questions in French and English. Mapping syntax to information structure. In C. Breul and E. Göbbel (eds.), *Comparative and contrastive studies of information structure*. Amsterdam: John Benjamins. 101-137.

Braunwald, Susan. 1985. The development of connectives. *The Journal of Pragmatics*, 9(4), 513-525.

Brown, Roger. 1973. A first language: The early stages. Cambridge, MA: Harvard University Press.

Bruening, Benjamin, & Thuan Tran. Wh-questions in Vietnamese. 2006. *Journal of East* Asian *Linguistics* 15(4), 319-341.

Chang, Lisa. 1997. *Wh-in-situ Phenomena in French*. University of British Columbia, Master thesis.

Cheng, Lisa L. S. 1991. *On the typology of wh-questions*. Massachusetts Institute of Technology dissertation.

Cheng, Lisa L. S. 2009. Wh-in-situ, from the 1980s to Now. *Language and Linguistics* Compass, *3*(3), 767-791.

Cheng, Lisa Lai-Shen and Johan Rooryck. 2000. Licensing wh-in-situ. *Syntax* 3: 1-19.

Chevrot, Jean-Pierre and Foulkes, Paul. 2013. Introduction: Language acquisition and sociolinguistic variation. *Linguistics*, 51(2), 251-254.

Chin, Simone, & Alan Kersten. 2010. The Application of the Less is More Hypothesis in Foreign Language Learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 32

Choi, Hye-Won. 1996. *Scrambling: Optimality-theoretic interaction between syntax and* discourse. Stanford University dissertation.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, Massachusetts: The MIT Press.

Chrysikou, Evangelia, Jared Novick, John Trueswell, & Sharon Thompson-Schill. 2011. The other side of cognitive control: can a lack of cognitive control benefit language and cognition?. *Topics in Cognitive Science*, *3*(2), 253-256.

Clark, Eve & Brian MacWhinney. 1987. The principle of contrast: A constraint on language acquisition. *Mechanisms of language acquisition*, 1-33.

Clark, Herbert, & Keith Brown. 2006. Context and common ground. *Concise* Encyclopedia *of Philosophy of Language and Linguistics (2006)*, 85-87.

Cochran, Barbara, Janet McDonald, & Susan Parault. 1999. Too smart for their own good: The disadvantage of a superior processing capacity for adult language learners. *Journal of Memory and Language*, *41*(1), 30-58.

Comyn, Marie. 2013. *Wh-in situ is accompanied by which formal features*? Ghent University, Master thesis.

Coveney, Aidan. 2020. L'interrogation directe. In *Encyclopédie grammaticale du français*. http://www.encyclogram.fr/notx/002/002_Notice.php (Last access: 05.06.2021)

Crain, Stephen, & Mineharu Nakayama. 1987. Structure dependence in grammar formation. *Language*, 63(3), 522-543.

Crisma, Paola. 1992. On the Acquisition of Wh-Questions in French. *Geneva Generative Papers*, 1, 115-122.

Culbertson, Jennifer. 2010. Convergent evidence for categorial change in French: From subject clitic to agreement marker. *Language*, *86*(1), 85-132.

Culbertson Jennifer, Paul Smolensky, & Geraldine Legendre. 2012. Learning biases predict a word order universal. *Cognition, 122*(3), 306-329.

Cuza, Alejandro, Lauren Miller, Rocio Tattam, & Mariluz Vergara. 2019. Structure complexity effects in child heritage Spanish: The case of the Spanish personal a. *International Journal of Bilingualism*, *23*(6), 1333-1357.

de Villiers, Jill, & Pyers, Jennie. 2002. Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive development*, *17*(1), 1037-1060.

de Villiers, Jill. 1995. Questioning minds and answering machines. In D. MacLaughlin, & S. McEwen (Eds.), *Boston University Conference on Language Development (BUCLD) 19*. Somerville, MA: Cascadilla Press. 20-36.

Demetras, Martha. 1989. Working parents' conversational responses to their two-year-old sons. University of Arizona doctoral thesis.

Demuth, Katherine, Jennifer Culbertson, & Jennifer Alter. 2006. Word-minimality, Epenthesis, and Coda Licensing in the Acquisition of English. *Language & Speech*, 49, 137-174.

Déprez, Viviane, Kristen Syrett, & Shigeto Kawahara. 2013. The interaction of syntax, prosody, and discourse in licensing French wh-in-situ questions. *Lingua*, *124*, 4-19.

