DISTRIBUTIONAL LEARNING OF SYNTACTIC GENERALIZATIONS

Daoxin Li

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2024

Supervisor of Dissertation:

Charles Yang, Professor of Linguistics, Computer and Information Science, and Psychology

Graduate Group Chairperson:

Meredith Tamminga, Associate Professor of Linguistics

Dissertation Committee:

Kathryn Schuler, Assistant Professor of Linguistics

Julie Anne Legate, Professor of Linguistics

Marlyse Baptista, President's Distinguished Professor of Linguistics

John Trueswell, Professor of Psychology

DISTRIBUTIONAL LEARNING OF SYNTACTIC GENERALIZATIONS

This dissertation is dedicated to my grandma, Shufen Wang,

who has always been believing in my wild academic dreams

ever since I first played the "pretending I'm a scientist" game at the age of three.

ACKNOWLEDGMENT

I've imagined this moment countless times, but when it is really coming it feels so unreal. When I started college, I didn't even know that the field of linguistics exists. Yet somehow I have found my passion in this field, gone on an academic journey which turns out to be the most exciting chapter of my life so far, and completed this dissertation with the help of so many amazing people who I never imagined that I could be fortunate enough to know.

First and foremost, my deepest gratitude goes to my advisor Charles Yang and my unofficial advisor Katie Schuler. My words can never do justice to their fundamental impact on me as well as on my research. I can hardly imagine any better graduate school experience than what I've had here at Penn, and Charles is surely the core figure among all the incredible people who make this possible. I still don't know how he has managed to keep this balance: I've fully enjoyed the freedom to explore all my crazy ideas and to take ownership of the work, but I've also received his invaluable guidance and support which have moulded me into a better speaker, writer, thinker, and researcher. I could have gone on for pages but since I have to limit myself, I guess nothing can summarize what he has taught me better than his own words: to do good science - even though it will be hard, to get the numbers right, and to never stop until getting to the bottom of a problem. Thank you also for all the beers and good food!

Katie has contributed an extensive amount of time and efforts to every step in my intellectual, academic and professional growth such that it's hard for me to remember she's not my official advisor. She introduced me to the fascinating world of artificial language learning experiments, and so much more. Our meetings where we discuss and brainstorm about every aspect of the research - as broad as implications to the big pictures of the field or as detailed as technical subtleties with our little alien in the experiment - are among my favorite memories of graduate school. She is a role model for me in terms of how to break disciplinary boundaries, to approach big questions using simple and beautiful methods, and to be an unwavering and supportive scientist, mentor, friend, and person.

I am also extremely grateful to my committee members, Julie Anne Legate, John Trueswell, and Marlyse Baptista, for their helpful and thought-provoking comments and discussion. As a

researcher I have been trained to avoid using words like “absolutely”, “definitely”, “the best”, but I *absolutely* believe that my dissertation committee is *definitely the best* team to work with on syntactic acquisition. Julie showed me the beauty of syntax - a subject that I was scared of when I came to Penn because I didn’t have a strong theoretical background and I was told the syntax courses at Penn were hard; yet now I’m hooked on it - and taught me to always consider all alternatives to my claim and never let go any possible weak points in my analysis. John welcomed me to his lab, a place of incredible intellectual stimulation and fun, and shaped how I think about language and about science in general. Every time I need to make a hard decision in either research or writing and ask myself “What would John do?”, the answer is always to pursue the most rigorous and creative option. Marlyse graciously accepted the role of chair on my proposal committee, and I dragged her into my dissertation committee, which I’m so thankful that she agreed to. Her thoughtful and kind words about my research and my professional development mean a lot to me. It’s a pity that I’m running out of my time at Penn and cannot work more with you!

It has also been a privilege to do my graduate study in the Department of Linguistics and in the amazing language research community at University of Pennsylvania. Thank you to Anna Papafragou, Martin Salzmann, Florian Schwarz, Beatrice Santorini, Dave Embick, Meredith Tamminga, and Dan Swingley for their valuable ideas and feedback in many different venues - lab meetings, seminars, classrooms, and personal discussions; to everyone involved in the ILST initiative, especially Yiran Chen, June Choe, Christine Soh Yue, Sarah Lee, Victor Gomes, Caroline Beech, Sandy LaTourrette, and Tyler Knowlton for many stimulating and productive discussions. Special thanks to Amy Forsyth, Colin Bonner and Jessica Marcus, who magically make our chaotic graduate life much more manageable. And I cannot thank Lila Gleitman enough for her profound influence on language acquisition research. I will never forget the first day I saw her in real life, when I wrote on my social media “OMG I saw a living legend today!” At the LSA meeting this year, I got a response for my work that I’ll cherish forever: “This is work that Lila would be interested in!” - What better response can I expect!

This dissertation would not have been possible without the help from all the members of the Child Language Lab, a group of inspiring and generous people, and the participants. Thank you to

Ariel Mathis, Yiran Chen, Alessandra Pintado-Urbanc, and Madison Paron for being amazing lab managers; to the undergraduate research assistants, especially Iris Zhong, Zaid Tabaza, and Milana Korobko, for their help with data collection; and to all the children and adults who participated in the studies.

Beyond Penn, I'm also grateful to my undergraduate mentors, Xiaolu Yang, Peng Zhou, Li Yin, and Zhongshe Lyu for introducing me to the intriguing field of psycholinguistics and continuing to support me even after I graduated; to Tom Roeper, whose immense passion for the field, extensive knowledge of child language, and genuine supports for students were crucial factors in my decision to pursue a PhD in linguistics; to Elissa Newport, Heidi Getz, and Rushen Shi for their insightful suggestions and kind encouragements; and to Lydia Grohe and Petra Schulz for enjoyable and fruitful collaborations. As I wrote down in my notebook after a conference, "I feel truly grateful that I've had the fortunate to get to know so many great scholars over the years! I only wish that I could work in this field forever."

Thanks are also due to all my friends and colleagues. A shout out to my wonderful cohort: Aini Li, Gwen Hildebrandt, Johanna Benz, Ruicong Sun, Ugurcan Vurgun, Hassan Munshi, and George Balabanian. I thank them and also Yiran Chen, May Chan, June Choe, Xin Gao, Christine Soh Yue, Karen Li, Annika Heuser, Jonathan Lee, Mingyang Bian, Chun-Hung Shih, Sarah Lee, Caroline Beech, Abimael Hernandez Jimenez, Veronica Lyu and many others for their precious friendship. I hope we can still have opportunities to explore local scenic spots and restaurants, make food (sorry I've been better at eating than cooking), play video games and board games (Mahjong included), share fan-fictions, and chat about all linguistic and non-linguistic stuff in the future. Thank you also to those ahead of me at Penn Linguistics for sharing their experience and wisdom, especially Aletheia Cui, Andrea Ceolin, Ryan Budnick, Spencer Caplan, Jordan Kodner, and Ollie Sayeed.

Finally, a million thanks to my family, especially my parents, Yuou Xia and Wanjun Li, and my grandparents, Shufen Wang and Qingbin Xia - using a popular term from the Chinese internet, I'm a 'small-town exam-taker', but they always believe that my life will not be limited to small towns or exam taking and they have been doing everything they can to support me; to Peng Yu

for being my ‘Source of the Peach Blossoms’; to Xuezhu Zheng, who has been there for me for the past 20 years despite the physical distance; to Xiaohan Zhao, who always believes in me even when I didn’t believe in myself; and to Ran Zuo and Rong Fu for going with me through all the ups and downs.

ABSTRACT

DISTRIBUTIONAL LEARNING OF SYNTACTIC GENERALIZATIONS

Daoxin Li

Charles Yang

During language acquisition, children are tasked with the challenge of determining which words can appear in which syntactic constructions. This has been long recognized as a learnability paradox. On one hand, there are generalizations that children must learn. On the other hand, language is known for its arbitrariness, so children also need to decide when not to generalize and just resort to memorization. Finally, the picture is further complicated by the lack of negative evidence during language acquisition. In this dissertation, by applying a generalization learning model, The Tolerance/Sufficiency Principle, I provide novel approaches to the acquisition of a range of syntactic generalizations.

Chapter 2 examines the acquisition of verb argument structure, where there are systematic syntax-semantics mappings. I argue that knowledge of such syntax-semantics mappings should not and need not be innate. Instead, I propose a computational model that can learn these mappings distributionally from modest-sized input data. I also conduct model comparisons to illustrate that the proposed model yields learning outcomes that are more accurate than a model which relies on Bayesian inference.

Chapter 3 moves on to a case where the relation between syntax and semantics is far less systematic - the acquisition of recursive structures. The rules for recursion differ across languages and structures. Through corpus analyses of different recursive structures across languages, I demonstrate that the rules for recursive embedding can be established through purely formal analyses of one-level embedding data, and the core semantic properties such as alienable possession vs. inalienable possession can be identified subsequently.

In Chapter 4, I conduct a series of artificial language learning experiments, which find that both adults and children can indeed use purely distributional cues to acquire recursive structures as predicted: They will allow recursive embedding in an artificial grammar when there are sufficient cues in the exposure supporting the generalization, even though they never hear recursively

embedded sentences in the exposure phase.

Ultimately, this dissertation aims to contribute quantitatively rigorous and psychologically real solutions to a well-known learning problem, offering new perspectives for the mechanisms of learning generalizations.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

An important task that children face in syntactic acquisition is to determine which words can appear in which syntactic constructions. For instance, some verbs can only occur in transitive frames (1), whereas some other verbs can only occur in intransitive frames (2).

(1) a. John touched Bill.

b. *John touched.

(2) a. Bill fell.

b. *John fell Bill.

A prominent view in linguistic literature is that there exist systematic mappings between syntax and semantics. For instance, in the well-known case of verb argument structure, it has been argued that the syntactic frame that a verb can appear in is a regular projection from its meaning (e.g., Gruber, 1965; Jackendoff, 1978; Pinker, 1987). To quote Zwicky (1971): "... if you invent a verb, say *greem*, which refers to an intended act of communication by speech and describes the physical characteristics of the act (say a loud, hoarse, quality), then you know that... it will be possible to greem (i.e., to speak loudly and hoarsely), to greem for someone to get you a glass of water, to greem to your sister about the price of doughnuts, to greem 'Etch' at your enemies, to have your greem frighten the baby, to greem to me that my examples are absurd, and to give a greem when you see the explanation" (Zwicky, 1971, p.232).

Indeed, empirical studies have confirmed that there often exist significant correlations between syntax and semantics cross-linguistically (e.g., Fisher et al., 1991), and that children form productive generalizations of such mappings from a young age (e.g., Naigles, 1990; Fisher, 1996, 2002b; Yuan and Fisher, 2009; Arunachalam and Waxman, 2010). However, on the other hand, it has also been well recognized that such syntax-semantics mappings are by no means perfect, either in a single language or across languages (e.g., Pinker, 1989; Fisher et al., 1991; Bowerman and Brown, 2008). Therefore, we need an account for how learners accurately identify the generalizations in their language in the face of both regularities and exceptions.

In this introduction, I will depict the challenge of learning generalizations, and review major previous attempts to solve the problem, showing that they are all inadequate. The current discussion will focus on the well-studied example of verb argument structure, but this problem is not limited to this case but is instead a common fact across different aspects of language acquisition. Then I will introduce an alternative, threshold-based approach to the acquisition of generalizations. By adopting this approach, in this dissertation, I will show how different types of syntactic generalizations can be accurately captured based on distributional cues available in simple child-directed speech.

1.1 Learning syntactic generalizations: A learnability paradox

It has been well recognized that the acquisition of linguistic generalizations poses a serious learnability paradox (e.g., Baker, 1979; Bowerman, 1987; Pinker, 1989). On one hand, language has *productive* rules, and children have been observed to construct such rules to acquire a language. On the other hand, language also exhibits *arbitrariness*: there are items that do not follow the rules, so children also need to learn when not to generalize. The challenge of language acquisition given this tension is further complicated by the *lack of negative evidence*.

Taking verb argument structure as an example. Verb argument structure is the specification of features of arguments required by a verb for the structure to be well-formed, such as the number and type of arguments. To start with, it is a prominent view in studies of verb argument structure that there are productive rules on the mappings between the meaning of a verb and the syntactic frame

it can appear in: the idea that the verb meaning maps regularly onto its syntactic frames is not only widely discussed in theories of adult grammars (e.g., Fillmore, 1968; Bresnan, 1979; Williams, 1981; Cullicover, 1988; Ladusaw and Dowty, 1988), but has been empirically supported as well. For instance, by asking participants how similar different words are syntactically and semantically, Fisher et al. (1991) confirmed that cross-linguistically, there are syntax-semantics mappings that are significant both linguistically and psychologically: in both English and Italian, verbs considered semantically similar also tend to appear in the same syntactic frames, e.g., *motion* verbs usually allow or require prepositional phrases (PP's) that encode the sources, paths, and/or goals of the movement, and verbs that describe *transfer*, or *change of possessor*, fit naturally into sentences with three NP arguments. Numerous studies with children have also demonstrated that the knowledge of syntax-semantics mapping for verbs is available at a young age. For example, in production tasks where children are taught only the meaning or only one syntactic frame of a novel verb, they can use the verb in other syntactic frames mirroring the mappings in real language (e.g., Maratsos et al., 1987; Pinker, 1987; Tomasello et al., 1997). And when children hear novel verbs in certain syntactic frame, they can infer the verb meaning using the syntactic cues, a learning strategy known as syntactic bootstrapping (Landau and Gleitman, 1985; Gleitman, 1990), e.g., when hearing a verb in a transitive frame, children will prefer a causative meaning to a non-causative meaning (e.g., Naigles, 1990, 1996; Naigles and Kako, 1993; Fisher et al., 1994; Fisher, 1996, 2002b; Lidz et al., 2003; Yuan and Fisher, 2009; Yuan et al., 2012; Arunachalam and Waxman, 2010, 2011).

On the other hand, though, it has been long observed that the mappings between syntax and semantics are not uniform and far from exhaustive. As pointed out by Fisher et al. (1991), "the idea that the structural and semantic partitionings of the lexicon are identical fails... in both directions" (Fisher et al., 1991, p.343). Firstly, there are a range of non-semantic factors that could also influence the structure, such as the overall syntactic and phonological architecture of a given language, and the morphosyntactic history of certain items. For example, for the English verb 'rain', it must be used with a subject such as 'It's raining' although the subject is not doing the raining. An expletive subject is needed simply due to the general requirement for an overt grammatical subject in English. Similarly, 'there' is also used as an expletive subject in the 'there

be...' construction in English. Another example they discussed was 'exit'. Unlike many other motion verbs in modern English, when describing the path of motion, it can directly take an object instead of taking a PP (e.g., 'exit the stage' vs. 'go off the stage'). This is due to historical reasons: In its source languages Latin and old French, 'ex-' is a morpheme that marks the path of motion, just as prepositions in English; as a result, 'exit' does not need another preposition to mark the path. There are many such irregular verbs in modern English as 'exit' which are residuals of historical productive processes, but it is now impossible for the child learner to predict whether a verb is irregular or not. Secondly, given the enormity of conceptual space, it is impossible for all semantic distinctions to be regularly mapped onto syntax. Instead, the structure only provides a very partial semantic partitioning of the verbs. Therefore, as put by Fisher et al. (1991), "at any real stage of a language there are bound to be irregularities in the mapping between semantics and clause structures" (Fisher et al., 1991, p.340), and children must determine when not to generalize despite the *productivity* aspect of language.

How do children strike the balance between the *productivity* and the *arbitrariness* of language and acquire the correct grammar? This problem is further complicated by the fact that it is agreed by linguists and psychologists that children do not have systematic access to negative evidence, i.e., information about what expressions are ill-formed in a language (Brown and Hanlon, 1970; Grimshaw and Pinker, 1989; Marcus, 1993). Negative evidence is considered necessary to distinguish between the target grammar G and a super-set grammar G' ($G \subset G'$) so that children can avoid or retreat from over-generalization (Brown and Hanlon, 1970; Braine, 1971; Baker, 1979). Thus, any accounts for the acquisition of generalizations must acknowledge all three aspects of the paradox. In the next section, I will introduce previous attempted solutions to the paradox, and show how they are all unsatisfactory.

1.2 Previous approaches to the paradox

One approach to the acquisition of verb argument structure is strict lexical conservatism (e.g., Baker, 1979; Fodor, 1985). It suggests that children will only add an argument structure to the lexical entry of a verb when this structure is attested in the children's input. This approach

agrees that children can abstract systematic generalizations from acquired items, but it argues that children will not extend such generalizations to novel words. Therefore, children will never face the problem of over-generalization.

This approach, however, has proven implausible given the empirical evidence that children are not conservative learners but do generalize to unattested items. One type of such evidence comes from experiments where children apply the productive generalizations to pseudo-words, as discussed in the last section. In addition, it is also evidenced by recurring reports of children's over-generalization in spontaneous speech. For instance, examples in (3) demonstrate children's generalization of the double object alternation, and (4) illustrates generalization of the causative alternation. Such over-generalizations are not limited in English, but have been found across languages (e.g., Hebrew (Berman, 1982); Hungarian (MacWhinney, 1985); French, Polish, and Turkish (Slobin, 1985)). Moreover, such over-generalization errors are not predicted by word frequency. For instance, for the causative alternation as shown in (4), some accounts suggested that the child should retreat from such errors earlier for verbs that occur in the intransitive frame with a higher frequency in the input (e.g., Brooks and Tomasello, 1999), but this prediction is not borne out: Although the verbs with which children make over-generalization errors differ drastically in frequency (Table 1.1, Irani, 2019), the errors observed in children's spontaneous speech all cluster together around the ages of 3-4, and there is no evidence that they stop producing errors from these verbs purely on the basis of hearing them in the intransitive frame a number of times.

(3) a. Mommy, fix me my tiger. (Adam, 5;2)

b. I write you something. (Eve, 2;3)

c. Jay said me no. (Ross, 2;8)

(Gropen et al., 1989)

(4) a. She came it over there. (Christy, 3;4)

b. Kendall fall that toy. (Kendall, 2;3)

c. He's gonna die you, David. (Hilary, 4+)

(Bowerman, 1982)

Verb	Raw frequency in CHILDES
disappear	153
stay	2,662
fall	2,819
go	55,689

Table 1.1: Raw frequency of pure intransitive verbs in child-directed speech (Irani, 2019).

Another approach to verb argument structure acquisition argues that children are endowed with universal and innate knowledge of syntax-semantic mapping, and that they use this innate knowledge to 'bootstrap' themselves into a fully abstract and adult-like grammar of argument structure. Specifically, there are two main theories of bootstrapping that highlight different starting points: Semantic bootstrapping focuses on children's use of semantics to bootstrap into syntax (e.g., Pinker, 1984, 1987, 1989), whereas syntactic bootstrapping focuses on children's use of syntax to bootstrap into verb meaning (e.g., Landau and Gleitman, 1985; Gleitman, 1990; Lidz et al., 2003).

However, this hypothesis for innate knowledge has been challenged by the well recognized problem of data coverage. First, even within the well-studied English vocabulary, researchers have been aware that the mappings between syntax and semantics are far from uniform but contain considerable irregularities, as discussed in the last section. Here are a few more well-known examples: Although the mapping between the double object construction ('John gave students a book') and the 'caused-possession' semantics (e.g., Green, 1974; Pinker, 1987) is widely acknowledged, many caused-possession verbs in English cannot be used in the double object construction such as 'donate', 'return' and 'provide' (Levin, 1993); and while many transitive verbs can be optionally intransitive such at 'eat', its near synonym 'devour' cannot. Moreover, cross-linguistic data have presented more variations. For example, ergative languages mark the roles of arguments differently from our familiar accusative languages - intransitive subjects typically pattern with transitive objects and differently from transitive subjects, so children acquiring ergative languages must learn to use different types of morphosyntactic cues to identify the rules in their language (e.g., Bavin and Stoll, 2013). And the productive rules for the number or type of verb arguments in English may simply not hold in other languages. For instance, while motion verbs in English usually allow or require PP's (Gruber, 1965; Talmy, 1975; Jackendoff, 1983, 1987), they cannot do so in languages like

Yukatek Mayan (Bohnemeyer, 2008); and while all languages seem to contain verbs with a caused-possession meaning, there are languages where such verbs are typically used in the transitive rather than double-object frame (e.g., Saliba (Margetts, 2008)), and there are also languages where the double-object construction is not allowed at all (e.g., Chamorro (Chung, 1998)); in Icelandic, the dative is surprisingly robust and productive and can appear in typologically rare semantic contexts such as the themes of motion verbs (e.g., Maling, 2002), and children acquire the language-specific productive rules from a young age (Nowenstein, 2023). Thus, despite cross-linguistic tendencies for syntax-semantics mappings, it is still necessary for the child learner to identify the rules from language-specific experience.

Yet other approaches question the absence of negative evidence. In particular, a prominent hypothesis is that children can use *indirect negative evidence* (Chomsky, 1981): "The child may learn that a certain string is not acceptable by the fact that it never occurs in a certain context" (Gold, 1967, p.454). That is, absence of evidence becomes evidence of absence. However, it has been debated whether indirect negative evidence will indeed be useful. For example, Pinker (1989) reviewed a range of such proposals and pointed out they failed to specify "under exactly what circumstances does a child conclude that a non-witnessed sentence is ungrammatical" (Pinker, 1989, p.17). It cannot be true that children rule out all non-witnessed sentences since they do need to generalize, so Pinker (1989) concluded that the indirect negative evidence approach is ineffective but is "virtually a restatement of the problem" (Pinker, 1989, p.17). A more recent formulation of indirect negative evidence arises from probabilistic learning models such as Bayesian inference (e.g., Xu and Tenenbaum, 2007): While the absence of a sentence in a small corpus may not be useful evidence for its ungrammaticality, its absence in a large corpus would provide stronger evidence. Therefore, the absence of expressions that fall in in the set difference between the target grammar G and its super-set G' may be conspicuous in a large enough sample of input, thus becoming evidence that such expressions are ungrammatical (e.g., Ambridge et al., 2008, 2009). However, studies on various structures have proven this idea implausible by showing that it is generally impossible to distinguish between ungrammatical expressions and grammatical expressions that are simply not attested, given the sparsity of input data (e.g., Yang, 2015; Irani, 2019). I will also discuss

the indirect negative evidence approach further in Chapter 2, where I show that a computational model (Alishali and Stevenson, 2008) which uses indirect negative evidence to learn verb argument structure rules fails to capture human learners' behavior. Therefore, the learning theory must account for when children will generalize and when not using positive evidence alone.

1.3 An alternative approach to learning generalizations

1.3.1 The Tolerance/Sufficiency Principle

In this dissertation, I will take an alternative, threshold-based approach to the acquisition of generalizations. In particular, I will adopt the Tolerance/Sufficiency Principle (TSP), (Yang, 2016), a cognitively-grounded model for the acquisition of productive rules. This section provides an introduction to the formulation, applications and properties of the TSP.

The TSP is built upon the prominent idea that in the presence of regular patterns and irregular exceptions, the more items following the regular pattern, the more efficient it is to form a productive rule (e.g., Aronoff, 1976; Bybee, 1995). In particular, the TSP offers a precise threshold for productive rule formation based on the Elsewhere Condition (Anderson, 1969), a psychological processing model: It has been confirmed by reaction time studies that exceptions are listed and have to be checked off first in word formation before the productive rule can be applied (e.g., Murrary and Forster, 2004). Therefore, as the list of exceptions grows longer, the psychological cost for checking them off before applying the elsewhere condition also increases, such that it may become more efficient to just list all the items. See Yang 2016, Chapter 3 for details. The Tolerance Principle thus provides the threshold for exceptions below which it is more efficient to have a productive rule:

(5) The Tolerance Principle: Let a rule R be defined over a set of N items in the input. R is productive if and only if e, the number of items not supporting R, does not exceed $\theta_N = N/lnN$.

The Sufficiency Principle (Yang, 2016) is a corollary of the Tolerance Principle. It is used to predict the threshold for how many rule-abiding items are sufficient to support a generalization in the face of items not (yet) attested against the rule. This dissertation will refer to the Tolerance

Principle and the Sufficiency Principle together as the TSP.

(6) The Sufficiency Principle: Let a generalization R be defined over a set of N items in the input. R is productive if and only if M, the number of items attested to follow R, exceeds N - N/lnN.

The TSP has been shown to successfully predict generalizations from corpus data in a wide range of linguistic phenomena across languages, including English past tense (Yang, 2016), German plurals (Yang, 2016; Belth et al., 2021), Icelandic gender assignment (Björnsdóttir, 2021) and case marking (Nowenstein, 2023), Spanish tense in bilinguals (Fernández-Dobao and Herschensohn, 2021), Dutch noun diminutives (van Tuijl and Coopmans, 2021), possessive suffixes in Northern East Cree (Henke, 2022), variation and change of verbal inflection in Frisian (Merkuur, 2021), allophonic restructuring of /æ/ in Philadelphia English (Sneller et al., 2019), past participles in Latin (Kodner, 2022), the historical change of English strong verb classes (Ringe and Yang, 2022), verb alternations in English (Irani, 2019), the Uniformity of Theta Assignment Hypothesis in English (Pearl and Sprouse, 2021), among others. Additional support can be found with artificial language learning experiments where the generalization threshold can be precisely manipulated. Schuler et al. (2016) exposed children to the singular and plural forms of 9 pseudo-nouns. In one condition, 5 nouns followed a regular rule of pluralization and 4 nouns were exceptions (5R/4E); in the other condition, 3 nouns followed the regular rule and 6 nouns were exceptions (3R/6E). Since $\theta_9 = 4.2$, the TSP predicts a productive rule in the 5R/4E condition but not in the 3R/6E condition. As predicted, when tested with novel items in a production task, children in the 5R/4E condition categorically applied the regular rule, whereas no children in the 3R/6E condition did, thus supporting the TSP (also see Schuler, 2017). Shi and Emond (2023) provided evidence that even infants can perform formal learning as predicted by the TSP. These authors designed two sets of stimuli, each of which consisted of 16 Russian sentences. In one set, 11 out of the 16 sentences followed a word order pattern; in the other, 10 out of the 16 items followed the pattern. For N=16 items, the critical threshold predicted by the TSP is θ_{16}=5. Therefore, 10 is not sufficient for generalization although it is the majority, but 11 is. Indeed, 14-month-old infants who never learned Russian generalized the pattern in the 11/16 condition, but not in the 10/16 condition.

The TSP has several desirable properties considering the design specifications of language acquisition. Throughout this dissertation I will show how those properties of the TSP also benefit the acquisition of the generalizations examined in this work. First, as shown in Table 1.2, the proportion of exceptions that can be tolerated decreases sharply as N increases, which means it will be easier to learn generalizations from a smaller vocabulary. This potentially explains how children are able to accurately learn the productive generalizations in their language at a very young age, based on their limited vocabulary and sparse input data (Yang, 2016, 2018).

N	θ_N	% of N
10	4	40.0
20	6	30.0
50	12	24.0
100	21	21.0
200	37	18.5
500	80	16.0
1000	144	14.4

Table 1.2: The maximum number of exceptions for a productive rule over N items.

Another important feature of the TSP is its developmental side: As children learn more words, the threshold will shift accordingly and the decision whether there is a productive rule may thus change. That is, even though children do not have a productive rule at a given point, they may end up identifying one if their increasing vocabulary contains sufficient new items supporting the rule; on the other hand, it is also possible that a productive rule that children discover at an earlier point will fall out of productivity if too many exceptions have been added to the vocabulary. Indeed, children's over-generalization errors have been observed in various linguistic phenomena before they retreat to the adult grammar (e.g., Baker, 1979; Pinker, 1989), and this trajectory can be successfully captured by the TSP (e.g., Yang, 2016; Irani, 2019; Belth et al., 2021). It makes sense to say that the over-generalizations children produce are not real mistakes; instead, the children are applying exactly the rules that are learnable from the positive evidence available to them at that point.

Another note on the TSP is that it is a formal learning model, which concerns only with data coverage but not with the nature of the hypothesis, as evidenced by Shi and Emond (2023)'s study:

The infants who did not know any Russian were able to acquire the generalization by just being exposed to the Russian sentences without knowing what they meant. In fact, Yang (2016) suggested that the TSP can be viewed as a general condition on the validation of a function that forms a mapping between two domains.

It is also worth noting that the TSP can apply recursively: Given a set of words, if there is a global rule that is sufficient to account for all the words, then a global rule will be learned and the irregular words will be lexically memorized. Otherwise, the set of words will be partitioned into subdivisions along some dimension - for language, it can be either linguistic (e.g., phonology, semantics) or non-linguistic dimensions (e.g., social hierarchy) - and the TSP recursively applies with the subdivisions. In this way, local productive rules can be discovered. If no productive rule can be found in the subdivisions either, then the learner will just resort to lexical memorization and not generalize beyond the input. The recursive application of the TSP resolves the issue that although a high type frequency is usually recognized as a prerequisite for productivity, it does not necessarily equate with productivity; instead, productivity can be observed in non-dominant classes. For example, there are several different suffixes for German noun pluralization, including '-(e)n', '-e', '-er', '-s' and the null $\emptyset$. Interestingly, the '-s' suffix has the smallest type frequency but is productive and applies to novel nouns (e.g., 'iPhones'). Multiple other suffixes are also productive but are conditioned on gender and/or phonology of the noun. Their productivity is evidenced by German-learning children's over-generalization in early language acquisition (Elsen, 2002; Kauschke et al., 2011). Belth et al. (2021) proposed a computational model which recursively searches for morphological productivity based on the TSP and successfully captures the developmental patterns of the nested productive rules for German noun plurals when trained on psychologically realistic data from child-directed input.

Finally, a crucial difference between the TSP and many existing language learning models is that the TSP does not determine the *best* grammar for a corpus but only an *adequate* one that covers a sufficient proportion of the data. As we will discuss more in later chapters, the search for the best grammar is not only computational intractable but also results in incorrect empirical results. Instead, the TSP does not need mapping rules to be perfect. A generalization can be

warranted in the presence of exceptions, as long as a sufficiently large number of items can be covered.

1.3.2 Applying the TSP to syntactic acquisition

This section reviews an example of applying the TSP to the acquisition of one type of syntactic generalization: learning the double-object generalization in English. I introduce the analysis by Yang (2016); these findings have been replicated by Subramanian (2019) using a modified, more conservative classification of these verbs.

In English, many but not all verbs that can be used in the 'to' dative construction can be also used in the double object construction, such as 'donate' and 'guarantee'. Children are observed to use novel verbs in the double object construction in experiments and overgeneralize this construction to verbs that cannot participate in it in spontaneous speech, such as 'say' 'whisper' and 'demonstrate' (e.g., Gropen et al., 1989; Conwell and Demuth, 2007), indicating that they have constructed a productive rule. Even after they learn the lexically arbitrary restrictions, they can still discover productive regularities within certain subclasses. For example, when new verbs such as 'fax' 'tax' and 'email' entered the English lexicon, the double object construction was available. Moreover, the learning cannot be entirely attributed to universal constraints given the cross- and within-linguistic differences in what kinds of verbs can be used so (e.g., Levin, 1993; Harley, 2002; Jelinek and Carnie, 2003; Jung and Miyagawa, 2004). Therefore, the English double object construction exhibits typical patterns of a learnability paradox.

To estimate English-speaking children's input experience, Yang (2016) constructed a five-million-word corpus of child-directed North American English from the CHILDES database and examined the distribution of verbs that participate in the double-object construction ('verb NP NP'). The corpus size is comparable to one year's linguistic input for working-class children. It was found that a total of 42 verbs were used in the double-object construction in the input, and 38 of them have a clearly identifiable semantics of 'caused possession', which is assumed identifiable to the learner (Grimshaw, 1990; Jackendoff, 1990; Pinker, 1989). Four exceptions are well below the threshold ($\theta_{42} = 11$), so the widely recognized semantic condition as a prerequisite for the

double-object construction as stated below (e.g., Gropen et al., 1989; Pinker, 1989; Pesetsky, 1995) can be learned from language-specific data and does not need to be baked into UG.

(7) In English, if a verb appears in the double-object construction, then it will have the semantics of caused possession, a tolerably low number of exceptions notwithstanding.

Having established (7), the child will consider its converse: whether caused possession can be established as a sufficient condition for the double-object construction. From the same CHILDES corpus, Yang (2016) found additional 11 verbs that have the semantics of caused possession but are not attested in the double-object construction: 'address', 'deliver', 'describe', 'explain', 'introduce', 'return', 'transport', 'ship', 'mention', 'report', 'say'. Some of the words here such as 'ship' can appear in the double-object construction but did not have the opportunity to do so in the child-directed corpus; whereas there are also words that do not allow the double-object construction such as 'say'. Children do not know this distinction, but the number of words that are attested in the double-object construction (38) is sufficient for children to acquire the generalization below, given the entire set of N=49 words ($\theta_{49} = 12$).

(8) In English, if a verb has the semantics of caused possession, then it can appear in the double-object construction.

This explains the well-documented over-generalization errors such as "Jay say me no", and also children's productive application of the double-object construction to novel verbs with the appropriate semantics in experimental settings (e.g., Conwell and Demuth, 2007).

The TSP can also account for how children retreat from the (over)generalization. According to Levin (1993), out of 253 English verbs with a cause-possession semantics, only 115 allow the double-object construction, which is far from the productivity threshold. Even when Yang (2016) trimmed the verbs to a relatively common set of 92 based on suitable frequency estimates, there are still 40 of them that cannot appear in the double-object construction, far exceeding the threshold ($\theta_{92} = 20$). Therefore, as the learner acquires more words, the generalization in (8) will no longer be productive, and the learner will lexicalize every verb that appears in the double-object construction instead.

The properties of the most frequent caused possession verbs can offer a learning-theoretic account for many well-documented regularities in the double object construction and its acquisition as well. For instance, some theoretical literature (e.g., Harley and Miyagawa, 2016) proposes that there are structural constraints on the length of the verbs that can participate in the double object construction. Yang (2016) showed that 50 out of the 92 most frequent caused possession verbs in the input are monosyllabic, among which 42 can take double objects. By contrast, only 10 out of the 42 polysyllabic verbs allow double objects. Therefore, it would be predicted that given a novel verb with a caused possession meaning, speakers will prefer a short verb to a long verb in the double object construction. Therefore, contra the proposals for structural constraints, this regularity is just a consequence of distributional learning. Also, although the entire class of caused possession verbs cannot be categorically used in the double object construction as the vocabulary grows, productivity can still hold in semantic subclasses of words as long as language learners can construct such subclasses (Pinker, 1989). Again, the fact that productivity is more likely to hold for a smaller value of N plays a role here.

Having illustrated the application of the TSP, I will show how it can be used to account for a broader range of syntactic generalizations in this dissertation.

1.4 Overview of the chapters

This dissertation is intended to expand our understanding of the learning mechanism of syntactic generalizations. I aim to provide novel solutions to the well-known learnability problem, showing that different types of syntactic generalizations are learnable from distributional cues in realistic child-directed speech.

In Chapter 2, I start with the acquisition of verb argument structure, where systematic mappings between syntax and semantics have been well-documented. I argue that even in this case where systematic syntax-semantics mapping is present, knowledge of such syntax-semantics mapping should not and need not be innate. Instead, I propose a computational model that can learn these mappings distributionally from modest-sized input data. I also conduct model comparisons to illustrate that the proposed model yields learning outcomes that are more accurate and more

consistent with human behavior than another model which relies on Bayesian inference.

Chapter 3 moves on to a case where the relation between syntax and semantics is far less systematic - the acquisition of recursive structures. The rules for recursion differ across languages. Some structures are freely recursive; some structures are regulated by semantic constraints which must be learned from language specific experience; and some structures in general cannot recurse. In addition, even for structures where recursion is productive, examples of recursive embedding are rare in children's input, which poses another acquisition challenge. In this work, I propose a new conceptualization of recursion based on its formal properties in non-recursively embedded data, which leads to a new theory of how it can be acquired. Through corpus analyses of different recursive structures across languages, I demonstrate that the rules for recursive embedding can be established through purely formal analyses of one-level embedding data, and the core semantic properties such as alienable possession vs. inalienable possession can be identified subsequently.

In Chapter 4, to determine whether the proposed distributional learning mechanism can indeed be helpful during language acquisition, I conduct a series of artificial language learning experiments, which find that both adults and children can use purely distributional cues to acquire recursive structures: They will allow recursive embedding in an artificial grammar when there are sufficient cues in the exposure supporting the generalization, even though they never hear recursively embedded sentences in the exposure phase.

Finally, Chapter 5 summarizes the findings and discusses the implications and future directions.

Chapter 2

DISTRIBUTIONAL LEARNING OF VERB ARGUMENT STRUCTURE

2.1 Introduction

This chapter investigates the acquisition of verb argument structure, i.e., the specification of features of arguments required by a verb for the structure to be well-formed, such as the number and type of arguments. As shown in Chapter 1, the argument structure of a verb is often associated with the semantics of the verb, and children are sensitive to such regularities from an early age. I reiterate some of the well recognized mappings in English below; as discussed above, syntax-semantics mappings are also found cross-linguistically although the exact form could differ across languages (e.g., Fisher et al., 1991). However, Chapter 1 has also shown it is agreed that such mappings are far from perfect. Therefore, children are tasked with the challenge of learning the regular mappings between syntax and semantics in their language as well as keeping track of the proportions of regular items and exceptions.

(9) Verbs that denote *causation* appear in the transitive frame, whereas verbs that do not cannot (e.g., Landau and Gleitman, 1985; Naigles, 1990).

(10) *Motion* verbs such as 'move' and 'walk' allow or require PP's that indicate the sources, paths, and goals of movement (e.g., Gruber, 1965; Talmy, 1975; Jackendoff, 1983, 1987).

(11) Verbs that describe *transfer*, or *change of possessor* such as 'give' and 'send' can take double objects (e.g., Jackendoff, 1978; Pinker, 1987).

(12) Verbs of *perception* and *cognition* such as 'see' and 'think' allow CP complements (e.g., Vendler, 1972).

In this chapter, I first review empirical findings and previous theoretical approaches for children's acquisition of syntax-semantics mapping rules for verb argument structure. I will show that previous solutions to the learning problem are inadequate. Then, I propose a model that automatically discovers regular mappings between syntax and semantics from simple child-directed speech. By training the model on realistic input data, I show that it can successfully learn many of the well documented mapping rules in English, reducing the need for any innate, universal theory of syntax and semantics mappings. I also demonstrate the plausibility of the proposed model through model comparisons, where the learning outcomes of the current model are not only more accurate but also more consistent with human behavior than a Bayesian inference model.