Derks, Peter, & Marianne Paclisanu. 1967. Simple strategies in binary prediction by children and adults. *Journal of Experimental Psychology,* 73(2), 278.

DeRoma, Cynthia. 2011. *Divide Et Impera: Separating Operators from Their Variables*. University of Connecticut doctoral thesis.

Dickinson, David, & Patton Tabors. 2001. *Beginning literacy with language: Young children learning at home and school*. Baltimore: Paul Brookes Publishing.

Donka Farkas, & Floris Roelofsen. 2017. Division of Labor in the Interpretation of Declaratives and Interrogatives. *Journal of Semantics* 34(2): 237–289.

Dörnyei, Zoltán. 2014. *The psychology of the language learner: Individual differences in* second *language acquisition*. Routledge.

Durrleman, Stephanie, Theodoros Marinis, & Julie Franck. 2016. Syntactic complexity in the comprehension of wh-questions and relative clauses in typical language development and autism. *Applied Psycholinguistics*, *37*(6), 1501-1527.

Ehrman, Madeline, Betty Leaver, & Rebecca Oxford, 2003. A brief overview of individual differences in second language learning. *System*, *31*(3), 313-330.

Elman, Jeffrey. 1993. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1), 71-99.

Ettlinger, Marc, Kara Morgan-Short, Mandy Faretta-Stutenberg, & Patrick Wong. 2016. The relationship between artificial and second language learning. *Cognitive science*, *40*(4), 822-847.

Faure, Richard, & Katerina Palasis. 2021. Exclusivity! Wh-fronting is not optional wh-movement in Colloquial French. *Natural Language & Linguistic Theory*, *39*(1), 57-95.

Fedzechkina, Maryia, Elissa Newport, & T. Florian Jaeger. 2017. Balancing effort and information transmission during language acquisition: Evidence from word order and case marking. *Cognitive science*, *41*(2), 416-446.

Ferdinand, Vanessa, Simon Kirby, & Kenny Smith. 2019. The cognitive roots of regularization in language. *Cognition*, 184, 53-68.

Frost, Adam, Simar Moussaoui, Jagjot Kaur, Samreen Aziz, Keisuke Fukuda, & Matthias Niemeier. 2021. Is the n-back task a measure of unstructured working memory capacity? Towards understanding its connection to other working memory tasks. *Acta Psychologica*, 219: 103398.

Frost, Ram, Blair Armstrong, & Morten Christiansen. 2019. Statistical learning research: A critical review and possible new directions. *Psychological Bulletin*, 145(12), 1128.

Gagliardi, Annie & Jeffrey Lidz. 2014. Statistical insensitivity in the acquisition of Tsez noun classes. *Language*, 90(1), 58–89.

Gass, Susan. 1997. *Input, Interaction, and the Second Language learner*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gershman, Samuel and David Blei. 2012. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Goldowsky, Boris, & Elissa L. Newport. 1993. Modeling the Effects of Processing Limitations on the Acquisition of Morphology: The Less is More Hypothesis. *Proceedings of the 24th Annual Child Language Research Forum,* 124-138.

Goldwater, Sharon. (2007). *Nonparametric Bayesian models of lexical acquisition*. Ph.D. thesis, Brown University.

Gopnik, Alison, Thomas Griffiths, & Christopher Lucas. 2015. When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*(2), 87-92.

Gotowski, Megan. 2017. Wh- In Situ Production in Child French. *Linguistica Atlantica*, *36*(2), 99-109.

Grice, Paul. 1975. Logic and conversation. In Peter Cole & Jerry Morgan. (eds.), *Syntax and Semantics*, Vol. 3, Speech Acts. New York: Academic Press. 41-58.

Grolla, Elaine. 2009. Speculations about the acquisition of wh-questions in Brazilian Portuguese. In Acrisio Pires and Jason Rothman (eds), *Minimalist inquiries into child and adult language acquisition: Case studies across Portuguese*. The Hague: Mouton de Gruyter, 85-104.

Gualmini, Andrea, Sarah Hulsey, Valentine Hacquard, & Danny Fox. 2008. The question-answer requirement for scope assignment. *Natural Language Semantics, 16*(3), 205–237.

Gunlogson, Christine. 2002. Declarative questions. *Semantics and linguistic theory*, 12, 124-143.