2.1.1 Acquisition of verb argument structure

Early spontaneous speech and elicited production studies have provided evidence that children's usage of syntactic frames in production is largely consistent with the syntax-semantic mapping rules in their language (Pinker et al., 1987; Gropen et al., 1989; Pinker, 1989). For example, children have been observed to overgeneralize the double object construction in spontaneous speech, (13). Gropen et al. (1989) found that 4 out of the 5 children that they studied made such over-generalization errors, and the semantics of the verbs can be characterized as either *change of possessor* or *benefactive*, mirroring the regular mappings in English. Such over-generalization errors occurred after grammatical usage of the double object construction, with the onset ranging from 2;3 to 4;1 for different children, and could stay for extended periods of time: for example, Adam from the Brown corpus was observed to produce his first over-generalization error at 4;1, and still did in his last file at 5;2. These authors also conducted an experiment where 5-8-year-old children were taught a novel verb that had a change of possessor meaning, and they found that the children would use the verb in both the double object frame and the 'to' dative frame. Similarly, for verbs

with a change of location meaning, children have been observed to over-generalize the 'verb - direct object - PP' frame to produce sentences such as "* I filled water into the glass" within the range of three to seven years of age, after which the errors would decline (e.g., Bowerman, 1982; Pinker, 1989; Gropen et al., 1991b).

(13) a. Pass me some more horsies. (Eva, 2;0)

b. Then put her some more. (Eva, 2;4)

c. I said her no. (Christy, 3;1)

d. Mattia demonstrated me that yesterday. (Damon, 8;0)

(Bowerman, 1987; Gropen et al., 1989)

There are also numerous studies demonstrating that children can use the syntactic frame where a verb appears to infer its meaning, a learning strategy named syntactic bootstrapping. In a seminal study by Naigles (1990), 25 months olds watched a video consisting of two simultaneous events: e.g., a rabbit repeatedly pushed a duck over (a causative event) while both the rabbit and the duck circled their arms independently (a non-causative event). One group of children heard a transitive sentence while watching the video ("The bunny is gorping the duck") while the other heard an intransitive sentence ("The bunny and the duck are gorping"). In the test, in a preferential looking paradigm, both groups saw the two events separated – pushing on one screen and arm-wheeling on the other – while hearing the prompt sentence "Where's gorping now? Find gorping!" It was found that the group that heard the transitive sentence looked longer at the causative event and the group exposed to the intransitive sentence looked longer at the non-causative event, suggesting that children can use the mapping rule between transitivity and causation to infer the meaning of a novel verb. This findings has been replicated and extended by extensive studies using similar paradigms (see Naigles and Swensen, 2007, for a review of earlier studies). Now it is established that infants as young as 19 months reliably show a preference for a causative event for a novel verb that was heard in a transitive sentence from simple dialogues without visual information for the events (Yuan and Fisher, 2009; Yuan et al., 2012; Arunachalam and Waxman, 2010; Arunachalam et al., 2013a; Messenger et al., 2015). Jin and Fisher (2014) found that even 15-month-olds demonstrated

this preference when tested with simplified visual stimuli. In addition to novel verbs, children can also adjust their interpretation of familiar verbs to fit a new syntactic frame: e.g., 2-4 year olds would act out a transfer event when hearing an intransitive verb used in a dative construction "Noah goes the elephant to the ark" (e.g., Naigles et al., 1992).

While most studies on syntactic bootstrapping were conducted with English-speaking children, there is also evidence that syntactic bootstrapping is not limited to English. 2-year-old children can use the transitive frame as a cue to causative meaning in different languages such as French (Naigles and Lehrer, 2002), Kannada (Lidz et al., 2003), and Mandarin Chinese (Lee and Naigles, 2008). In addition, language specific properties also play a role. For example, different from English, Turkish has less strict word order rules but is morphologically richer, where case marking provides important information about argument roles. Göksun et al. (2008) found that Turkish-speaking 2-year-olds were much better at the syntactic bootstrapping task when case-marking was present. In Japanese, where a range of word orders are possible and case marking is rare and variable, 2-year-olds needed both case and word order information to succeed in the task (Matsuo et al., 2012). In another study, Arunachalam et al. (2013b) found that in contrast to English-speaking children, 2-year-old Korean-speaking children performed better in verb learning when the novel verb appeared in a sparse linguistic context, which the authors attributed to the fact that rich linguistic contexts are less frequent in Korean since the language allows extensive argument omission. Nowenstein (2023) examined morphosyntactic bootstrapping in Icelandic, a language with both rich morphological case marking and a relatively rigid word order, and observed that children can use case to determine verb meaning when word order is uninformative, and that case morphology can be as salient as the number of arguments in specific contexts.

In addition to mapping rules between syntax and semantics, children also exhibit knowledge of construction alternations. In this work I will focus on the causative-incohative alternation, such as "Tom broke the vase" and "The vase broke". Bowerman (1982) documented her children's over-generalization errors, as in (14). (15) shows all the verbs overgeneralized in a causative frame in Bowerman 1982 and their semantic subclasses summarized by Pinker (1989). These errors typically emerged after 2 years of age and may extend beyond 7 years of age. Experimental studies have

provided additional evidence for the productivity of this alternation. For example, Maratsos et al. (1987) conducted an experiment where they introduced a novel intransitive verb 'fud' to adults and children aged 4;6 to 6;2. The verb referred to a dough-like substance being converted into strands by a machine. It was found that the participants could spontaneously produce the verb in a causative frame even when they never heard it in that frame. In another study, Gropen et al. (1991a) also found that children would use an intransitive verb transitively when it depicted a direct causative event.

(14) a. She came it over there. (Christy, 3;4)

b. Kendall fall that toy. (Kendall, 2;3)

c. He's gonna die you, David. (Hilary, 4+)

(Bowerman, 1982)

(15) a. Directed motion: *come, go, fall, rise, drop*

b. Going out of existence: *die, disappear, vanish*

c. Being/staying: *stay, be, spell, sound, wait*

d. Possession: *have, take*

e. Psychological: *remember, watch, guess, wish, feel, ache, learn*

f. Involuntary emission: *sweat, blood*

g. Internally caused state change: *bloom*

h. Semivoluntary expression of emotion: *laugh, cry, giggle*

i. Voluntary action: *eat, drink, sing, talk, swim, climb*

(Pinker, 1989, p.303)

Studies have also shown that children are aware of the semantic constraints for verbs that can participate in this alternation. For example, Brooks and Tomasello (1999) taught 4-year-old children two novel intransitive verbs by having an apparatus performing the novel actions, one with manner of motion semantics (e.g., *roll*) and the other with directed motion semantics (e.g., *come*). They found that children used the first verb in the causative frame but not the other. Kline and

colleagues demonstrate that children were sensitive to the semantic distinction between unergative intransitives and unaccusative intransitives and that external causation was crucial for preschoolers' causativization of novel intransitive verbs (Kline and Demuth, 2014; Kline et al., 2017).

Overall, different lines of work have converged on the finding that children have productive knowledge of verb argument structure at a young age. Knowledge of the transitive-causation mapping rule is particularly early and robust. These findings hold across languages; on the other hand, it is worth noting that children are also sensitive to language specific patterns, indicating the role of learning from language specific experience.

2.1.2 Previous approaches

In this section I review the major previous approaches to the acquisition of verb argument structure. First, as we discussed earlier, the strict lexical conservatism (e.g., Baker, 1979; Fodor, 1985) is clearly untenable given children's productive knowledge.

A prominent idea in the acquisition of verb argument structure is that children break into the system using innate linking rules between syntax and semantics. According to the semantic bootstrapping approach (Pinker, 1984, 1989), children start with focusing on the semantics of the event denoted by a verb, and they use the semantics to bootstrap themselves into syntax. By comparison, the syntactic bootstrapping approach (Gleitman, 1990) proposes that children first attend to the syntactic frame that a verb appears in, which is used as a 'syntactic zoom lens' (Fisher et al., 1994) to infer the meaning of the verb. The empirical findings in the last section provide clear evidence that children have productive knowledge of the mapping rules between syntax and semantics and that they are able to use one to infer the other. However, this does not necessarily mean that such knowledge is innate. Given the data coverage problem both across- and within-languages as discussed in Chapter 1, the innateness approach cannot be adequate.

An alternative view, the usage-based approach, argues that children have no innate knowledge but instead learn verb argument structure rules from the input using domain general cognitive mechanisms (e.g., Tomasello, 2000). For example, through analysis of children's longitudinal data, some argued that children's usage of verbs was restricted and did not easily generalize before 2 years

of age (Tomasello, 1992; Lieven et al., 1997; McClure et al., 2006). In some elicited production tasks on construction alternations such as the causative-incoative alternation and active-passive alternation, the rate that children used a novel verb in a construction that they never heard was low (see Tomasello, 2000, for a review). Moreover, adding a training phase with familiar verbs that could participate in these alternations was helpful, which was taken as evidence that learning of the alternation rules is influenced by input frequency (Childers and Tomasello, 2001; Abbot-Smith et al., 2004). Such findings, however, have been claimed by others to be effects of priming and effects of performance and processing demands, not necessarily evidence for children's lack of productive knowledge (e.g., Fisher, 2002a).

A particularly relevant aspect of the usage-based theory is the indirect negative evidence approach (e.g., Tomasello, 2000; Goldberg et al., 2004; Ambridge et al., 2008), which has been claimed to offer a solution to the learnability paradox by taking absence of evidence as evidence of absence. This approach relies on input frequency: Over time, an infrequent form will be overridden by a more frequent form in the input. In particular, I will introduce the theories of statistical preemption and entrenchment. The statistical preemption approach argues that a frequent form in the input rules out the infrequent form hypothesized by the learner with a similar meaning (e.g., Bates and MacWhinney, 1987; Goldberg, 1995). For example, learners may hypothesize a transitive form of the verb 'die'. They will not hear this form in the input. On the other hand, 'kill' has a similar meaning to the transitive 'die' and is more frequent in the input. Therefore, learners will eventually replace the transitive 'die' that they hypothesized with 'kill'. By comparison, according to the entrenchment approach, a verb frequently used in one frame in the input will be evidence against its usage in other frames (e.g., Braine and Brooks, 1995; Ambridge et al., 2008). For example, some studies suggested that intransitive-only verbs with higher token frequencies in the input would be less acceptable in the transitive frame compared to less frequent verbs (Ambridge et al., 2008, 2009; Ambridge and Lieven, 2011).

However, empirical studies on different cases have shown that it is generally impossible to distinguish between ungrammatical expressions and grammatical expressions that simply do not get the opportunity to be attested in the input. I will introduce Bowerman and Croft (2008)'s

and Irani (2019)'s analyses of the causative generalization; see Marcus 1993, Yang 2015 for more discussion.

Firstly, for the preemption approach, since a competing form with same meaning is considered crucial, we can make a distinction between two different types of verbs that are overgeneralized in the causative frame: verbs that have a suppletive lexical causative form (e.g., 'die', with a suppletive lexical form 'kill') and verbs that do not (e.g., 'disappear'). For the former type of verbs, according to the preemption approach, learner will make causative errors before they learn the suppletive causative form of the verb. However, through examination of children production data in CHILDES, Irani (2019) found that children have learned the competing forms such as 'take' (as a competing form for transitive 'go') and 'kill' (as a competing form for transitive 'die') well before they produce over-generalization errors and they use both forms at the same time, showing no evidence for preemption. Moreover, the preemption approach would predict that causative errors with verbs that have more frequent suppletive lexical forms should disappear earlier, since more frequent verbs are expected to be learned earlier. However, this is not the case. For instance, 'take' and 'kill' are considered to be the suppletive causative forms of 'go' and 'die' respectively; 'take' is much more frequent than 'kill' in the input, yet children are observed to make causative errors with 'go' and 'die' at the same time.

Next, for verbs without a suppletive lexical causative form, it is argued that causative errors will be preempted by the 'make'-causative (e.g., 'make something disappear' will preempt 'disappear something') (MacWhinney, 1987). However, the meaning of the 'make'-causative is known to differ from the lexical causative (Fodor, 1970; Ammon, 1980; Bowerman and Croft, 2008). In this regard, Bowerman and Croft (2008) suggested that the 'make'-causative is a weaker cue than the lexical causative. Hence, it was predicted that causative errors with verbs that do not have a suppletive lexical causative form will persist longer due to the lack of a strong competing form. This, however, was not borne out. In Bowerman and Croft (2008)'s analysis of causative errors made by two children, C and E, no significant difference was observed between the rates of causative errors with verbs with and without a suppletive form (Figure 2.1-2.2). Furthermore, children produced the 'make'-causative early, which co-existed with their causative errors, questioning whether the

'make'-causative preempts the overgeneralized causative form at all.

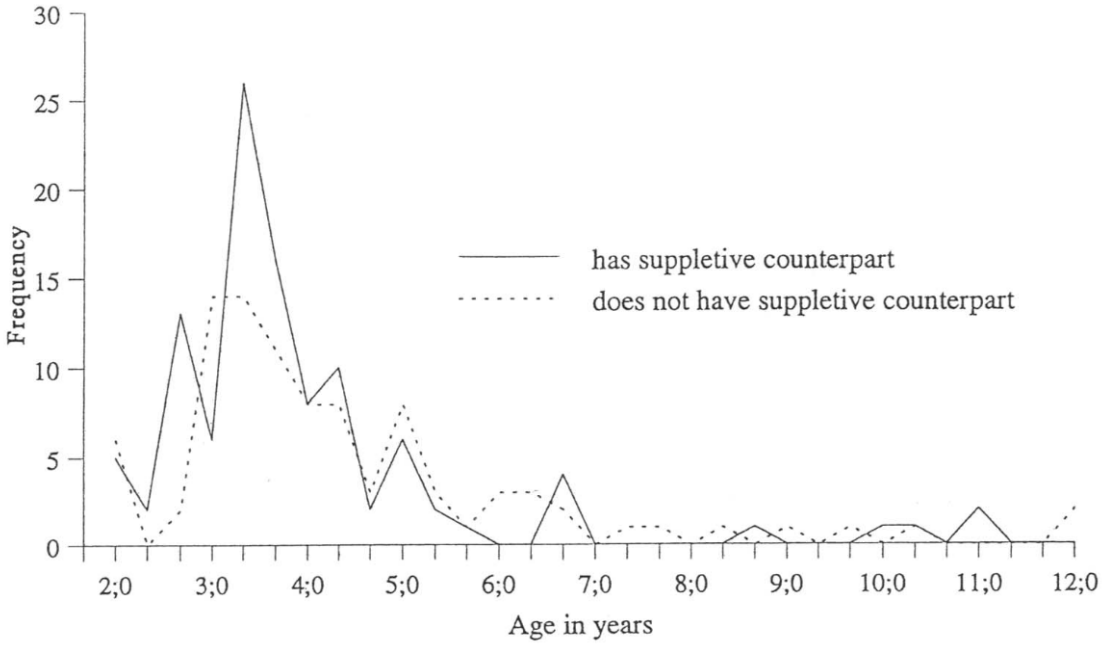

Figure 2.1: Token frequency of novel lexical causatives with and without suppletive counterparts in child C's speech over time (Bowerman and Croft, 2008, p.297).

Finally, according to the entrenchment approach, which relies on statistics of token frequency, children should retreat from causative over-generalizations earlier for intransitive verbs of higher frequency in the input. Again, this was not borne out in examination of children's causative errors in CHILDES and the verb frequencies in child-directed speech. For instance, as pointed out earlier, Irani (2019) observed that the errors in children's spontaneous speech all clustered around the ages of 3-4 although the verbs with which children make over-generalization errors differ drastically in input frequency (Table 1.1, repeated here as Table 2.1).

Verb	Raw frequency in CHILDES
disappear	153
stay	2,662
fall	2,819
go	55,689

Table 2.1: Raw frequency of pure intransitive verbs in child-directed speech (Irani, 2019, p.88)

In summary, I have argued that previous solutions to the acquisition of verb argument structure

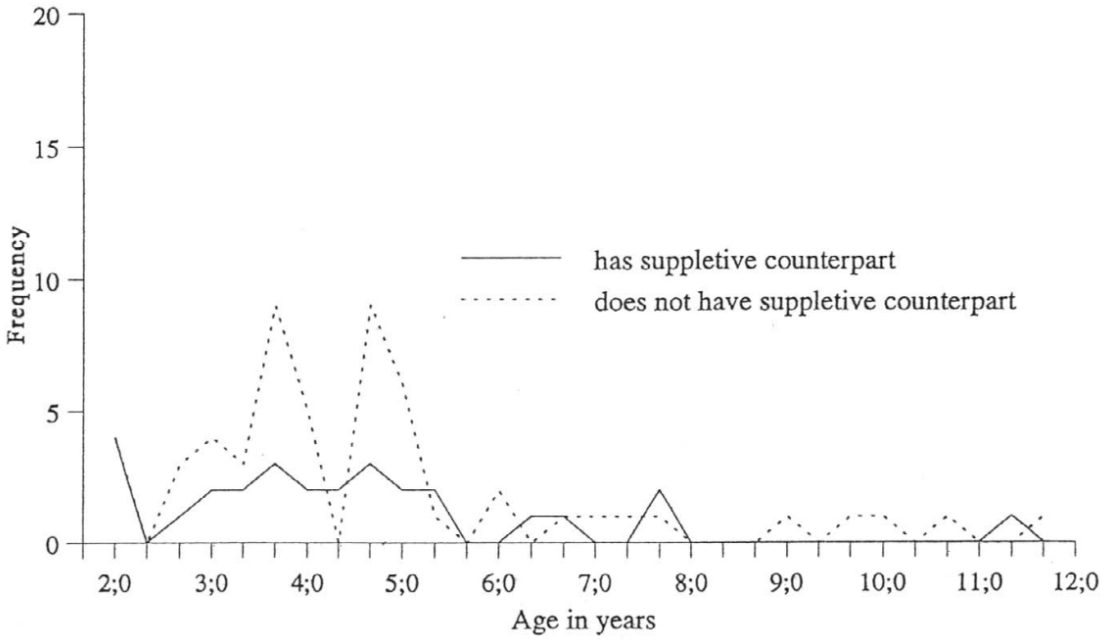

Figure 2.2: Token frequency of novel lexical causatives with and without suppletive counterparts in child E's speech over time (Bowerman and Croft, 2008, p.297).

learning are inadequate: Knowledge of the regular rules should not be innate given the problem of data coverage; and children cannot be conservative learners either because they do generalize to unobserved items; moreover, the learning account cannot reply on negative evidence, which has proven ineffective. In this chapter, I propose a model that automatically discovers regular mappings between syntax and semantics only from positive evidence in simple child-directed speech.

2.2 Model description

In this section I describe a model that automatically and recursively hypothesizes and evaluates productive rules between syntax and semantics based on the TSP. The input data are presented in pairs: it includes the verb, the syntactic frame it appears in, and the semantic feature of the event it depicts.

Before describing the details of the model, I start with the basic assumptions. First, following Jackendoff (1990), I assume that children can construct a conceptual structure to represent an event upon observing an event, and that the semantic structures of verbs are essentially of the

same kind as the conceptual representations by means of which events are represented. And if children hear a sentence while observing the event, they will establish a link between the linguistic description and the conceptual structure of the event. This assumption is shared by virtually all approaches to language acquisition. Note that this is different from the claim that there is innate linking knowledge between syntax and semantics.