Gussenhoven, Carlos, and Wencui Zhou. 2013. Revisiting pitch slope and height effects on perceived duration. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, 1365-1369.

Hamann, Cornelia. 2006. Speculations About Early Syntax: The Production of Wh-questions by Normally Developing French Children and French Children with SLI. *Catalan Journal of Linguistics, 5*(1), 143-189.

Hamblin, Charles. 1973, 'Questions in Montague English'. *Foundations of Language*, 10:41–53.

Hamlaoui, Fatima. 2011. On the role of phonology and discourse in Francilian French wh-questions. *Journal of Linguistics* 47(1), 129-162.

Hansen, Laura, Pedro Macizo, Jon Duñabeitia, David Saldaña, Manuel Carreiras, Luis Fuentes, and Teresa Bajo. 2016. Emergent bilingualism and working memory development in school aged children. *Language Learning* 66(2) 51-75.

Hanten, Gerri, & Randi Martin. 2000. Contributions of phonological and semantic short-term memory to sentence processing: Evidence from two cases of closed head injury in children. *Journal of Memory and Language* 43(2), 335-361.

Harness, Ashley, Lorri Jacot, Shauna Scherf, Adam White, & Jason Warnick. 2008. Sex differences in working memory. *Psychological reports*, *103*(1), 214-218.

Hendricks, Alison Eisel, Karen Miller, & Carrie Jackson. 2018. Regularizing unpredictable variation: Evidence from a natural language setting. *Language Learning and Development*, *14*(1), 42-60.

Hendriks, Petra, & Charlotte Koster. 2010. Production/comprehension asymmetries in language acquisition. *Lingua*, *120*(8), 1887-1897.

Hilbert, Sven, Tristan Nakagawa, Patricia Puci, Alexandra Zech, & Markus Bühner. 2015. The digit span backwards task: Verbal and visual cognitive strategies in working memory assessment. *European Journal of Psychological Assessment*, *31*(3), 174.

Hopp, Holger, Michael T. Putnam, & Nora Vosburg. 2019. Derivational complexity vs. transfer effects: Long-distance wh-movement in heritage and L2 grammars. *Linguistic Approaches to Bilingualism*, *9*(3), 341-375.

Huang, Cheng-Teh. 1982. *Logical relations in Chinese and the theory of grammar*. Massachusetts Institute of Technology dissertation.

Hudson Kam, Carla, & Elissa Newport. 2005. Regularizing Unpredictable Variation: The Roles of Adult and Child Learners in Language Formation and Change. *Language Learning and Development*, 1, 151–195.

Hudson Kam, Carla, & Elissa Newport. 2009. Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology,* 59(1), 30–66.

Hudson Kam, Carla. 2019. Reconsidering retrieval effects on adult regularization of inconsistent variation in language. *Language Learning and Development,* 15(4), 317-337.

Humphreys, Gina. 2012. *Linking sentence production and comprehension: The neural mechanisms underlying production and comprehension control processes.* University of York doctoral thesis.

Hupp, Julie, & Melissa Jungers. 2013. Beyond words: Comprehension and production of pragmatic prosody in adults and children. *Journal of Experimental Child Psychology*, *115*(3), 536–551.

Ito, Kiwako, Nobuyuki Jincho, Utako Minai, Naoto Yamane, & Reiko Mazuka. 2012. Intonation facilitates contrast resolution: evidence from Japanese adults and 6-year olds. *Journal of Memory and Language,* 66, 265–284.

Iwata, Seizi. 2003. Echo questions are interrogatives? Another version of a metarepresentational analysis. *Linguistics and Philosophy* 26(2), 185-254.

Jaeggi, Susanne, Martin Buschkuehl, Walter Perrig, & Beat Meier. 2010. The concurrent validity of the N-back task as a working memory measure. *Memory* 18(4), 394-412.

Jakubowicz, Celia, & Nelleke Strik. 2008. Scope-marking strategies in the acquisition of long distance wh-questions in French and Dutch. *Language and Speech*, 51(1–2), 101–132.

Jakubowicz, Celia. 2005. The Language Faculty: (Ab)normal Development and Interface Constraints. *Presentation at GALA 2005*, Siena.

Jakubowicz, Celia. 2011. Measuring derivational complexity: New evidence from typically developing and SLI learners of L1 French. *Lingua 121*(3), 339-351.

Johnson, Mark. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, 20–27.

Kane, Michael, Andrew Conway, Timothy Miura, & Gregory Colflesh. 2007. Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental psychology: learning, memory, and cognition*, 33(3): 615.

Klima, Edwards, & Ursula Bellugi. 1966. Syntactic regularities in the speech of children. In J. Lyons & R. Wales (eds.), *Psycholinguistic Papers*. Edinburgh : Edinburgh University Press.

Kuczaj, Stan. 1977. The acquisition of regular and irregular past tense forms. *Journal of* verbal *learning and verbal behavior*, *16*(5), 589-600.

Kuczaj, Stan. 1977. The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16, 589–600.

Lacoste, Véronique & Lisa Green. 2016. Sociolinguistic and formal approaches. *Linguistic Variation*, 16(1), 1-11.

Lee, Chia-ying, Timothy O'Donnell, & James Glass. (2015). Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403.

Legendre, Geraldine, Paul Hagstrom, Anne Vainikka, & Marina Todorova. 2002. Partial constraint ordering in child French syntax. *Language Acquisition*, *10*(3), 189-227.

Lessa-de-Oliveira, Adriana. 2003. Aquisição de constituintes-QU em dois dialetos do Português Brasileiro. *Torno da Língua (gem): Questões e Análises*, 85.

Lieven, Elena, & Michael Tomasello. 2008. Children's first language acquisition from a usage-based perspective. In P.Robinson and N.Ellis (eds), *Handbook of Cognitive Linguistics and Second Language Acquisition*, 168-196.

Lightfoot, David. 1999. *The Development of Language: Acquisition, Change and* Evolution. Malden, MA and Oxford: Blackwell.

Lukasik, Karolina, Minna Lehtonen, Anna Soveri, Otto Waris, Jussi Jylkkä, & Matti Laine. 2018. Bilingualism and working memory performance: Evidence from a large-scale online study. *PloS one* 13(11).

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Marinis, Theodoros, & Sharon Armon-Lotem. 2015. Sentence repetition. *Assessing* multilingual *children: Disentangling bilingualism from language impairment*: 95-124.

Mathieu, Eric. 2004. The mapping of form and interpretation: the case of optional wh-movement in French. *Lingua* 114: 1090-1132.

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, & Morgan Sonderegger. 2017. Montreal Forced Aligner [Computer program]. Version 0.9.0, retrieved from http://montrealcorpustools.github.io/Montreal-Forced-Aligner/.

McDonough, Laraine, Soonja Choi, & Jean Mandler. 2003. Understanding spatial relations: Flexible infants, lexical adults. *Cognitive psychology*, *46*(3), 229-259.

Mencarelli, Lucia, Francesco Neri, Davide Momi, Arianna Menardi, Simone Rossi, Alessandro Rossi, & Emiliano Santarnecchi. 2019. Stimuli, presentation modality, and

load-specific brain activity patterns during n-back task. *Human brain mapping*, *40*(13), 3810-3831.

Miller, Kimberly, C. Price, M. Okun, H. Montijo, & D. Bowers. 2009. Is the n-back task a valid neuropsychological measure for assessing working memory? *Archives of Clinical* Neuropsychology, *24*(7), 711-717.

Morgenstern, Aliyah, & Christophe Parisse. 2007. Codage et interprétation du langage spontané d'enfants de 1 à 3 ans. *Corpus n°6 "Interprétation, contextes, codage*", 55-78.

Mycock, Louise. 2013. Discourse functions of question words. Proceedings of the *LFG13* Conference. Stanford, CA: CSLI Publications.

Newport, Elissa. 1993. Modeling the effects of processing limitations on the acquisition of morphology: the less is more hypothesis. *Proceedings of the twenty-fourth annual child language research forum. Center for the Study of Language*, 124.

Nguyen, An, & Geraldine Legendre. 2022. The acquisition of wh-questions: Beyond structural economy and input frequency. *Language Acquisition*, *29*(1), 79-104.

Nishigauchi, Taisuke. 2006. Short answers as focus. *Theoretical and applied linguistics at Kobe Shoin* 9, 122.

Noh, Eun-Ju. 1998. Echo questions: Meta-representation and pragmatic enrichment. *Linguistics and Philosophy*, 603-628.

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

Oiry, Magda. 2011. A case of true optionality: Wh-in situ patterns like long movement in French. *Linguistic Analysis,* 37(1), 115-142.