I also assume that young children can identify semantic/conceptual primitives such as *act, causation, motion*, and *transfer*, which is also generally agreed in the literature of verb argument structure (e.g., Jackendoff, 1978, 1983, 1987; Landau and Gleitman, 1985; Pinker, 1989; Naigles, 1990) and has been supported by studies with prelinguistic infants. For example, infants by 6 months exhibit nearly all types of motion perception despite the complex nature of motion (e.g., Bertenthal et al., 1987; Aslin and Shea, 1990; Johnson and Mason, 2002; von Hofsten et al., 2007). Another long line of literature has found that very young infants can discriminate events based on causality (e.g., Leslie, 1982, 1984; Saxe and Carey, 2006): For instance, in a habituation looking paradigm, 4 month olds could distinguish between an event where a red brick moved toward, collided with, and launched a green brick (direct causation), and an event where the green brick did not immediately move off upon the collision (no direct causation) (Leslie, 1982). Some have even argued that this ability is available to newborns (e.g., Mascalzoni et al., 2013). Similarly, previous studies have also shown that infants can represent possession and ownership relations early on (see Nancekivell et al., 2019, for a review).

Another assumption is that children have knowledge of the syntactic categories in their language at a young age. There is ample evidence that very young children already know a great deal about the formal grammatical system of their language (e.g., Shi et al., 2006; Shi and Melançon, 2010). In this work I use terms such as NP, VP and PP as a matter of convenience, but this does not mean that I believe these categories must be universal or innate. Instead, studies have shown that categorization is learnable from the input (Mintz et al., 2002; Mintz, 2003; Gerken et al., 2005; Reeder et al., 2013; Schuler et al., 2017; Liang et al., 2022).

Finally, I also assume children can identify some familiar nouns at an early point in language development. This assumption is again widely held, evidenced by findings on children's early

acquisition of nouns (see Gentner and Boroditsky, 2001, for a review). With these assumptions, I do not mean that children must have perfect commanding of such knowledge. Recall that the TSP does not pursue the best grammar but only an adequate one. Therefore, occasional errors with representing the syntactic frame or the semantic primitives can be tolerated. Moreover, I recognize that these different processes such as the acquisition of categories and the meanings of nouns are non-trivial and may interleave with the acquisition of verb argument structure. In this work I am focusing on the learning of verb argument structure but future work should investigate how these processes integrate.

Now I will describe the model. The model prioritizes formal properties during learning: It starts with the most frequent syntactic form and tests whether it can cover all items based on the TSP. If so, then the model will just learn this catch-all form as a productive rule and memorize all items not following the rule. Otherwise, the data will be sub-divided. A crucial property of the model is its greediness, so it will always pursue the most frequent feature first, and only move on the next feature if the most frequent feature proves insufficient. It first tests whether a semantic feature as a prerequisite for the syntactic form can be established using the TSP: in all verbs attested in the most frequent syntactic form, whether a TSP-majority of them have the most frequent semantic feature attested with this syntactic form (e.g., 'double object $\rightarrow$ transfer' in the example in Chapter 1). If so, then the rule that maps from the semantic feature to the syntactic form will be proposed for evaluation. To test this proposed rule, the model will examine all verbs with that semantic feature to see whether the proportion of verbs used in that syntactic frame reaches the productivity threshold predicted by the TSP. If so, this 'semantics $\rightarrow$ syntax' mapping will be learned as a productive rule, and the verbs covered by this rule, including the memorized irregular items, will be removed from the data, and the model will go through the same process again with the remaining verbs; otherwise, it will proceed to test the next most frequent feature. This procedure recursively applies until the most frequent form is productive across the board or no more productive rule can be found. The main steps are summarized in Figure 2.3.

Belth et al. (2021) applied similar algorithms to model the acquisition of morphology. When presented with realistic data from child-directed input, the model successfully learned the rules for

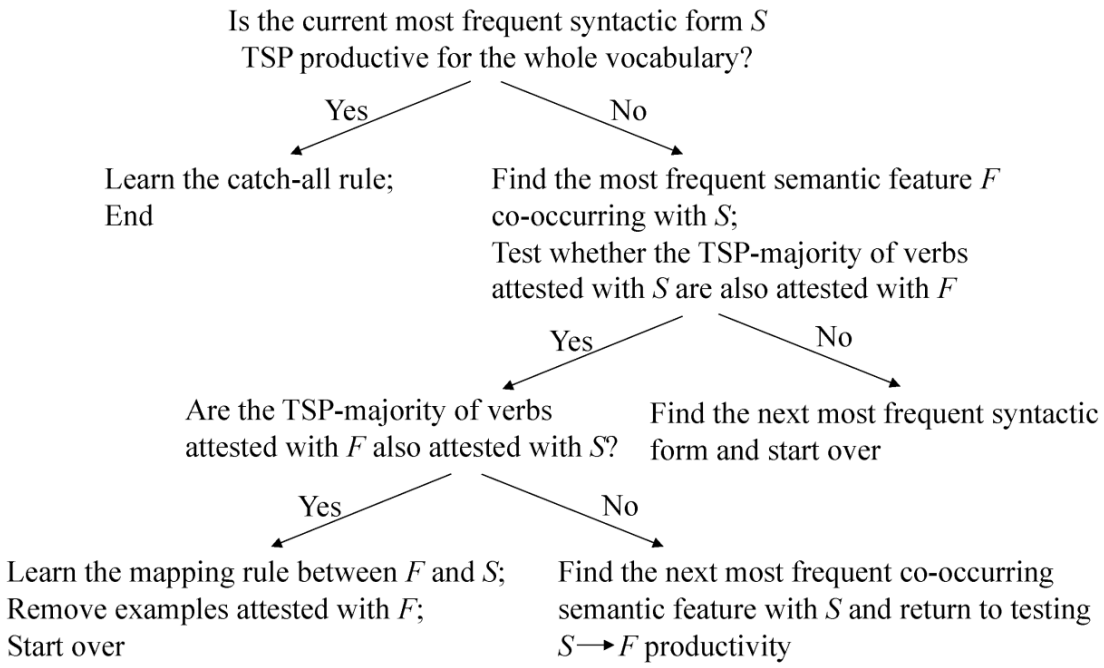

Figure 2.3: A decision tree for the model in the process of rule proposal and evaluation.

noun plurals, past tense, and progressive in English and noun plurals in German. In this work I will show that this algorithm can also capture the acquisition of regular mapping rules in the domains of syntax and semantics.

2.3 Simulation

2.3.1 Data

This section presents a simulation of the model with child-directed English data. The data were extracted from the input to Alex in the Providence corpus (Demuth et al., 2006), a corpus of longitudinal recordings of monolingual English-speaking children from 1-3 years during spontaneous interactions with their parents at home. I used this corpus because it provides video recordings that enable us to code the observable semantic features of the events accompanying the verbs to be learned. For Alex, the corpus contains 51 one-hour video recordings with two-week intervals from 1;4 to 3;5.

To approximate a young language learner's vocabulary, I focused on the 60 most frequent verbs

in early child English (Rowe and Goldin-Meadow, 2009; Carlson et al., 2014) and extracted caregivers' basic-structure sentences containing those verbs. Sentences that had no clearly observable accompanying events were not included. This practice is not just for coding feasibility but also supported by findings from word learning literature that learning instances in natural input differ in their referential clarity and that the highly informative "gems" are particularly helpful for word learning (e.g., Cartmill et al., 2011; Trueswell et al., 2016). This ended up with 1752 sentences. The verbs are provided in (16).[1] For regular verbs, their various forms were automatically included; for irregular verbs, I only included the irregular forms that are highly frequent and therefore should be learned by children at a young age (Rowe and Goldin-Meadow, 2009; Carlson et al., 2014). Note that the irregular forms and their stem still count as one word in the calculation (e.g., 'eat' and 'ate'); the last step is just used to determine which irregular forms should be included in the data.

I then manually coded the syntactic frame that the verb appears in, and the semantic features observable in the accompanying video. The semantic features I used include: *act, causation, motion, transfer, change of state,* and *creation*, which are assumed identifiable to young learners in the literature (e.g., Jackendoff, 1978, 1983, 1987; Landau and Gleitman, 1985; Pinker, 1989; Naigles, 1990). For example, for a scene where the mother said "bite your cookie" as the child did so, the input pair will be "(bite, V NP, (act, causation))" - "act" for an action and "causation" for having an affected object, which is the cookie in this case. The syntactic frames that each verb was attested in are provided in Appendix A. Additionally, the immediate situational context is not a completely reliable cue for verb meaning: There are 302 sentences where the accompanying observable event did not match the event described by the verb, making up for around 20% of all sentences in the data. For example, in Line 553 of Video 020125, the mother said "Let's open our house" when picking up a toy house from the table instead of indeed opening it. Those sentences were not excluded a priori, since children do need to overcome this problem, a topic that has been discussed extensively in word learning literature (see Gleitman and Trueswell, 2020, for a review).

(16) Verbs included in the input data; irregular forms are in parentheses:

[1] Verbs that are provided in Carlson et al. 2014 but have been well known to be abstract and lack reliable correlates in the world were not included, such as mental verbs (e.g., 'think') and perception verbs (e.g., 'see') (e.g., Gillette et al., 1999; Medina et al., 2011).

bite, break (broke), bring, build, call, carry, catch, clean, climb, close, come (came), cry, cut, dance, draw, drink, drive, drop, eat (ate), fall (fell), fly, give (gave), go (went, gone), hide, hit, hold, hug, jump, kick, kiss, knock, move, open, paint, pick, play, press, pull, push, put, reach, read, ride, run, say (said), sing, sit, stand, step, take (took), talk, tell (told), throw, tickle, touch, turn, walk, wash, wipe, write

2.3.2 Results

When trained on all data, the first rule that the model learned was in (17). It was learned through the following process: Given the vocabulary size $N = 59$, the most frequent syntactic frame is 'V NP', with 42 words attested in it. It does not meet the threshold for a catch-all rule since at least 45 are needed, so the vocabulary is split immediately. The most frequent semantic feature co-occurring with 'V NP' is 'act', with 39 words attested with this feature. 'V NP $\rightarrow$ act' is productive based on TSP. However, the other way around is not true, since there are 16 words that also have an 'act' feature but are not attested in the 'V NP' frame, such as unergative verbs 'run', 'sit' and 'walk'. Therefore, the model moves on to the next most frequent semantic feature with 'V NP'. Note that although 'V NP' cannot be a sufficient condition for 'act', it is a necessary condition since 'V NP $\rightarrow$ act' is productive. Therefore, this one-way productive rule can still guide language acquisition. For example, in comprehension, a transitive frame can help children narrow down the possible meaning of a novel verb: they will prefer an 'act' meaning to a 'non-act' meaning. But this rule will not be sufficient for production: given a novel verb with an 'act' meaning, the children will not be able to determine whether it can be productively used in a transitive frame.

(17) If a verb has the semantics of causation, it can appear in the transitive frame.

Now we continue with the process of searching for productive rules. The current most frequent semantic feature with 'V NP' is 'causation'. There are in total 46 words that have a 'causation' meaning, and 37 of them are attested in the 'V NP' frame, meeting the productivity threshold. Thus, 'V NP' and 'causation' are bi-directionally productive. The model thus learned the rule in (17), removed all examples with a 'causation' feature in the 'V NP' frame, and started over. Note that some words can appear in multiple frames and/or with different semantics features. For

example, the verb 'drop' has appeared in the transitive frame with a causation meaning, and also in the intransitive frame without a causation meaning. When we removed entries covered by the established 'causation → V NP' rule, only the first entry of 'drop' was removed; the entry without a causation meaning or transitive frame remained in the data.

Then, the following rules are learned in similar ways as (17); I include the number of verbs attested with the syntactic frame, with the semantic feature, and with both the syntactic frame and the semantic feature in the parenthesis:

(18) If a verb has the semantics of act, it can appear in the intransitive frame. (V: 37, act: 41, overlap: 32)

(19) If a verb has the semantics of change of state, it can appear in the intransitive frame. (V: 7, change of state: 4, overlap: 4)

(20) If a verb has the semantics of creation, it can appear in the double object frame. (V NP NP: 5, creation: 4, overlap: 4)

These rules are all consistent with English speakers' knowledge. The words supporting these rules, though, are in general of much lower type and token frequencies than those supporting rule (17). Therefore, depending on individual children's linguistic experience, they may or may not get sufficient input to pick up these words and consequently discover these rules at a very early stage. This is consistent with children's actual behavior: They readily exhibit knowledge of (17) by two years of age (e.g., Yuan and Fisher, 2009; Arunachalam and Waxman, 2010; Yuan et al., 2012; Arunachalam et al., 2013a; Messenger et al., 2015), whereas the other rules are known to be acquired later than that (e.g., Gropen et al., 1989; Campbell and Tomasello, 2001).

When trained on data where non-matching examples are excluded, the model learned additional meaningful rules including (21-22), which are also well documented rules in English. The reason why these rules were not learned in the previous simulation is that care givers frequently move an object when referring to it in the early input even though the sentence they uttered did not describe an event of motion or transfer, as in the "Let's open our house" mentioned earlier.

(21) If a verb has the semantics of motion, it can take PP's. (V PP: 12, motion: 11, overlap:

11; V NP PP: 7, caused motion: 8, overlap: 7)

(22) If a verb has the semantics of transfer, it can appear in the dative construction. (V NP to NP: 4, transfer: 7, overlap: 4)

In summary, these are all syntax-semantics mapping rules that have been well documented in verb argument structure literature as reviewed in earlier sections. Another well-documented rule that was not learned by the model is the mapping from the semantics of transfer to double object construction. Upon checking the input data, this is accounted for by the fact that this early vocabulary contains a high proportion of verbs with transfer semantics but cannot take double objects, such as 'throw', 'say' and 'talk'. Again, this is consistent with the fact that this rule is acquired later: For example, some 4 year olds still struggle with novel verbs in the double object frame in both comprehension and production (e.g., Conwell and Demuth, 2007; Rowland and Noble, 2010; Rowland et al., 2014; Arunachalam, 2017).

We now focus on rule (17), which is acquired early by children and indeed learned early by our model. Despite the modest vocabulary size and limited input, the learning is robust. Figure 2.4 shows the probability of learning the 'causation → transitive' rule with different sizes of random input sample from 100 runs: For each number on the x-axis, I randomly drew this number of sentences from all input data, trained the model, and repeated this 100 times to model different children who receive different input; the numbers in parenthesis are the mean vocabulary size of the 100 random samples. The y-axis indicates the percentage of "children" who learn the 'causation → transitive' rule. As can be seen, almost half of children already learn the rule with only 200 input sentences; with a vocabulary approaching 60 verbs, essentially all children can learn the productive rule. This developmental pattern is consistent with what we know about language acquisition: There is significant variation in child learners' linguistic experience, and it thus follows naturally that their early vocabularies can differ significantly from each other (e.g., Kidd et al., 2018; Frank et al., 2021); however, once their vocabulary reaches a modest size, they acquire highly consistent grammars at early stages of acquisition (Brown, 1973): "The end result is a high degree of uniformity in both the categorical and variable aspects of language production, where individual variation is reduced below the level of linguistic significance." (Labov, 2012, p.265).

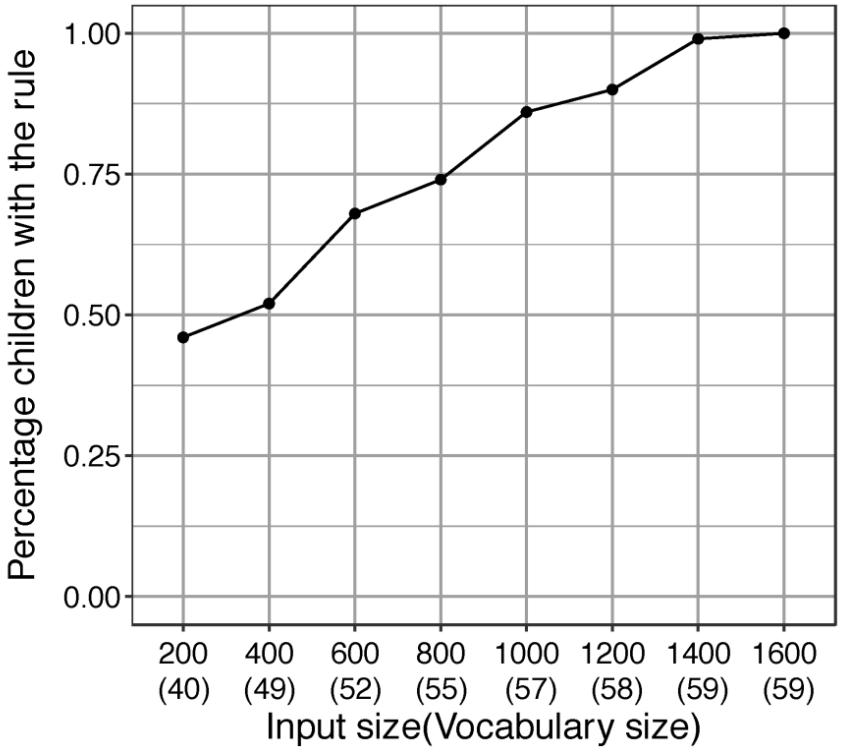

Figure 2.4: Percentage of "children" who have learned the 'causation → transitive' rule with different input and vocabulary sizes.

This model also captures children's well-documented causative over-generalizations, as in (23). Figure 2.5 shows the percentage of the 100 "children" who have learned causative over-generalization as a productive rule, i.e., for verbs that are attested in the intransitive frame with a *change of state* meaning, the TSP majority of them are also attested in the transitive frame. As shown in the figure, more children will make over-generalization errors with increasing input sentences and vocabulary sizes. The vocabulary we work with is representative of children around three years of age, which is indeed within the age range when causative over-generalizations are typically observed (Bowerman, 1982; Pinker, 1989). With increasing vocabularies, it is expected that there will be more verbs not following this rule, and children will thus retreat from this over-generalization (Irani, 2019).

(23) a. Kendall fall that toy. (Kendall, 2;3)

b. I'm gonna ... disappear something under the washrag. (E, 3;7)

c. You ached me. (Rachel, 4;1)

d. He's gonna die you, David. (Hilary, 4+)

(Bowerman, 1982)

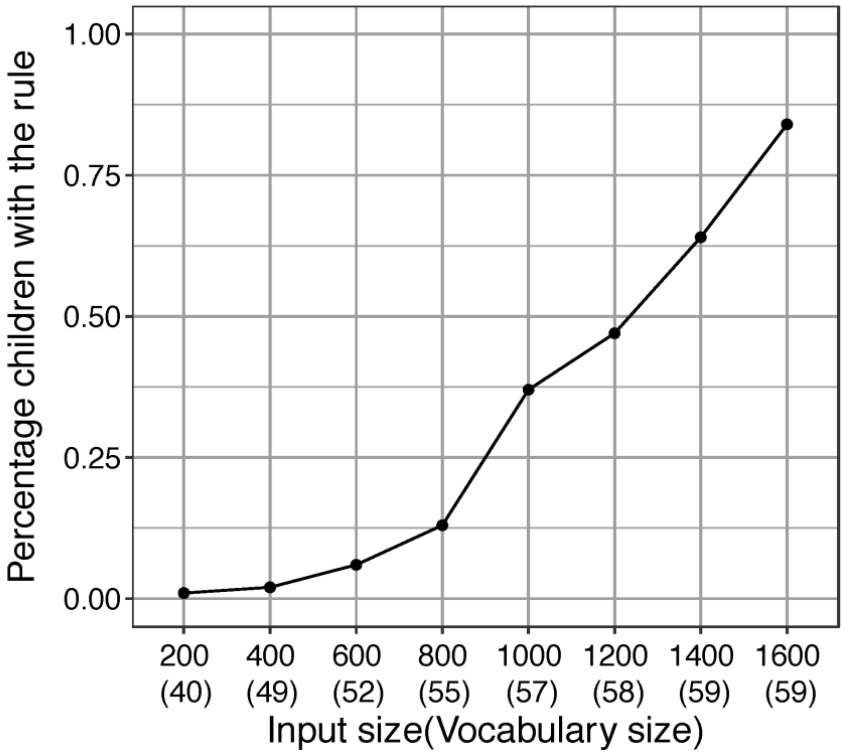

Figure 2.5: Percentage of "children" who have learned the causative over-generalization rule with different input and vocabulary sizes.