Omaki, Akira, & Jeffrey Lidz. 2015. Linking parser development to acquisition of syntactic knowledge. *Language Acquisition*, *22*(2), 158-192.

Owen, Adrian, Kathryn McMillan, Angela Laird, & Ed Bullmore. 2005. N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human brain mapping, 25(1), 46-59.

Palasis, Katerina, Richard Faure, & Frédéric Lavigne. 2019. Explaining variation in wh-position in child French: A statistical analysis of new semi-naturalistic data. *Language Acquisition*, 26:2, 210-234.

Park-Johnson, Sunny. 2017. Crosslinguistic influence of wh-in-situ questions by Korean-English bilingual children. *International Journal of Bilingualism*, 21(4). 419–432.

Pelegrina, Santiago, Teresa Lechuga, Juan García-Madruga, Rosa Elosúa, Pedro Macizo, Manuel Carreiras, Luis Fuentes, & Teresa Bajo. 2015. Normative data on the n-back task for children and young adolescents. *Frontiers in psychology* 6: 1544.

Pelucchi, Bruna, Jessica Hay, & Jenny Saffran. 2009. Statistical learning in a natural language by 8-month-old infants. *Child Development*, 80(3), 674-685.

Perfors, Amy, Joshua Tenenbaum, & Terry Regier. 2011. The learnability of abstract syntactic principles. *Cognition.* 118(3):306-338.

Perfors, Amy. 2012. When do memory limitations lead to regularization? An experimental and computational investigation. *Journal of Memory and Language, 67*(4), 486–506.

Perkins, Lauren, Naomi Feldman, & Jeffrey Lidz. 2017. Learning an input filter for argument structure acquisition. In *Proceedings of the 7th Workshop on Cognitive Modeling and Computational Linguistics,* 11–19.

Pierrehumbert, Janet. 1980. *The Phonology and Phonetics of English Intonation*. Massachusetts Institute of Technology doctoral thesis.

Pires, Acrisio, & Heather Taylor. 2007. The syntax of wh-in-situ and common ground. *Proceedings from the Annual Meeting of the Chicago Linguistic Society*, 43(2), 201-215.

Platzack, Christer. 1996. The initial hypothesis of syntax. *Generative perspectives on language acquisition*, 369-414.

Polišenská, Kamila. 2011. The influence of linguistic structure on memory span: repetition tasks as a measure of language ability. City University London PhD dissertation.

Pozzan, Lucia, & Virginia Valian. 2016. Asking questions in child English: Evidence for early abstract representations. *Language Acquisition*, *24*(3), 209–233.

Prasada, Sandeep, & Stenven Pinker. 1993. Generalisation of regular and irregular morphological patterns. *Language and cognitive processes*, *8*(1), 1-56.

Prévost, Philippe, Laurice Tuller, Marie Barthez, Joelle Malvy, & Frédérique Bonnet-Brilhault. 2017. Production and comprehension of French wh-questions by children with ASD. *Applied Psycholinguistics*, *38*(5), 1095–1131.

Prévost, Philippe, Nelleke Strik, & Laurie Tuller. 2014. Wh-questions in child L2 French: Derivational complexity and its interactions with L1 properties, length of exposure, age of exposure, and the input. *Second Language Research*, *30*(2), 225-250.

Prince, Alan, & Paul Smolensky. 2004. *Optimality Theory: Constraint interaction in generative grammar*. John Wiley & Sons.

Puskas, Genovéva. 1992. The wh-criterion in Hungarian. *Rivista di grammatica generative,* 141-186.

Reis, Marga. 2012. On the analysis of echo questions. *Tampa Papers in Linguistics* 3.1–24.

Rizzi, Luigi. 1990. *Relativized minimality*. Cambridge: MIT Press.

Rizzi, Luigi. 1996. Residual verb second and the wh criterion. In Adriana Belletti & Luigi Rizzi (eds.), *Parameters and Functional Heads*, 63-90. Oxford: Oxford University Press.

Roeper, Thomas, & Jill de Villiers. 2011. The Acquisition Path for Wh-Questions. *Studies in Theoretical Psycholinguistics Handbook of Generative Approaches to Language Acquisition*, 189–246.

Rohde, D., & Plaut, D. 1999. Language acquisition in the absence of negative evidence: How important is starting small? *Cognition*, 72, 67–109.

Rooth, Mats. 1992. A theory of focus interpretation. *Natural Language Semantics* 1:75-*116*.

Rowland, Caroline. 2007. Explaining errors in children's questions. *Cognition*, 104(1), 106-134.

Rowland, Caroline, & Julian Pine. 2000. Subject–auxiliary inversion errors and wh-question acquisition:'What children do know?'. *Journal of child language*, 27(1), 157-181.

Rowland, Caroline., Julian Pine, Elena Lieven, & Anna Theakston. 2005. The incidence of error in young children's wh-questions. *Journal of Speech, Language, and Hearing Research,* 48(1), 384-404.

Sachs, Jacqueline. 1983. Talking about the there and then: The emergence of displaced reference in parent–child discourse. In K. E. Nelson (Ed.), *Children's language*, Vol. 4, Hillsdale, NJ: Lawrence Erlbaum Associates.

Saffran, Jenny, Richard Aslin, & Elissa Newport. 1996. Statistical learning by 8-month-old infants. *Science,* 274(5294), 1926-1928.

Saldana, Carmen, Kenny Smith, Simon Kirby, & Jennifer Culbertson. 2010. Is Regularization Uniform across Linguistic Levels? Comparing Learning and Production of Unconditioned Probabilistic Variation in Morphology and Word Order. *Language Learning and Development* 17(2),158-188.

Sanborn, Adam, Thomas Griffiths, and Daniel Navarro. 2010. Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

Sato, Yosuke, & Jian Ngui. 2017. Wh-questions in Singapore English tell us what about questions *with* declarative syntax?. *Glossa: a journal of general linguistics*, *2*(1).

Saylik, Rahmi, Evren Raman, & Andre Szameitat. 2018. Sex differences in emotion recognition and working memory tasks. *Frontiers in psychology*, *9*, 1072.

Schneider, Jordan, Lauren Perkins, & Naomi Feldman. 2019. A Noisy Channel Model for Systematizing Unpredictable Input Variation. *Proceedings of the 44th Annual Boston University Conference on Language Development*. Cascadilla Press. 533-547.

Schwab, Jessica, Casey Lew-Williams, & Adele Goldberg. 2018. When regularization gets it wrong: Children over-simplify language input only in production. *Journal of Child Language, 45*(5), 1054–1072.

Seidl, Amanda, George Hollich, & Peter Jusczyk. 2003. Early understanding of subject and object wh-questions. *Infancy*, *4*(3), 423-436.

Sekerina, Irina, & John Trueswell. 2012. Interactive processing of contrastive expressions by Russian children. *First Language*, *32*(1–2), 63–87.

Seshadri, Shreyas, Ulpu Remes, & Okko Räsänen. 2017. Comparison of non-parametric Bayesian mixture models for syllable clustering and zero-resource speech processing. In *INTERSPEECH 2017.*

Sethuraman, Jayaram. 1994. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650.

Shlonsky, Ur. 1997. *Clause structure and word order in Hebrew and Arabic: An essay in* comparative *Semitic syntax*. Oxford University Press.

Shlonsky, Ur. 2012. Notes on wh in situ in French. In: Brugé, L., Cardinaletti, A., Giusti, G., Munaro, N. & Poletto, C. *Functional Heads*. Oxford : Oxford University Press.

Sikansi, Nilmara. 1999. A aquisição das interrogativas-Q do Português do Brasil [The acquisition of wh-questions in Brazilian Portuguese]. *General Examination Paper*.

Singleton, Jenny & Elissa Newport. 2004. When learners surpass their models: the acquisition of American Sign Language from inconsistent input. *Cognitive psychology*, *49*(4), 370–407.

Sobin, Nicholas. 1990. On the syntax of English echo questions. *Lingua* 81(2-3), 141-167.

Sobin, Nicholas. 2010. Echo questionsin the minimalist program. *Linguistic Inquiry*, 41(1). 131–148.

Soveri, Anna, Jan Antfolk, Linda Karlsson, Benny Salo, & Matti Laine. 2017. Working memory training revisited: A multi-level meta-analysis of n-back training studies. Psychonomic *bulletin & review*, *24*(4), 1077-1096.

Stalnaker, Robert. 1978. Assertion. In *Pragmatics: Syntax and Semantics*, volume 9. ed. P.Cole, New York: Academic Press.