2.3.3 Discussion

In summary, this simulation showed that the proposed model can learn many of the well-documented syntax-semantics mapping rules from realistic children input data, reducing the need for universal, innate knowledge of verb argument structure. It also captures children's over-generalization errors. The proposed model is not limited to this one case; future research could investigate its use in other fields of generalization learning. In terms of verb argument structure, it would be desirable to test the model with larger corpora and vocabulary size to see its acquisition of more mapping rules and depict the developmental trajectory, which should also enable us to see the retreat from over-generalizations, and to apply the model in different languages that have different mapping rules

from English. It is also worthwhile to directly compare the current model against other models, which I do in the next section.

2.4 Model comparisons

The Bayesian inference is a learning algorithm that has been widely applied to model language learning in different domains, such as phonetics and phonology (e.g., Feldman et al., 2009; Feldman, 2011; Feldman et al., 2013), word segmentation (e.g., Goldwater, 2006; Goldwater et al., 2009), word-meaning mapping (e.g., Xu and Tenenbaum, 2007; Frank et al., 2009), and syntactic structure (e.g., Perfors et al., 2006, 2011). In this section, I replicate a computational model that learns verb argument structure based on this algorithm (Alishali and Stevenson, 2008, henceforth A&S), and compare its performance to our model. The point of the comparisons, though, is not about the A&S model per se; instead, we are interested in what insights we can obtain from the model comparisons regarding how generalizations are formed. In particular, our proposed model relies on type frequency from positive evidence for generalization; by contrast, the Bayesian inference is a typical algorithm that relies on token frequency and can use indirect negative evidence. Therefore, the comparisons can shed lights on our understanding of the mechanism of generalization.

2.4.1 Introduction

The A&S model extracts argument structure *frames* (i.e., the syntax-semantics pairing of a verb) from the input, and then presents the frame to unsupervised Bayesian clustering, which groups the new frame together with an existing group of frames (which the authors call a *construction*) that probabilistically has the most similar properties to the new frame. Ultimately, the model forms different constructions, which are essentially a probabilistic association between syntactic frames and semantic features. Regular mappings rules should be reflected by stronger probabilistic associations. Figure 2.4 illustrates a portion of the acquired lexicon, with the lexical entries of verbs containing frames and their links to constructions.

Specifically, given a new frame F, which consists of a verb and its syntax-semantics pairing, the model groups it into construction k by finding the k with the maximum probability given F:

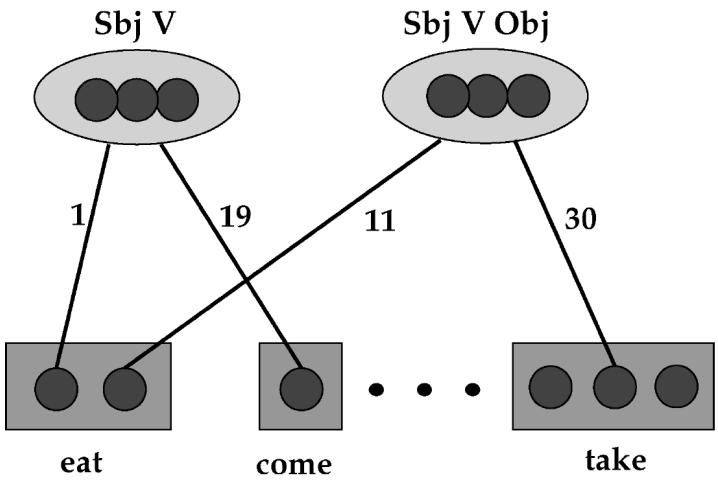

Figure 2.6: A portion of the lexicon showing an intransitive construction, with the predominant syntactic pattern SV; and a transitive construction, with the predominant syntactic pattern SVO. Numbers on links represent the observed frequency of the verb in a frame that participates in that construction. For example, 'eat' has been seen once in an intransitive construction and 11 times in a transitive construction.

(24)

$$\textbf{Best construction}(F) = \underset{k}{\operatorname{argmax}}\, P(k|F)$$

In (24), k ranges over the indexes of all constructions; a new construction is represented by an index of 0. Using Bayes's rule, the conditional probability of a construction given a frame is calculated as follows, where *P(F)* is dropped since it is constant for all k.

(25)

$$P(k|F) = \frac{P(k)P(F|k)}{P(F)} \propto P(k)P(F|k)$$

Following a used-based approach, A&S assumed that the prior probability of a construction is proportional to the total token frequency of frames associated with the construction: A new frame will more likely come from a construction that has been observed more frequently. Therefore, $P(k)$ is given by the following, where n is the total number of observed frames, and n_k is the number of frames participating in construction k:

(26)

$$P(k) = \frac{n_k}{n+1}$$

For a new construction ($k = 0$), its prior probability is as below, with the idea being that as children receive increasing amount of exposure, it will be less likely that they will observe a totally new construction:

(27)
$$P(0) = \frac{1}{n+1}$$

Next, for the conditional probability of a frame given a construction, for calculation feasibility, A&S assumed that the component features are independent. Therefore, the conditional probability is expressed as the product of the conditional probabilities of its component features:

(28)
$$P(F|k) = \prod_{i \in FrameFreatures} P_i(j|k)$$

where j is the value of the i-th feature of the frame F. In this model, the features include the verb, the syntactic frame, the number of arguments, which can be easily extracted from the syntactic frame, and the semantic primitives. Therefore, $P_i(j|k)$ is the probability of having value j on feature i in construction k. This probability is estimated using a smoothed maximum likelihood formulation:

(29)
$$P_i(j|k) = \frac{\mathbf{count}_i^k(j) + \lambda}{n_k + \lambda\alpha_i}$$

where n_k is the number of frames participating in construction k, and $\mathbf{count}_i^k(j)$ is the number of these frames that have value j for feature i. The smoothing parameter α_i is set to an estimate of the number of possible values that feature i can take on. This means that for a new construction, where $\mathbf{count}_i^0(j)$ and n_0 are both 0, all feature values have the same estimated conditional probability $1/\alpha_i$. Parameter λ is set to a small constant so that constructions with no members that have value j for feature i will have a low but non-zero probability. Following A&S, I set λ to 10^{-4}.

To examine the knowledge acquired by the model and compare it to human's knowledge, A&S formulated language comprehension and production as a prediction process of finding the most probable values for missing features given the available features in a frame and the association

probabilities learned during the training phase: In comprehension, the model predicts the most probable semantic features given the syntactic frames; in production, it predicts the most probable syntactic frames given the semantics. In specific, the prediction is made as follows:

(30)

$$\textbf{Best Value}_i(F) = \underset{j}{\text{argmax}}\, P_i(j|F) = \underset{j}{\text{argmax}} \sum_k P_i(j|k)P(k|F)$$

where F is a partial frame with missing features, i is the missing feature missing in that frame, j ranges over possible values for feature i, and k ranges over all constructions. The first component of the sum, $P_i(j|k)$, is calculated in the same way as in the learning phase, (29). $P(k|F)$ is calculated by using Bayes's rule and dropping the constant $P(F)$:

(31)

$$P(k|F) \propto P(k)P(F|k)$$

Then $P(k)$ and $P(F|k)$ are calculated as in the learning phase ((26) and (28) respectively).

Note that in addition to syntactic and semantic features, the verb itself is also regarded as a feature in the model. Therefore, the prediction will depend on both verb-specific knowledge and general knowledge of the constructions. I will later return to discussions on how this algorithm influences the learning outcomes.

Overall, a fundamental difference between the Bayesian model (and most usage-based learning accounts) and our proposed model is that the former relies on token frequency, whereas in the latter model, only type frequencies are used in learning and evaluation of generalizations. Of course I do not deny the role of token frequency in language acquisition. For instance, the TSP has token frequency in its derivation (Yang, 2016); and children must hear a word frequently enough to establish it in their vocabulary (e.g., Yu and Smith, 2007; Trueswell et al., 2013). However, my argument here is that token frequency itself does not participate in the process of generalization. A high type frequency has long been associated with productivity (e.g., Plunkett and Marchman, 1991; Bybee, 1995; Baayen and Renouf, 1996; Pierrehumbert, 2003): It is conceivable that a form attested with one highly frequent word will not be convincing evidence that the form can apply to *all* words. And this idea has been supported by realistic data. For instance, it has been know

that a consistent trajectory in morphology acquisition for English speaking children is that the regular noun plural '-s' is acquired earlier than the regular past tense '-d' (Brown, 1973). This can be captured by their type frequency but not token frequency: According to Yang (2016), noun plurals appeared 69,246 times in all child-directed North American English corpora from CHILDES at the time (five million words), which is way less frequent than past tense (89,030). However, when we look at type frequency, irregular nouns are much less frequent than irregular verbs, in particular for the most frequent words. For example, only four nouns out of the twenty most frequent plural nouns in the child-directed corpora examined by Yang (2016) are irregular; whereas for the twenty most frequent past tense verbs, only three are regular. Empirical experiments have provided additional evidence for the crucial role of type frequency in generalization. In the study by Schuler et al. (2016), 7-8-year-old English speaking children learned a noun plural rule (adding 'ka') in an artificial language. There were 9 pseudo-nouns in the learning phase. In the test, the children generalized the rule when 5 nouns obeyed the rule and did not generalize when 3 nouns obeyed the rule, although in the second language the 'ka' marker was still the most frequent marker in the learning phase in terms of token frequency (58% of tokens). In another study, Koulaguina and Shi (2019) exposed 14-month-old French-learning infants to Russian sentences, some of which followed a word-order shift rule. They found that the infants generalized the rule when there were 8 out of 10 sentences in the learning phase following the rule, although the token frequency of rule-following sentences and not-rule-following sentences were kept the same (each of the 8 rule-following sentences were repeated 4 times and each of the 2 not-rule-following sentences were repeated 16 times). By contrast, the infants did not generalize the rule when 8 out of 16 sentences followed the rule and all sentences were repeated 4 times. Therefore, given the difference status of type vs. token frequency, I predict that this difference between the two models will lead to differences in the knowledge they acquire. In particular, the A&S model would be more vulnerable to patterns that are of high token frequency but not necessarily generalizable.

Another difference between the models, which is related to the point on token vs. type frequency, is that the A&S can use indirect negative evidence: According to its algorithms, the more frequent a construction is heard in the input, the more likely it will be used in the future. Thus, high token

frequency of one form will be taken as evidence against alternative forms. By comparison, our proposed model only relies on positive evidence - generalizations are only found when there is a sufficiently high type frequency in the input. We will return to discussions on how these differences impact the learning outcomes.

2.4.2 Results

I trained the A&S model with the same input data to Alex. To examine the acquired knowledge, following A&S, I tested the model's usage of novel words by presenting the model with an input pair of a pseudo-word where either the syntactic frame or the semantic features were missing, and the model would compute the probabilities of different possible values for the missing element.

First, I presented to the model input pairs in different syntactic frames where the semantic features were omitted. Table 2.2 shows the semantic features of the highest probabilities associated with each of the major syntactic frames when the model was trained on either all data or only clean data where non-matching examples were excluded. The probability is in the parentheses. As the table shows, the model did learn some associations that are consistent with mapping rules in English, such as the strong associations between 'V NP' and 'causation', between 'V' and 'change of state', between the double object construction/the dative construction and 'transfer', and between the constructions that take PP's with 'motion'. A difference between this model and the model we proposed is that the former learns probabilistic associations instead of categorical rules that are necessary and sufficient. As a result, all these frames have a strong association with 'act' and 'causation' although these features may not be sufficient to distinguish one frame from another. To further explore what this means, I tested the model's 'comprehension' and 'production' in the same way as A&S. I focused on the transitive and intransitive constructions since these are robustly learned by children at a very early stage.

I first replicated the syntactic bootstrapping test to examine 'comprehension'. This was conducted by having the model compute the probabilities for four different test input pairs in Table 2.3. Recall that in a typical syntactic bootstrapping experiment, children would hear a novel verb in either a transitive or intransitive frame while watching a causative scene and a non-causative

Syntactic frame	All data	Clean data
V NP	act (2.9×10^{-7}) causation (1.7×10^{-7}) communication (1.6×10^{-7})	act (3.3×10^{-7}) causation (2.5×10^{-7}) communication (2.0×10^{-7})
V	act (2.8×10^{-7}) causation (1.2×10^{-7}) change of state (7.3×10^{-8})	act (3.3×10^{-7}) causation (1.4×10^{-7}) change of state (8.8×10^{-8})
V NP NP	act (1.0×10^{-7}) causation (5.5×10^{-8}) transfer (4.7×10^{-8})	act (1.1×10^{-7}) transfer (9.5×10^{-8}) causation (5.8×10^{-8}
V NP to NP	act (6.0×10^{-8}) causation (5.4×10^{-8}) transfer (5.0×10^{-8})	transfer (1.5×10^{-7}) act (8.2×10^{-8}) causation (8.2×10^{-8})
V NP PP	act (5.8×10^{-8}) causation (5.8×10^{-8}) caused motion (5.7×10^{-8})	act (9.0×10^{-8}) causation (8.9×10^{-8}) caused motion (8.8×10^{-8})
V PP	act (1.7×10^{-7}) motion (9.2×10^{-8}) causation (6.0×10^{-8})	act (1.5×10^{-7}) motion (1.4×10^{-7}) causation (7.0×10^{-8})

Table 2.2: The probability of semantic features in different syntactic frames.

scene. The causative scene that children see in the experiment was represented by the features 'act & causation', while the non-causative scene was represented by 'act'. If the model has learned the mapping rules, then given the transitive frame, the probability of 'act & causation' should be higher than that of 'act'; and given the intransitive frame, the probability of 'act' should be higher than that of 'act & causation'. However, as Table 2.3 shows, while this is borne out for the transitive frame, the probabilities of 'act' and 'act & causation' are essentially the same given the intransitive frame. Upon examination of the input data, this result is explained by the high token frequency of words used with an intransitive frame in a causative scene (e.g., "I eat."). A&S reported higher probabilities for both kinds of matched pairs, but their training data contained much fewer words and sentences (13 verbs, 500 sentences); and more importantly, they used artificially generated data instead of data that were coded from videos. Therefore, our training data are likely more representative of children's actual input. Our model, by contrast, did not learn any productive rule between the intransitive frame and the causation meaning because the type frequency of such optional transitive verbs is not high: Among 37 words with an 'act & causation' meaning, only less than half of them (17) were attested in the intransitive frame.

Test pair	Probability (all data)	Probability (clean data)
'V NP' – 'act & causation' (matched)	5.9×10^{-8}	7.1×10^{-8}
'V NP' – 'act' (unmatched)	1.9×10^{-9}	2.3×10^{-9}
'V' – 'act' (matched)	5.9×10^{-8}	7.1×10^{-8}
'V' – 'act & causation' (unmatched)	5.9×10^{-8}	7.1×10^{-8}

Table 2.3: Probabilities for matched and unmatched utterance-scene pairs.

As in A&S, I also conducted the 'syntactic bootstrapping' test with varying amounts of learning. Table 2.4 presents the results; the column name indicates the number of training sentences randomly drawn from all input data; the values are averaged over 10 simulations. As the table shows, while the exact values differ from those in Table 2.3, the pattern that the model showed no preference between a causative event and a non-causative event given an intransitive frame is stable.

Test pair	400	800	1200	1600
'V NP' – 'act & causation' (matched)	2.6×10^{-7}	1.3×10^{-7}	8.6×10^{-8}	6.5×10^{-8}
'V NP' – 'act' (unmatched)	8.3×10^{-9}	4.2×10^{-9}	2.8×10^{-9}	2.1×10^{-9}
'V' – 'act' (matched)	2.6×10^{-7}	1.3×10^{-7}	8.6×10^{-8}	6.5×10^{-8}
'V' – 'act & causation' (unmatched)	2.6×10^{-7}	1.3×10^{-7}	8.6×10^{-8}	6.5×10^{-8}

Table 2.4: Probabilities for matched and unmatched utterance-scene pairs with varying amounts of learning.

This pattern which is at odds with human learners' behavior is also found in the A&S model's 'production', which was reflected by the probabilities of different syntactic frames given the semantic features of a novel word. Table 2.5 shows all the syntactic frames with a probability over 1×10^{-8} given 'act' (a non-causative scene) or 'act & causation' (a causative scene). Given just 'act', as expected, the intransitive frame is dominant; however, given 'act & causation', the probabilities of 'V' and 'V NP' are essentially the same, also due to the high token frequency of optional transitive verbs. This would predict that given a novel verb in a causative scene, children would productively use it in an intransitive frame, which is not what studies have found (e.g., Tomasello et al., 1997). Again, this result is stable with varying amounts of learning (Table 2.6, also averaged across 10 simulations; only displaying the syntactic frames with the highest probability).

Finally, A&S proposed another account for causative over-generalization: Unaccusative verbs may be used in scenes where there is a causative agent, e.g., for 'fall', there may be someone who caused the object to fall. Therefore, given the acquired knowledge that verbs of causation can be

Semantic features	All data	Clean data
act	V (5.9×10^{-8})	V (7.1×10^{-8})
act & causation	V (5.9×10^{-8}) V NP (5.9×10^{-8}) V NP to NP (1.0×10^{-8})	V NP (7.1×10^{-8}) V (7.1×10^{-8})

Table 2.5: The probability of syntactic frames given different semantic features.

Semantic features	400	800	1200	1600
act	V (2.7×10^{-7})	V (1.4×10^{-7})	V (8.6×10^{-8})	V (6.5×10^{-8})
act & causation	V (2.7×10^{-7}) V NP (2.6×10^{-7})	V (1.4×10^{-7}) V NP (1.3×10^{-7})	V NP (8.6×10^{-8}) V (8.6×10^{-8})	V (6.6×10^{-8}) V NP (6.5×10^{-8})

Table 2.6: The probability of syntactic frames given different semantic features with varying amounts of learning.

used in the transitive frame, children would use 'fall' in 'V NP'; and they would retreat when they get more input for the word, because as the token frequency increases, knowledge of the individual word will have a stronger influence based on the model's algorithm. However, an issue with this account is that there are unaccusative verbs that do not necessarily have a causative agent, such as 'disappear' or 'die', but children have been observed to overgeneralize them as well (e.g., Bowerman, 1982). Moreover, the A&S model would predict more frequent verbs to retreat earlier, which is not attested. For example, Ross from the MacWhinney corpus has been found to overgeneralize verb such as 'go' 'fall' and 'disappear' around the ages 3-4 although their frequency in the input differs drastically (Irani, 2019). This observation is consistent with our model, since once a productive rule is learned, it applies to all words regardless of their token frequency.

2.4.3 Discussion

In summary, in this section I compared the A&S model against our proposed model by training them with the same realistic children input data. It is found that the A&S model learned a very strong association between the intransitive frame and the causation meaning, which is of high token frequency in the input but not a productive rule in English. Moreover, the predictions that the A&S model makes for the trajectory of over-generalization errors are not borne out in child language. Overall, the learning outcomes of the proposed model are more accurate and in line with children's developmental patterns than the A&S model.

As discussed earlier, the point of the model comparisons is not about this specific A&S model per se, but about the insights that we can obtain regarding the mechanism of learning generalizations. The A&S model has properties which fundamentally differ from our proposed model but have been widely applied in language learning models from the machine learning framework: It learns probabilistic associations instead of categorical rules; it relies on token frequency; and it makes use of indirect negative evidence. These properties are not only unsupported by empirical findings in language and language acquisition (e.g., children form a categorical distinction between productive and unproductive processes (e.g., Berko, 1958; Xu and Pinker, 1995); children rely on type frequency to form generalizations (e.g., Schuler et al., 2016; Koulaguina and Shi, 2019); indirect negative evidence cannot distinguish between ungrammatical expressions and grammatical expressions that just happen to be unattested (e.g., Marcus, 1993; Yang, 2016; Irani, 2019)), but more crucially, as we have seen, they lead to inaccurate learning outcomes. Therefore, the results highlight the necessity to develop language learning models that are empirically plausible and testable.

These properties of the A&S model that we discussed above - reliance on token frequency and indirect negative evidence, and learning probabilistic associations - are often related to each other and indeed they usually co-exist in virtually all computational language models based on the Bayesian inference. However, future research could tease them apart to examine their individual influences on the learning outcome. For instance, one could adjust the algorithm of the A&S model so that it calculates type frequency rather than token frequency while keeping all the other aspects unchanged. It is predicted that the acquired knowledge of the adjusted A&S model will still be inaccurate since all of these features are in odds with empirical facts of the language learning mechanism.

2.5 General discussion

In this chapter, I propose a model that learns regular rules for verb argument structure. By applying the TSP, which offers a precise productivity threshold, I demonstrate that the well-documented syntax-semantics mapping rules are learnable from language specific modest-sized input data, as long as children are able to identify conceptual/semantic primitives, track formal cues in different

categories, and recognize some familiar nouns – which we have evidence that young children can do given independent lines of prior work (e.g., Leslie, 1982; Jackendoff, 1990; Gentner and Boroditsky, 2001; Shi and Melançon, 2010). By contrast, a computational model which learns verb argument structure based on Bayesian inference (A&S) is sensitive to statistical trends that may not be necessarily consistent with interpretable productive rules, and it makes predictions that do not match human behavior.

Overall, the findings in this chapter reduce the need for an innate and universal theory of syntax and semantics mappings. I would like to make it clear that this does not replace syntactic bootstrapping: there is ample evidence that young children have productive and robust knowledge of syntax-semantics mapping and that they can use the knowledge to guide the learning of novel words. Instead, my goal is to explain where such knowledge comes from. The innateness approach would be difficult to maintain given the level of idiosyncrasies within and across languages (e.g., Pinker, 1989; Levin and Rappaport Hovav, 1995). The learning of form-feature mapping rules has actually been extensively studied in other linguistic domains such as morphology (e.g., Berko, 1958; Rumelhart and McClelland, 1986; Marcus et al., 1992; Bybee, 1995; Xu and Pinker, 1995; Baayen and Renouf, 1996); in this work I argue that mapping knowledge in the syntax-semantics domain also should not and need not be innate. On the other hand, the model comparisons in this chapter also demonstrate that the learning algorithms should make psychological commitments and be subject to empirical examinations against what we know about child language development.

This chapter opens up many possible future directions. First, in future research, it would be desirable to run the model with increased sample size and vocabulary size. It is predicted that the model will be able to learn more mapping rules in English and exhibit the U-shape developmental trajectory of making over-generalizations and then retreat to adult grammar. One can use some existing annotated corpora as the training data (e.g., Pearl and Sprouse, 2013). There might be occasional errors in the annotation but that would be fine and even welcome, since children may also make errors during language learning and the proposed model should be able to tolerate the errors as long as enough data can be covered. It would also be interesting to test the model in languages that have different verb argument structure rules than English. For instance, there are languages that

allow extensive argument omission, such as Mandarin and Korean. Experiments have found that young children learning these languages can also use the number of arguments to infer the meaning of novel verbs even though the arguments are not always present in caregiver speech (e.g., Lee and Naigles, 2008); on the other hand, these children seem to rely less on the realization of arguments to learn novel verbs in such languages (e.g., Arunachalam et al., 2013b; Fisher et al., 2020). Thus, it would be worthwhile to examine how the model captures such language-specific properties. In addition to syntax-semantics mappings, the model can also be adapted to model generalization learning in other domains, such as syntax-syntax mappings for construction alternations, and other generalization phenomena in syntax, morphology, or phonology.

Chapter 3

DISTRIBUTIONAL LEARNING OF RECURSIVE STRUCTURES

3.1 Introduction

This chapter examines another case of syntactic generalization, where the relation between syntax and semantics is less systematic, but the vocabulary needs to be partitioned into subsets for a productive rule to hold: recursive structures. While the ability to form recursive structures is considered innate and universal (Yang, 2013; Berwick and Chomsky, 2017), it has been well observed that languages differ regarding what structures allow recursive embedding (e.g., Roeper, 2011). Therefore, the ability of recursion cannot be freely applied. Instead, it must be learned from language-specific experience which structures allow recursive embedding and under what conditions.

In this chapter, I will argue that the acquisition of recursive structures is also a problem of learning generalizations: It is not learning whether a structure per se allows recursive embedding, but learning whether there are words that allow recursive embedding for a given structure. Therefore, just as the problem of learning verb argument structure, the acquisition of recursive structures also requires a learning mechanism that identifies productive regularities from input data. In this chapter, I will provide a solution that connects those questions. I will present a proposal for how recursive structures should be formulated, which leads to a theory of how they are acquired by

children. It will be demonstrated that the recursivity of a structure can be established as a quantitative generalization through distributional learning, from which the necessary semantic properties can be readily identified.

3.1.1 The acquisition challenge

Many linguists and cognitive scientists agree that the ability for recursion is a crucial part of the language faculty and is universal across languages (Yang, 2013; Berwick and Chomsky, 2017). However, languages differ regarding the domains of recursive embedding, which must be learned from language specific experience. For instance, consider the following examples of English possessives: While both the *s*-possessive and the *of*-possessive can express the general 'possession' meaning, the *s*-possessive but not the *of*-possessive can be infinitely embedded when expressing ownership.

(32) a. the man's neighbor's book

b. *the book of the neighbor of the man

But this does not mean the *of*-possessive cannot be recursively embedded. As in the examples below, it does allow embedding for a restricted class of words:

(33) The top of the tip of his hat (attested in CHILDES input)
The end of the story of the kitty (attested in CHILDES input)
The color of the cover of the book
The middle of the third inning of the deciding game
The son of the President of the Union of Retired Professors

It is well known that the meanings expressed by these structures are quite varied; for example, as in all the sentences above, the *of*-possession is largely limited to inalienable possession, where the possessor and possessee have a permanent association such as kinship or part-whole relationship, whereas the *s*-possessive can be used to describe alienable possession (see Rosenbach, 2014, for extensive discussion). So how can English-speaking children learn these rules for recursive embedding?

Such knowledge cannot be universal or innate given the cross-linguistic differences. For instance, different from the English examples above, the *s*-possessive in German can only be used with a narrow set of words such as proper names and some kinship terms and usually cannot embed at all (Weiß, 2008; Pérez-Leroux et al., 2022), while the possessive with the preposition *von* 'of' is freely recursive, (34); and in Mandarin Chinese, noun embedding takes place freely when the possessive marker *de* is present but is much more restricted without *de*, (35) (Li and Thomson, 1981).

(34) a. das Buch von dem Nachbarn von dem Mann
the book of the neighbor of the man
'the book of the neighbor of the man'

b. Marias/Vaters/*Manns Buch
Maria's/father's/*man's book
'Maria's/father's/man's book'

c. *Peters Nachbars Buch
Peter's neighbor's book
'Peter's neighbor's book'

(35) a. na ren de linju de shu
the man GEN neighbor GEN book
'the man's neighbor's book'

b. *na linju shu
the neighbor book
'the neighbor's book'

c. *na ren shu
the man book
'the man's book'

Beyond the case of possessives, Table 3.1 shows some other well-documented cross-linguistic differences in rules for recursive embedding (e.g., Bauer, 1978; Everett, 2005; Haspelmath, 2016). Overall, although the ability of recursion may be universal, a structure cannot be recursive by default. Therefore, we need a theory for children's psychological procedure of discovering the specific rules for recursive embedding in their language.

Structure	Language	Productive recursion?
Compounds	Germanic languages	Yes
	Romance languages	No
Adjectives	English	Prenominal
	French	Postnominal
Serial verbs	English	No
	Bantu, Mandarin	Yes
Clauses	Germanic, Romance	Yes
	Pirahã	Disputed

Table 3.1: Cross-linguistic differences in recursive embedding.

3.1.2 Acquisition of recursive structures

Early work on children's acquisition of recursive structures has reported difficulty with recursion. Corpus and experimental studies on a range of structures including adjectives, possessives, nominal compounds, prepositional phrases, and sentence complements across languages have found that recursively embedded structures are challenging for young children in both comprehension and production, which was interpreted as evidence that children's early grammar lack recursion (e.g., Matthei, 1982; Gentile, 2003; Roeper and Snyder, 2005; Roeper, 2007, 2011; Hollebrandse et al., 2008; Fujimura, 2010; Hiraga, 2010; Limbach and Adone, 2010; Pérez-Leroux et al., 2012). However, debates exist as to whether the observed difficulty is due to immature grammar or processing limitations, since even adults experienced difficulties when performing tasks that involved recursive embedding in previous studies (e.g., Limbach and Adone, 2010; Pérez-Leroux et al., 2012, 2018). More recent studies which provided a felicitous context for recursion have found higher rates of success among young children (Pérez-Leroux and Roberge, 2018; Giblin et al., 2019; Robinson, 2022). For instance, Giblin et al. (2019) used a Truth Value Judgment Task and designed the scenario to maximize the felicitous use of two-level recursive possessives: For example, in a trial there were two characters that each had a frog, and both characters and both frogs each had a cookie. One of the frogs had their cookie stolen. A blinded-folded puppet would provide a false statement of what happened using one-level possessive (e.g., "The pirate's cookie was stolen"), and participants needed to decide whether to accept or reject the description, and if they rejected the puppet's description they were asked to justify their rejection (e.g., "No, the pirate's frog's

cookie was stolen"). It was found that 4-year-old English- and Mandarin-speaking children could successfully comprehend and produce sentences with recursive possessive phrases. Robinson (2022) used similar methods to study the acquisition of recursive prepositional phrases and also found much higher rates of success than previous studies. Hall and Pérez-Leroux (2022) found three- to five-year-old children's interpretation of recursively embedded NPs was highly accurate. Therefore, evidence suggests that children do have the grammar for recursion at an early age although they may be constrained by processing limitations.

Importantly, children can distinguish between the productive and non-productive structures for recursive embedding at an early age. Pérez-Leroux and Roberge (2018)'s study on English recursive NPs found that *of*, which is less productive than *-s*, emerged later in children's production and was never overgeneralized. Pérez-Leroux et al. (2022) examined German-speaking children's acquisition of recursive possessives. There are several structural options for possessives in German, including genitive case, possessive *-s*, relative clauses, and *von*-prepositional phrases, of which the Saxon possessive *-s* in general cannot embed. In an elicited production task, they found that similar to adults, five-year-old German-speaking children used double *von* most commonly; moreover, they never embedded *-s* possessives. In another study on Spanish NP recursion, Pérez Leroux et al. (2018) observed that both adults and four- to six-year-old children preferred to use the productive linker *de*, and rarely used PPs, which is largely limited to comitatives for inalienable relations in Spanish.

In summary, although recursive structures are complex and pose a high processing load, children do acquire them very early and can distinguish between the productive structures and the structures that are constrained. A crucial question, then, is how they acquire such knowledge from the input.

3.1.3 Previous approaches

The innateness approach contends that since recursion is a core design feature of the language faculty, it should be an innate linguistic property and emerge spontaneously in language development (Crain, 1991; Berwick and Chomsky, 2017). However, even though the ability of recursion may be innately available, as we have seen in this chapter, languages differ regarding the exact rules

for recursive embedding, so children still must learn from language specific experience in which domains they can apply the ability of recursion.

An existing proposal in the literature for the learnability problem is that explicit evidence of recursive embedding is necessary for the acquisition of recursive structures (Roeper and Snyder, 2004; Roeper, 2007, 2011): For instance, German-speaking children can avoid mis-analyzing the *s*-possessive structure as recursive because they never hear self-embedding of the structure, whereas English-speaking children will hear examples from the input where the *s*-possessive is recursively embedded. We argue that this approach cannot be adequate. First, multi-level embedding, like all complex linguistic structures (Jelinek, 1998; Yang, 2013), is rare and may not reliably appear in the input. For example, in a corpus of 13.5 million words of child-directed English from the CHILDES database (MacWhinney, 2000), which roughly corresponds to 2 years of language input (Swingley, 2009), we found approximately 150 doubly-embedded *s*-possessive structures, the majority being of the form "Proper-name's Kinship-term's name" as in "Cindy's child's name" and "Bruce's mommy's name"; not a single triple-level (or deeper) embedding example is found. Previous studies across structures and languages have all confirmed the rarity of recursively embedded examples in children's input (e.g., Pérez-Leroux et al., 2018; Giblin et al., 2019). Second, and more important, there is no principled reason why the presence of N-level embedding would ensure even *(N+1)*-level embedding, never mind infinite embedding. Infinite embedding, it would seem, must be acquired on "shallow" data, likely only level-one data, the only kind that is abundantly available in the input data. To do so requires an alternative conceptualization of recursion, which I will introduce in the next section.

3.2 Proposal

The proposal consists of two logically independent components. I first put forward a conceptualization of recursion, which reduces embedding of infinite depth to a structural property that holds in strictly one-level data. I then provide an account of how this distributional property is acquired, as a special instance of productivity in language (Yang, 2005, 2016).

3.2.1 Recursion as structural substitutability

Recursion in language has traditionally been construed as the self-embedding of a linguistic object (see Huijbregts, 2019, for a review). For example, the possessive structure 'the man's neighbor' can be expressed by a rewrite rule 'NP → NP's NP', where the first NP self-embeds another NP to derive 'the man's neighbor's book'. Once a rule such as 'NP → NP's NP' is available, infinite embedding immediately follows. For the learner, the problem is how recursive rules become part of their grammar.

As we discussed earlier, the mere attestation of multilevel embedding, however deep, cannot be sufficient to support recursion. Moreover, this conceptualization does not encode lexical restrictions on recursion. As we have reviewed, recursive embedding in natural language is not a binary choice - it is not that a structure can either freely recurse or cannot recurse at all. Instead, there are many structures like the English *of*-possessive which is recursive for a restricted class of words, and this traditional conceptualization cannot answer how the rules for such structures are represented and learned.

In this section I propose a new conceptualization of recursion: I argue that recursion does not derive from the self-embedding of a linguistic object (e.g., NP) but from a property, dubbed *structural substitutability*, that concerns *two* positions in the formal linguistic structure.

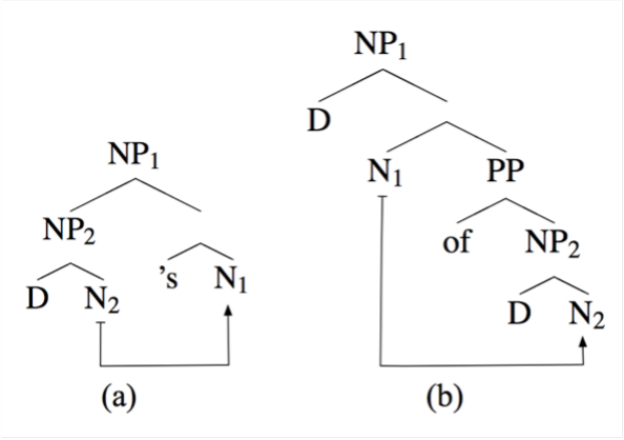

Figure 3.1: Syntactic representations of *s*-possessive and *of*-possessive in English.

Figure 3.1 shows the syntactic representation of the English *s*-possessive and *of*-possessive. I

use N_1 and N_2 to denote the two head nouns. In both the *s*-possessive and the *of*-possessive, N_1 is the head of the structure. Thus, a structure as in Figure 3.1 is recursive if N_1 and N_2 are substitutable: nouns used in one position can also be used in the other, indicated by a directional arrow ($\mapsto$). For example, as will be discussed in details in later sections, $N_2 \mapsto N_1$ holds for the English *s*-possessive (Figure 3.1a): nouns in N_2 can also be used in N_1, i.e., the possessor can be possessed. Therefore, the structure is recursive. The *of*-possessive (Figure 3.1b) is recursive because $N_1 \mapsto N_2$: nouns used in N_1 can also be used in N_2, i.e., the possessee can possess. Note that the directionality of substitutability is different for the two structures, which is a distributional property that English-learning children must acquire.

This proposal applies to all recursive structures where the two structural positions are in a selectional relation so that they can "see" each other. The studies in this chapter will all focus on such structures; and in the general discussion section we will touch on recursive structures without a selectional relation. In brief, we believe they can also be learned distributionally but in a slightly different way.

The only representational prerequisite of the current proposal is that the substitutable element is the head of the structure. For instance, the structure N_2's N_1 is really an instance of N_1, the head. Therefore, representationally, N_1 and N_2's N_1 are the same thing, i.e., a syntactic objected headed by N_1. The substitutability criterion thus follows because the notion of the head establishes an equivalence relation between a head noun and all syntactic objects headed by that noun, given that children have Merge. We can see the necessity of the head requirement from a counter example: consider 'NP_1-V-NP_2' structures in English, e.g., 'dogs chase cats'. It is possible for the two NPs to be substitutable, but that will not lead to recursion (e.g., '* dogs chase cats chase rats') because neither NP is the head of the structure.

Importantly, under the current conception of recursion, there cannot be embedding of finite depth (say, up to level 2 or 7): a structure is either not recursive or infinitely recursive (see also Huijbregts, 2019). In principle, a structure may be recursive for only one lexical item. Imagine a hypothetical variety of English where $N_2 \mapsto N_1$ holds in the *s*-possessive only for the noun 'mother'. That is, the language contains expressions such as 'mother's car', where 'mother' is in N_2, and 'the

student's mother', where 'mother' is in N_1; no other noun appears in both positions. Nevertheless, the structure is recursive albeit only for a single item: 'the mother's mother', 'the 40-year-old mother's 80-year-old mother's 120-year-old mother', etc. are immediately available and recursion ensues. If multiple nouns enable $N_2 \mapsto N_1$, then learner will seek to form generalizations about what makes these nouns eligible for N_2 (and thus recursion-triggering). If the generalization turns out to be valid in a way that will be specified in later sections, then novel, and potentially infinite many, items that follow the generalization can also be used recursively. The learning of recursion, then, becomes the problem of learning the lexicon for which structural substitutability holds.