Stepanov, Arthur, & Wei-Tien Dylan Tsai. 2008. Cartography and licensing of wh-adjuncts: a cross-linguistic perspective. *Natural Language & Linguistic Theory* 26(3), 589-638.

Suzuki, Norio. 2012. Reflections on Chomsky's strong minimalist thesis II: What UG Should be Like in the Context of Plato's and Darwin's problems. *Yokohama: Shumpusha Publishing*.

Takahashi, Mari. 1991. The acquisition of echo questions. In Thomas Maxfield and Bernadette Plunkett (eds.). *Papers on the acquisition of wh: Proceedings of the UMass Roundtable 1990.* Amherst: GLSA.

Thompson, Susan, & Elissa Newport. 2007. Statistical learning of syntax: The role of transitional probability. *Language learning and development* 3(1), 1-42.

Tomasello, Michael. 2005. Beyond formalities: The case of language acquisition. *The Linguistic Review,* 22, 183-197.

Valian, Virginia, & Lyman Casey. 2003. Young children's acquisition of wh-questions: The role of structured input. *Journal of Child Language, 30,* 117-143.

Valian, Virginia, Ingeborg Lasser, & Deborah Mandelbaum. 1992. Children's early questions. In *17th Annual Boston University Conference on Language Development, Boston, MA*.

Valian, Virginia. 1991. Syntactic subjects in the early speech of American and Italian children. *Cognition*, *40*(1-2), 21-81.

Van Hout, Angeliek, Alma Veenstra, & Sanne Berends. 2011. All pronouns are not acquired equally in Dutch: Elicitation of object and quantitative pronouns. In *the 4th Conference on Generative Approaches to Language Acquisition North America,* 106-121.

Van Houten, Lori. 1986. *Role of maternal input in the acquisition process: The communicative strategies of adolescent and older mothers with their language learning children*. Paper presented at the Boston University Conference on Language Development, Boston.

van Kampen, Jacqueline. 2004. An acquisitional view on optionality. *Lingua*, *114*(9-10), 1133-1146.

Vieira, Clariana, & Elaine Grolla. 2020. The pragmatics of wh-in-situ questions in Brazilian Portuguese: data from child and adult language. *Proceedings of the 44th Boston University Conference on Language Development*, 677-690.

Von Eckardt, Barbara, & Mary Potter. 1985. Clauses and the semantic representation of words. *Memory & Cognition*, *13*(4), 371-376.

Weinberg, Amy. 1990. Markedness versus maturation: The case of subject-auxiliary inversion. *Language Acquisition* 1(2), 165-194.

Weist, Richard & Andrea Zevenbergen. 2008. Autobiographical memory and past time reference. *Language Learning and Development, 4*(4), 291 – 308.

Westergaard, Marit. 2009. Microvariation as diachrony: A view from acquisition. *Journal of Comparative* Germanic *Linguistics* 12.1: 49-79. Westergaard, Marit. 2014. Linguistic variation and micro-cues in first language acquisition. *Linguistic Variation*, *14*(1), 26-45.

Wilde, Nancy, Esther Strauss, & David Tulsky. 2004. Memory Span on the Wechsler Scales. *Journal* of *Clinical and Experimental Neuropsychology*, 26(4), 539-549.

Wonnacott, Elizabeth. 2011. Balancing generalization and lexical conservatism: An artificial language study with child learners. *Journal of Memory and Language*, *65*(1), 1-14.

Yang, Charles. 2002. *Knowledge and Learning in Natural Language*. Oxford University Press, Oxford.

Yang, Charles. 2010. Three factors in language variation. *Lingua*, *120*(5), 1160-1177.

Yang, Charles. 2016. *The price of productivity.* MIT press.

Yip, Virginia, & Stephen Matthews. 2000. Syntactic transfer in a Cantonese–English bilingual child. *Bilingualism: Language and Cognition,* 3(3), 193-208.

Yip, Virginia, & Stephen Matthews. 2007. *The bilingual child: Early development and language contact*. Cambridge: Cambridge University Press.

Yuan, Boping. 2015. The effect of computational complexity on L1 transfer: Evidence from L2 Chinese attitude-bearing wh-questions. *Lingua*, *167*, 1-18.

Yurovsky, Daniel, Ty Boyer, Linda Smith, & Chen Yu. 2013. Probabilistic cue combination: Less is more. *Developmental Science* 16(2), 149-158.