While the proposal reduces the problem of infinite recursion to level-one structural substitutability, the learnability problem may still seem intractable. To know that $N_2 \mapsto N_1$ holds for the English *s*-possessive, for example, is to know that *all* nouns that can appear in N_2 can also appear in N_1. Of course, just the fact that an N_2 noun *can* be used in N_1 does not mean it will be in fact used as such, not least in a modest-sized child-directed input corpus. Therefore, when there are words attested in N_2 but not in N_1 in the input, one possibility is that $N_2 \mapsto N_1$ does not hold; it is also possible that $N_2 \mapsto N_1$ does hold - it may be completely acceptable for the unattested words to appear in N_1 but they just did not get the opportunity to do so. Thus, it still requires a leap of faith for the child learner to determine whether to form a generalization for recursion.

3.2.2 Productivity and generalization

As the second part of the proposal, I argue that the property of structural substitutability can be learned from distributional cues as an instance of learning productive generalization. The generalization problem is not unique to the acquisition of recursion but encompasses every aspect of language acquisition (and learning more generally): How does a grammar of infinite capacities arise from a finite sample of data that embodies the grammar?

Again, I use the TSP to predict the acquisition of this generalization. Recall that the TSP states that given N items, in order for a generalization R to generalize productively, the proportion of supporting items in the input data must exceed a precisely specified threshold, i.e., N - N/lnN. Therefore, for recursive structures, if a sufficiently large proportion of words in children's vocabulary

are attested in both positions in the input, then children will learn substitutability as a productive generalization, thereby ensuring recursion; otherwise, substitutability and recursion will be limited to the lexicalized attested items. Specifically, to test whether the generalization $N_2 \mapsto N_1$ is productive, for example, in our calculation we will use the number of words attested in N_2 in child-directed speech as N, and examine whether the number of words attested in both N_1 and N_2 exceeds the productivity threshold predicted by the TSP based on the value of N.

As discussed earlier, a critical property of the TSP is that the threshold for generalization is lower as a proportion of N when N is smaller, which may provide an account for why young children, who know relatively few words, can nevertheless accurately extract the productive rules in their language. I will show that learning recursion via structural substitutability benefits from a small vocabulary in a similar way.

3.3 Corpus study 1: Possessive

In this section I present corpus analyses on possessives in English, Mandarin Chinese and German as a test case for the proposal. Note that I use the term *possessive* here only to refer to the formal syntactic structures; it is well known that the meanings expressed by these structures are quite varied and not limited to canonical notion of possession such as ownership (Quirk et al., 1985). This study focuses on the formal and semantic properties of these structures available in the early stages of language acquisition: as we will see, they provide the core element for the extension of the possessives in broader usage.

3.3.1 English

We start with the English possessives. As discussed earlier, both the *s*-possessive and the *of*-possessive in English can be recursively embedded but in different conditions. Therefore, the question here is to identify the different properties of the nouns that enable recursion in the two structures. I will show that the different properties are learnable from one-level child-directed data.

3.3.1.1 Methods

Child-directed English data were extracted from the CHILDES database (MacWhinney, 2000). The English input corpus contains 12.6 million words, which constitute approximately two years of input for typical American English-learning children.

Due to their structural simplicity, the English possessives were extracted via a regular expression pattern matcher, followed by manual inspection that eliminated a small number of annotation errors. To approximate a young language learner's vocabulary, the analysis was confined to the possessives where the N_1 and N_2 positions are occupied by the 50 most frequent nouns in early child English (Rowe and Goldin-Meadow, 2009; Carlson et al., 2014). All in all, there are 1,070 *s*-possessives and 1,158 *of*-possessives. The search procedure returned many expressions that are typically analyzed as measure or classifier phrases (e.g., 'two cups of water', 'a piece of paper'). These were not eliminated a priori: Measure phrases also need to be learned by children, and it is useful to investigate how they are distinguished from the other uses of the *of*-possessive, which have the same formal structure. As discussed in Section 3.2, I assume that the child attends to the nouns that enable structural substitutability in order to identify the condition on recursion. Thus, I assume that the syntactic representation in Figure 3.1 is available to the learner: statements such as $N_1 \mapsto N_2$, which make reference to the structural configuration, can be formulated and evaluated.

3.3.1.2 Results

The Venn diagrams in Figure 3.2 displays the counts of the N_1 and N_2 nouns in the two possessives and the quantitative relation between the N_1 and N_2 sets, which plays a critical role in the determination of structural substitutability.

The two structures are analyzed separately by mimicking the computational process that the child learner may follow. The results clearly point to the conditions under which the two possessive structures can, and cannot, be used interchangeably, a topic of much previous research.

For the *s*-possessives, as Figure 3.2 (left side) illustrates, there are 22 N_2 nouns in the *s*-possessive, of which 18 also appear in the N_1 position. The four that fail are 'girl', 'bear', 'fish',

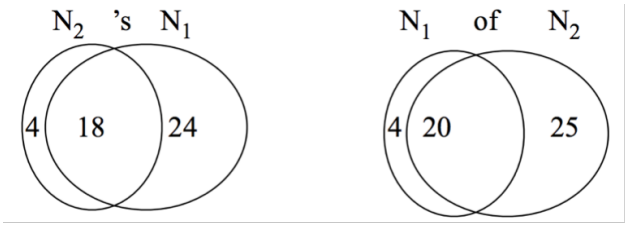

Figure 3.2: Set relations between the N_1 and N_2 nouns in the English possessives. The numbers indicate the cardinality of N_1, N_2, and the cardinality of their intersections.

and 'color', which clearly can appear in N_1 but simply did not have the opportunity to do so in the child-direct speech corpus. Nevertheless, the TSP threshold for generalization is met (θ_{22}=7): $N_2 \mapsto N_1$ thus holds. In other words, if a noun is used in N_2, it can be used in N_1 as well, thereby enabling recursion. Note that substitutability in the other direction, i.e., $N_1 \mapsto N_2$ is untenable: 18 out of 42 is not anywhere near sufficiency under the TSP (31 is required).

Having identified the condition for recursion, the learner can now discover what makes nouns eligible for N_2 thereby triggering recursion. At least in the child-directed corpus, almost every *s*-possessive expresses the meaning of possession, which can be divided into two kinds. The first kind can be called internal possession: N_1 is a kin, body part, attribute, characteristic, and other inherent property of N_2, perhaps as an extension of inalienability and part-whole relations, which are often grammatically marked in the world's languages (e.g., Hyman et al., 1970; Dol, 1999). Here we find 'dog's hair', 'man's son', 'baby's name', etc. The second kind can be referred to as external possession, where N_2 expresses ownership and other contingent relations with N_1. This is sometimes referred to as 'alienable possessives' in literature (e.g., Nichols and Bickel, 2013). Here we find 'baby's cat', 'daddy's bed', 'mommy's lunch', etc.

Previous corpus analyses of adult speech and written corpora have found that the *s*-possessive strongly favors animate nouns in the N_2 position (e.g., O'Connor et al., 2013). Indeed, the statistical tendency is overwhelming in terms of token frequency: 99% of the *s*-possessives have an animate noun in the N_2 position. However, inanimate nouns are not uncommon in terms of types, which is the quantity over which productivity is calculated under the TSP. In fact, 12 out of 22 N_2 nouns are inanimate, even though most appear only once. The animate nouns, despite their abundance, cannot be regarded as a sufficient condition for N_2 while relegating the inanimate

nouns as lexicalized exceptions (θ_{22}=7). The semantic property of internal and external possession, however, does appear sufficient. When an inanimate noun is used in the N_2 position, it is always a case of internal possession (e.g., 'car's name') or anthropomorphic extension (e.g., 'bus's house', 'train's way', 'flower's face', 'next door's cat'), which is consistent with the well-documented pattern in literature (Belvin, 1993; Harley, 1998).

In comparison to the *s*-possessive ($N_2 \mapsto N_1$) the structural substitutability in the *of*-possessive goes in the direction of $N_1 \mapsto N_2$. As illustrated in Figure 3.2 (right side), 24 nouns are used in N_1 of which 20 appear in N_2 as well, easily clearing the TSP threshold (θ_{24}=7). Thus, the *of*-possessive is recursive.

Again, let us consider the eligibility of nouns in N_1, the crucial condition for recursion. 4 of the N_1 nouns are measure words as in 'a piece of fish', 'a bit of cheese', 'two cups of juice', and 'a box of food'. The remaining Ns all express internal possession in the sense defined earlier: N_1 is a part, component, attribute, characteristic, and other inherent property of N_2. Examples include 'picture of the boy', 'middle of the night', 'time of the day', 'day of the week', 'color of the flower', 'head of the man', etc. Importantly 20 out of 24 guarantees productivity by the TSP (again, θ_{24}=7): internal possession can be regarded as the productive condition for N_2 nouns, with the 4 measure words lexicalized as exceptions.

Unlike the *s*-possessive, noun animacy does provide a categorical criterion for the N_2 position in *of*-possessives: 22 out of 24 nouns are inanimate, and the two animate exceptions are both nouns referring to humans, in a fixed expression: 'daddy/man of the house'. This usage is also consistent with the semantic characterization of internal possession — if the child understood it as intended: the 'man' or 'daddy' is understood as the (senior) male member of a family as opposed to some male or father residing in the house. The inanimacy condition is likely to remain productive even if we expand the learner's vocabulary. The entire 12.6-million-word corpus contains 630 N_1 nouns in the *of*-possessive, only 17 are animate: 'man', 'daddy', 'mother', 'baby', 'friend', 'boss', 'prince', 'fan', 'director', 'boy', 'girl', 'principal', 'president', 'father', 'family', 'cousin', and 'author'. These are numerically well below the TSP threshold and can be lexicalized as exceptions, or learned as a group (i.e., nouns referring to humans). Notably, nouns referring to animals, of which there are

many, never make an appearance in N_1. On the other hand, it is possible that some kinship terms, always animate, may be used in the *of*-possessive. Although none is found in the input corpus, both 'the son of the mother' and 'the mother of the child' can be found in adult corpora. One reason may be that mother and child inherently possess each other by definition. Another, simpler, reason may be that learners eventually encounter expressions such as 'mother of Dragons' and '(in) the name of the mother': the attestation of 'mother' in both N_1 and N_2 guarantees recursion by fiat according to our formulation.

From the child-directed input analysis, it can be concluded that in addition to measure phrases,[1] the relation of internal possession between N_1 and N_2 enables $N_1 \mapsto N_2$ and therefore recursion for the *of*-possessive. Moreover, N_1 must be inanimate as a rule. Note that external possession (e.g., ownership), freely available in the *s*-possessive, cannot be expressed with the *of*-possessive. This can be seen in the contrast between 'the screen of the laptop' and '*the monitor of the laptop' (meaning the external display): the screen is a built-in and thus inherent component of the laptop but the monitor is a detachable and independent device.

We can now account for *of*-possessive embedding in (33), repeated below:

(36) The top of the tip of his hat (attested in CHILDES input)
The end of the story of the kitty (attested in CHILDES input)
The color of the cover of the book
The middle of the third inning of the deciding game
The son of the President of the Union of Retired Professors

These examples all obey the criterion that N_1 is an internal possession of N_2. Their rarity is likely due to the fact that it is unusual for multiple nouns to form an appropriate Russian-dolls-like chain of internal possession, but it is not difficult to construct examples that obey such relations thereby extending recursion: 'the hue of the color of the cover of the book', 'the middle of the third inning of the deciding game of the World Series', etc.

The distributional analysis of the possessive structures in child-directed English produces the

[1]Measure words that appear in both N_1 and N_2 positions appear to allow recursion as our formulation predicts. From 'the price of a pint' (N_2) and 'two pints of beer' (N_1), recursion follows for 'pint', as 'an order of three pints of beer' is possible.

following conditions for recursive embedding, correctly capturing restrictions on those structures recorded in literature (e.g., Rosenbach, 2014). Only valid generalizations are listed. Exceptions can nevertheless trigger recursion if attested in both N_1 and N_2 sufficiently often to be lexicalized as such. However, those will depend on individual linguistic experiences and could vary across people. The generalizations below, by contrast, are robust and shared among speakers.

(37) a. *s*-possessive (N_2's N_1) is recursive if N_2 and N_1 are related by internal or external possession (i.e., the possessor can embed).

b. *of*-possessive (N_1 of N_2) is recursive if inanimate N_1 and N_2 are related by internal possession (i.e., the possessee can embed).

Although the conditions above are established on a very small set of child-directed data, their applicability is considerably broader. For example, they account for a well-known semantic difference between the two possessives (Chomsky, 1970). The *s*-possessive 'the man's picture' has two readings: a depiction of the man or an item in the man's collection, i.e., internal and external possession (37a). For the *of*-possessive 'the picture of the man', by contrast, only the internal possession reading (37b) is available. We must note that these conditions are, and probably can only be, approximate. On the one hand, they are extracted from child-directed input data, which does not contain many of the rare and non-canonical uses of the possessives noted in the literature. On the other, and more fundamentally, these conditions depend on the language user's understanding of noun concepts and their relations, which can be flexible and fluid as the anthropomorphized examples in the input corpus illustrate (e.g., 'the flower's face', 'the truck's home').

Figure 3.3 illustrates the semantic scope of the two possessives: for internal possession, both possessive structures are possible, as illustrated by the examples. Again, there are strong statistical tendencies that favor one (underlined) over the other. Recall that N_1 position in the *of*-possessive is overwhelming inanimate and the N_2 position in the *s*-possessive is overwhelming animate. However, as discussed earlier in light of type frequencies and the TSP, the inanimacy condition in the former is a categorical one whereas the animacy condition in the latter is merely a matter of preference. Other factors contribute to the language user's preference when both options are grammatically available; see O'Connor et al. 2013 for a detailed quantitative analysis. The current study can be

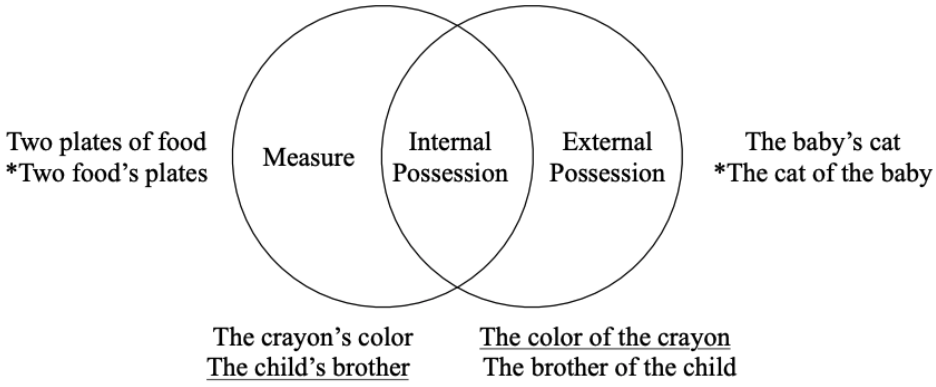

Figure 3.3: The semantic conditions for the English possessives. The underlined expressions are statistically preferable.

viewed as a prerequisite: it identifies the conditions on the possessives — and thus when they do (and do not) overlap. The preferences can then be learned through other learning mechanisms.

3.3.2 Mandarin Chinese

We now turn to the Mandarin possessives. To reiterate the empirical facts, there is a possessive marker *de* in Mandarin. The possessive structure can freely embed when the marker is used. In limited cases the marker can be omitted, but that structure in general cannot embed except for highly restricted vocabulary, mostly kinship terms (Li and Thomson, 1981).

3.3.2.1 Methods

I analyzed all the annotated CHILDES corpora (1.6 million words of input) and focused on the 56 nouns that were representative of three-year olds' vocabulary (Hao et al., 2008). I extracted all the N_2 *de* N_1 and N_2 N_1 possessive structures[2] that contain those words in either N_1 or N_2 position.[3]

[2]This excludes N_2 N_1 structures appearing at the beginning of sentences, which are topicalization instead of possessives (Li and Thomson, 1981).

[3]Due to smaller corpus size, this departs from the English analysis, since otherwise there will be little data available.

3.3.2.2 Results

The results are given in Figure 3.4: Bidirectional substitutability ($N_1 \leftrightarrow N_2$) can be maintained for N_2 *de* N_1, as the 28 nouns used in both N_1 and N_2 are sufficient for $N_1 \mapsto N_2$ as well as $N_2 \mapsto N_1$ (θ_{38}=10). The bidirectionality of substitutability means that no additional constraints on nouns are necessary. In the construction without *de*, however, only 9 nouns appear in both positions, falling far short of the TSP threshold for either direction of substitutability. Thus children will learn these nouns lexically: close kinship terms (e.g., 'father') and body parts (e.g., 'head'). Indeed, recursion with such limited vocabulary is possible: 'baba' ('dad') appears in both N_1 ('baba toufa', 'dad's hair') and N_2 ('baobao baba', 'baby's dad'), so 'baobao baba toufa' ('baby's dad's hair') follows immediately without the use of the possessive *de*.

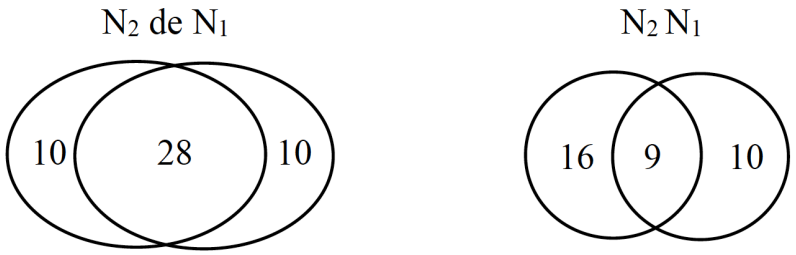

Figure 3.4: Set relations between the N_1 and N_2 nouns in the Mandarin possessives. The numbers indicate the cardinality of N_1, N_2, and the cardinality of their intersections.

The distribution of the Chinese possessives is very similar to the German possessives: In Li et al. (2021), we investigated German *von*-possessive and *s*-possessive using 5 CHILDES corpora (yielding a dataset of 3.5 million words of input).[4] Since there is no established study on the nouns in German-speaking children's early vocabulary, we searched for the 50 most frequent nouns in the input and extracted all possessive structures containing these in N_1 or in N_2 position. This resulted in a total of 368 *s*-possessives and 888 *von/vom*-possessives. The results are shown in Figure 3.5. It is found that in the *von*-possessive, similar to Chinese N_2 *de* N_1, it is evident that N_1 and N_2 are bidrectionally substitutable ($N_1 \leftrightarrow N_2$). There are 39 nouns used in both positions, exceeding the generalization threshold for both $N_1 \mapsto N_2$(θ_{46}=12) and $N_2 \mapsto N_1$ (θ_{42}=11). Again, because substitutability is bidirectional in the German *von*-possessive, the learner would not even need to

[4]I thank Lydia Grohe and Petra Schulz for the analysis of German data.

investigate the properties of the nouns further. For the German Saxon *s*-possessive the picture is very different: there are 5 N_2 nouns, all of which appear in the N_1 position as well. Accordingly, the child can conclude that the *s*-possessive is not generally recursive but may lexicalize the five attested nouns that do appear in both positions. Lexicalization in the absence of productivity, however, requires extensive exposure; it is thus possible that not all speakers allow the recursive embedding of these 5 nouns. Although some kinship terms such as 'Mama' ('Mom') and 'Papa' ('Daddy') seem to allow recursion for some speakers, it is unclear whether the first N_2 receives a proper name reading in these cases. These findings are in line with Pérez-Leroux et al. (2022)'s results from elicited production: when prompted to produce recursive possessives, children at age 5 provided recursive *von*-structures such as 'der Ballon von dem Affen von dem Clown' ('the balloon of the monkey of the clown) and not a single recursive *s*-possessive. In addition, the Saxon *s* was used only very rarely even at level one.