Zhu, Zihui & Nguyen, An. 2022. The interaction between structure, discourse, and prosody in wh-questions in English. Talk at the *Chicago Linguistics Society 58,* Chicago, IL.

Zimmermann, Michael & Katharina Kaiser. 2019. Refining current insights into the *wh*-in-situ interrogative construction in French: The case of Contemporary Hexagonal French. *Romanistisches Jahrbuch* 70(1): 123–157.

Zuckerman, Shalom. 2001. *The acquisition of" optional" movement*. University of Groningen Doctoral thesis.

Appendices

Appendix A

Stimuli in the comprehension task (Experiment 1)

Question	Answer for a PQ	Answer for an EQ
The map is where?	On the fridge in the kitchen	In the room with a plant
The boy bought what?	A watermelon cake	A cake with fruit
Kate went on the ferris wheel with who?	With her mom	With her parent
The boy got what?	An ice-cream	A sweet treat
The two kids chose to play what?	A bounce house	A jumping game
The thief is caught by who?	A fireman	A man in uniform
The man got the apples from where?	7-11	A store
The girl got what	A bunny	A white pet
The man went where	To the hospital	To the white building
The squirrel is hiding where?	Behind a rock	Behind something big
The winner is who?	Batman	The man with a mask
The dog belongs to who?	The doctor	The girl in blue

Appendix B

Stimuli in the production task (Experiment 1)

Prompt	Target
Let's ask Beeple about the food the boy is eating.*	The boy is eating what? / What is the boy eating?
Let's ask Beeple if he knows about the zoo.*	The two boys are going where? / Where are the two boys going?
Let's ask Beeple about the person the girl is hugging.	The girl is hugging who? / Who is the girl hugging?
Let's ask Beeple about the game the boy is playing.	The boy is playing what? / What is the boy playing?
Let's ask Beeple if he knows about the fruit on the table.	That is what on the table / What is that on the table?
Let's ask Beeple about the place the children are playing at.	The children are playing where? / Where are the children playing?
Let's ask Beeple about the person the fairy is talking to.	The fairy is talking to who? / Who is the fairy talking to?
Let's ask Beeple about the thing the boy has.	The boy has what? / What does the boy have?
Let's ask Beeple about the place the cat is at.	The cat is where? / Where is the cat?
Let's ask Beeple about the place the man gets his books from.	The man gets his book from where? / Where does the man get his book?
Let's ask Beeple about the person the girl gives the balloon to.	The girl gives the balloon to who? / Who does the girl give the balloon to?
Let's ask Beeple about the lady the kids are playing with.	The kids are playing with who? / Who are the kids playing with?

*Practice trial

Appendix C

Intake size and alpha value in the Data-Alpha condition

Run #	Intake size	Alpha value
1	50	0.01
2	100	0.02
3	200	0.05
4	400	0.1
5	800	0.2
6	1200	0.4
7	1500	0.5
8	1800	0.75
9	2000	1

Appendix D

Participants' comments on the ALL tasks

I definitely noticed a difference in the language when someone was on the phone, versus asking a question to someone in person. But I could not put my finger on the exact rule.
Yes, I tried to pay attention to whether the person being talked to could see the picture.
I noticed that there was direct face-to-face speech vs when over the phone it was like a teaching context
It changed the way the questions were asked based on whether or not the person was face to face or on the phone.
I noticed when questions were asked over the phone, that questions concerning objects were formatted differently.

Appendix E

Sentences used in SRT, adapted from Van Hedger

Trial	Sentence
1	the birds were singing all day long
2	the paper was under the chair
3	the sun was shining throughout the day
4	he entered about eight o'clock that night
5	the pretty house on the mountain seemed empty
6	the lady followed the path down the hill toward home
7	the island in the ocean was first noticed by the young boy
8	the distance between these two cities is too far to travel by car
9	a judge here knows the law better than those people who must appear before him
10	there is a new method in making steel which is far better than that used before
11	this nation has a good government which gives us many freedoms not known in times past
12	the friendly man told us the directions to the modern building where we could find the club
13	the king knew how to rule his country so that his people would show respect for his government
14	yesterday he said that he would be near the village station before it was time for the train to come
15	his interest in the problem increased each time that he looked at the report which lay on the table
16	riding his black horse, the general came to the scene of the battle and began shouting at his brave men