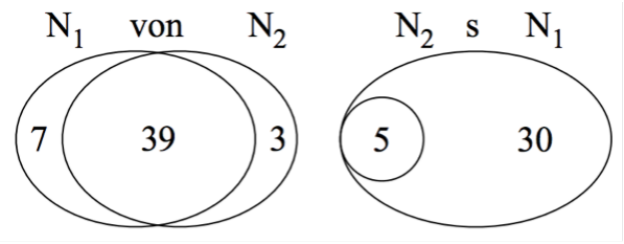

Figure 3.5: Set relations between the N_1 and N_2 nouns in German possessives. The numbers indicate the cardinality of N_1, N_2, and the cardinality of their intersections.

3.3.3 Discussion

In summary, the proposed approach to recursion has correctly captured the conditions for recursive embedding for different possessive structures in a typologically diverse range of languages – English, Mandarin Chinese, and German. First, for freely recursive structures without any restriction - German *von*-possessive and Mandarin *de*-possessive, we found N_1 and N_2 are bi-directionally substitutable, so children can learn that those structures can be freely embedded. English *s*-possessive and *of*-possessive are both one-way substitutable: It is $N_2 \mapsto N_1$ for the *s*-possessive and $N_1 \mapsto N_2$ for the *of*-possessive, where N_1 is the possessee. Therefore, those structures should only

be recursive for the types of words eligible for N_2 in the *s*-possessive and for N_1 in the *of*-possessive, and children need to discover what nouns are eligible for those positions and thus trigger recursion. Through semantic analyses of attested words in the input, we showed that children can discover the well-documented restrictions on those structures for recursive embedding, such as external vs. internal possession and animacy. Finally, for German *s*-possessive and the possessive without *de* in Mandarin, the proportion of nouns appearing in both positions fail to meet the threshold of productivity for each direction, so depending on individual's linguistic experience, those structures will either be recursive only for the highly limited words attested in both positions, or not recursive at all because lexicalization in the absence of productivity is hard. Overall, findings from the corpus study suggests that recursion, as structural substitutability, can be learned as a productive generalization from one-level language specific input.

3.4 Corpus study 2: VV

To further test the proposal, this section looks at a structure in a different domain that can also be recursively embedded: VV constructions in Mandarin Chinese. Apart from predicate-object relations, there are two major types of Mandarin VV constructions: resultative verb compounds (RVC), and serial verb constructions (SVC). In an RVC, the second element signals some result of the action or process conveyed by the first element, such as 'da-po' ('hit-break'), 'zou-lai' ('walk-come') (Li and Thomson, 1981). An SVC is a mono-clausal construction where two or more verb phrases are juxtaposed to describe a single event, without any marker in between, e.g., 'hui jia chi fan' (lit. 'go home eat meal'),'kai che shang ban' (lit. 'drive car go-to work') (Chao, 1980; Li and Thomson, 1981; Zhu, 1982; Bisang, 2009; Haspelmath, 2016). RVCs generally cannot recurse[5] while SVCs can, e.g., 'fang xue hui jia xie zuo ye' (lit. 'leave school go home do homework'), but RVCs can be embedded inside serial verb constructions, e.g., 'pao-lai wan' (lit. 'run-come play'). In this section I examine how the conditions of recursively embedding Mandarin VV constructions can be learned.

[5] An exception which only works for a small set of words is that in directional RVCs, the second constituent can be an RVC that consists of a directional verb and 'lai' ('come') or 'qu' ('go'), such as 'zou-jin-lai' ('walk-enter-come') (Li and Thomson, 1981).

3.4.1 Methods

Because the Mandarin CHILDES corpora are too small to represent children's VV input (1.6 million words, approximately three months of input), I used the dialog corpus from the Beijing Language and Culture University Corpus Center (BCC) (http://bcc.blcu.edu.cn/) instead (Xun et al., 2016). The dialog corpus contains sentences from social media and movie subtitles, comprising around 600 million characters, which means around 375 million words given the word:character ratio in Mandarin. I chose the dialog corpus among all the corpora from the BCC corpus because its content is supposed to be more similar to everyday conversations that children will hear compared to the other corpora (e.g., newspapers, literature). Although this corpus is not child-directed, Kodner (2019) demonstrated that for high token frequency items, non-child text corpora are effectively interchangeable with child-directed speech corpora for the purpose of estimating child lexical experience. Since I will only focus on highly frequent verbs which are representative of three-year olds' vocabulary, the BCC corpus can thus be used to model children's language acquisition.

I used the top 48 verbs that three-year-old children can reliably produce (Hao et al., 2008). I extracted all V_1 (N) V_2 examples that are not in a VO relation[6] from the input where both the V_1 verb and the V_2 verb are from our verb list.

3.4.2 Results

Firstly, when all the examples are considered, V_1 and V_2 are bidirectionally substitutable: as Figure 3.6 shows, 38 easily clears the threshold for 42 ($\theta_{42} = 11$), so VV constructions as a whole is freely recursive. RVCs, however, can be distinguished from the remaining VV examples by their potential form: A unique property of RVCs is that they allow the insertion of *de* 'obtain' or *bu* 'not' between their two constituents to produce a potential meaning: The insertion of *de* gives the compounds an affirmative potential meaning, (38a) while the insertion of *bu* gives a negative potential meaning, (38b), (Li and Thomson, 1981).

[6]Including VO examples in the analysis does not change the result. There is only one verb in our verb list, namely *xihuan* 'like', that exclusively appear in VO examples and not in other VV constructions. Furthermore, given children's sensitivity to argument role relations and the syntactic frames that verbs appear in (e.g., Gleitman et al. 2005), they should be able to distinguish this mental state verb from the other verbs in the verb list.

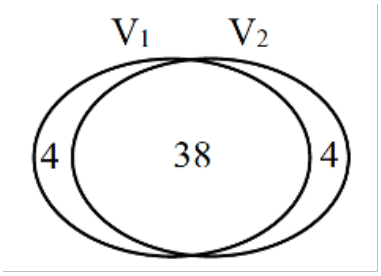

Figure 3.6: Set relations between the V_1 and V_2 nouns in Mandarin VV constructions. The numbers indicate the cardinality of V_1, V_2, and the cardinality of their intersections.

(38) a. ta tiao-de-guo-qu.
3SG jump-obtain-cross-go
'S/He can jump across.'

b. ta tiao-bu-guo-qu.
3SG jump-NEG-cross-go
'S/He cannot jump across.'

Therefore, the potential form provides an unambiguous way to distinguish RVCs from other VV constructions. In our data, 32 out of the 59 RVC examples have appeared in the potential form, and they exhibit patterns that are generalizable to almost all the other RVCs. As has been well recorded in literature (e.g., Li and Thomson, 1981), an important type of RVCs is directional RVCs, where the first verb implies a displacement, and the second verb signals the direction in which some entity moves as the result of the displacement, such as 'pao-lai' ('come running', lit. 'run-come'). In particular, there are three major types of displacement verbs in directional RVCs: motion verbs, such as 'zou' ('walk'), 'pao' ('run'); verbs that may cause the direct object to undergo displacement, such as 'song' ('send, give'), 'da' ('hit'); and verbs of inquiry, perception, or saying when the second element is 'chu-lai' ('out', lit. 'exit-come'). The last type makes metaphorical RVCs, where 'chu-lai' means 'find out' or 'come out' instead of 'exit' in the physical sense, e.g., 'kan-chu-lai' ('find out by watching', lit. 'watch-exit-come'), 'shuo-chu-lai' ('get something out by saying it', lit. 'say-exit-come'). Indeed, in our verb list, all five motion verbs ('zou' ('walk'), 'pao' ('run'), 'tiao' ('jump'), 'fei' ('fly'), 'pa' ('crawl')), 7 out of 8 verbs that may cause the direct object to undergo displacement (e.g., 'bao' ('hold'), 'na' ('take'), 'mai' ('buy'), etc.), all four perception verbs (e.g., 'kan' ('watch'), 'mo' ('touch'), etc.) and all three saying verbs ('ku' ('cry'), 'chang'

('sing'), 'jiao' ('yell')) appear in the V_1 position of RVCs in potential forms. Four out of five verbs or verb compounds that can signal direction ('hui-jia' ('come-home'), 'chu-lai' ('exit-come'), 'zou' ('walk'), 'fei' ('fly')) appear in the V_2 position of RVCs in potential forms. The only other verb attested in V_2 of RVCs in potential forms is 'ku' ('cry'), which is used in causal RVCs, where the event described by V_1 causes the event described by V_2, such as 'kan-ku' (lit. 'watch-cry', 'watch something and as a result cry'), 'da-ku' (lit. 'hit-cry', 'hit somebody and as a result somebody cries'). For either RVCs in potential forms or all RVCs, the number of verbs attested in both V_1 and V_2 positions fails to reach the productivity threshold ($\theta_{20} = 6, \theta_{24} = 7$) (Figure 3.7), suggesting RVCs cannot recursively embed.

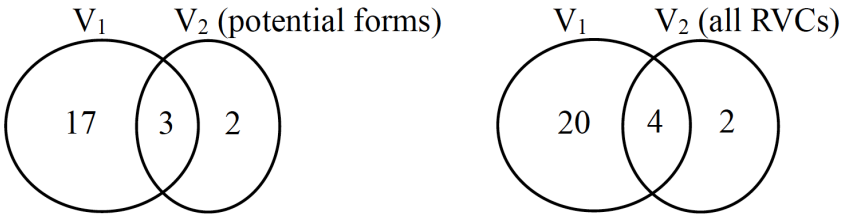

Figure 3.7: Set relations between the V_1 and V_2 nouns in Mandarin RVCs. The numbers indicate the cardinality of V_1, V_2, and the cardinality of their intersections.

The set relations between V_1 and V_2 verbs in VV constructions after excluding RVC examples are shown in Figure 3.8, where $V_1 \mapsto V_2$ is productive ($\theta_{34} = 9$) but $V_2 \mapsto V_1$ is not ($\theta_{42} = 11$). Therefore, SVCs are recursive, and the learner needs to discover what verbs are eligible for V_1 and thereby trigger recursion. Consistent with the literature, the semantic relation between the events described by the two verb phrases can be divided into three kinds (Chao, 1980; Li and Thomson, 1981; Zhu, 1982). First, the two events occur consecutively; this kind can sometimes be difficult to distinguish from the second kind, where the first event is done for the purpose of achieving the second event. Such examples include 'hui jia chi fan' (lit. 'go home have meal', 'come home and/to have meal'), 'guan deng shui-jiao' (lit. 'turn-off light sleep', 'turn off the light and/to go to bed',), 'chu-lai wan' (lit. 'exit-out play', 'come out and/to play'). For those SVCs, VP_1 is usually a VO phrase or a compound that describes a telic event. In the other kind of SVCs, the first VP, which is an atelic event or state, is the circumstances under which the event in the second VP occurs, such as 'kai che hui jia' (lit. 'drive car come home', 'drive home') and 'zou-lu shuai-dao' (lit. 'walk

fall', 'fall while walking').

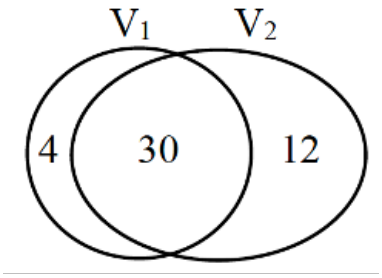

Figure 3.8: Set relations between the V_1 and V_2 nouns in Mandarin SVCs. The numbers indicate the cardinality of V_1, V_2, and the cardinality of their intersections.

These results are consistent with findings from acquisition literature that the rules for VV constructions are available to Mandarin-speaking children pretty early. For RVCs, they are spontaneously used in abundance from around 1;7 (e.g., Xu, 2006; Chen, 2008; Deng, 2010, 2019); moreover, experiments using novel verbs showed that 3-4-year-old children can comprehend and produce different types of RVCs following the rules discussed above and they never used the structure recursively (e.g., Deng, 2010, 2019). For SVCs, they are also observed to emerge in children's spontaneous speech after around 1;5 (e.g., Yang, 2006) and are used productively. For instance, one of the children Kang from the Erbaugh corpus (Erbaugh, 1992) produced an innovative SVC example which cannot be heard from the input, suggesting the acquisition of a productive rule: He produced "zuo ji-cheng-che zhuan qian" (lit. 'take taxi make money', 'make money by taking taxi') at the age of 3;2 when playing going to work and then taking taxi home. This semantic anomalous yet grammatical SVC was due to his misunderstanding of the game: he thought he was paid for taking the taxi instead of for work. (39) shows more examples of children's spontaneous production of recursive SVCs from CHILDES which were not attested in their input (MacWhinney, 2000).

(39) a. lai chitang he shui xi-zao
come pond drink water bathe
'come to the pond and/to drink water and then have a bath'

b. hui jia mai tang-guo gei ni chi
return home buy sugar give you eat
'come home and/to buy sugar for you to eat'

c. xi-zao ca-ca chi fan
bathe wipe-wipe eat meal
'have a bath and then dry off and then have meal'

d. na qian hui-lai mai dong-xi
take money return-come buy things
'return with money and/to buy things'

e. yong zhua yu de dong-xi na-shang-lai huan-gei ta
use catch fish DE thing pick-up-come return-give him
'use the thing for fish-catching to pick it up and give it back to him'

3.4.3 Discussion

Our first corpus study looked at recursive structures in the NP domain. To further test the proposal, we conduct Corpus Study 2 in a different linguistic domain, using VVs in Mandarin as a test case. We found that the substitutability rule $V_1 \mapsto V_2$ is productive in one-level data of SVCs; and further semantic analysis correctly captured the rules for SVCs. Therefore, Mandarin-speaking children can acquire the recursive embedding rules for SVCs from distributional cues. This is in sharp contrast to languages where VVs cannot embed. For example, in English VV constructions, the V_1 position is largely limited to 'come' and 'go', so no productive generalization can hold. Overall, the results demonstrate that our new approach to recursion can accurately account for the learnability of recursive structures in different domains and languages as long as they meet the prerequisites for a selectional relation and head substitutability.

3.5 General discussion

In this chapter, I proposed a new approach to recursion which has two distinct components. The conception of recursion as structural substitutability (e.g., $V_1 \mapsto V_2$) transforms the problem of infinite embedding to the problem of productivity that can be resolved on level-one data. This in turn calls for a theory of learning and generalization, with the TSP on offer in our proposal. Through corpus analyses on a range of constructions from different languages, I demonstrate that detection of structural substitutability - and thus recursion - is possible with simple level-one embedding

data. For freely recursive structures that pose no semantic constraints, a sufficiently large number of words are attested in both structural positions in the input to meet the productivity threshold so that bidirectional substitutability can be established. For structures that can be recursively embedded in certain conditions, those semantic constraints are also learnable from simple one-level data consisting of children's modest vocabulary and do not need to be innate. And finally, for structures that generally do not allow recursion, the number of words attested in both positions fails to reach the productivity threshold; a learner may still learn that a limited set of words can nevertheless trigger recursion in those structures if the words are attested in both positions frequently enough in that learner's input. Overall, the results suggest that recursive structures can be learned distributionally from language-specific experience.

I would like to make it clear what this proposal is *not* about. It is not about acquiring the ability of recursion, or Merge. We assume children already bring this ability to the table. Instead, we are interested in how children learn from language-specific experience in which situations this ability of recursion can be applied. This current proposal is also not a theoretical analysis of why languages allow or do not allow recursive embedding in certain structures. We are aware of a number of such analyses (e.g., Hoekstra, 1984; van Riemsdijk, 1988, 1998; Richards, 2006; den Dikken and Dékány, 2018), and our proposal is not necessarily inconsistent with them, but we are focusing on a different problem, namely how children discover the rules for recursive embedding in their language from the finite input data available to them.

We regard the notion of structural substitutability as crucial for all treatments of recursive structures that meet the selectional prerequisite. Recursive structures without a selectional relation are not the focus of the present work, but we believe they are also learnable as a productive generalization from distributional cues. For instance, consider the familiar case of CP recursion, such as 'I think she believes ...'. In our view, the crucial element that decides whether such embedding is possible is the main verb (e.g., 'think' and 'believe' are fine, but 'put' and 'eat' not). The verbs in this configuration, however, clearly cannot select each other since they are in different CPs. Therefore, in order to learn this recursion, learners will need to learn (i) there are verbs that can take CP complements; (ii) those same verbs can be the main verb in an embedded CP. Therefore,

to learn (i) and (ii), children still need to learn that the two verb positions are substitutable. It turned out that every CP-taking verb in our English child-directed corpus embeds each other, easily clearing the TSP threshold, and the child can conclude that all CP-taking verbs allow recursion. Thus, the general idea of substitutability and learning it as a productive generalization still works, but applies in slightly different ways when there is no selectional relation between the crucial positions.

A notable finding from this chapter is that the basic properties of the recursive structures can be acquired on very simple data such as our child-directed input corpora. This may not be surprising upon reflection. After all, language acquisition is remarkably early and accurate: the necessary distributional evidence must be readily available in the limited input that children receive. More interestingly, recursive structures may be learnable only if children have a very small vocabulary of extremely frequent words. In order for structural substitutability to hold, the vast majority of the items in one position must also be attested in the other. Given the sparsity of linguistic distributions, only highly frequent items will have the opportunity to appear in both structural positions. Thus, as the size of the lexicon increases, the intersection of the two sets will shrink as a proportion: the identification of productive processes becomes impossible under the TSP. Indeed, the generalizations in (37) no longer hold when all input nouns are included in the calculation: Less does seem to be more (Newport, 1990).

In summary, our approach to recursion and its acquisition embodies of a view of language acquisition that prioritizes the formal properties of grammatical structures. Recall the requirement that, as a rule, only inanimate nouns can appear in the possessee (N_1) position of the *of*-possessive. This generalization was formed after the establishment of $N_1 \mapsto N_2$, a purely formal and quantitative condition that all N_1 nouns can appear in N_2 but crucially not vice versa. No such semantic refinement is necessary for the German *von*-possessive and the Chinese *de*-possessive where N_1 and N_2 are bidirectionally substitutable. The core semantic properties of the possessives can be identified subsequently, provided that children have the requisite conceptual understanding of possession and ownership (Nancekivell et al., 2019). To be sure, the primacy of formal structure has been recognized since the foundation of modern linguistics (Chomsky, 1955): our proposal provides a

psychological procedure for the discovery of formal structures, from which their semantic content can be established. Or, as Chomsky (1957) remarks, "How are the syntactic devices available in a given language put to work in the actual use of this language?" (Chomsky, 1957, p.93)

In future research, other tests of the full range of the proposal will be welcome and necessary. It would be desirable to evaluate the proposal on a broader range of linguistic phenomena across domains and languages, not only beyond those examined in this dissertation, but also including structures where the constraints on recursion may be more complex and where the structural positions are not in a selectional relation such as CP recursion and PP recursion. Moreover, since the current approach relies on formal distributional regularities and is independent of the substantive form that the items appear in, there is nothing in its nature that would constrain it to the domain of language, and thus it would be interesting to test whether it applies in other domains as well. For instance, recursion has been studied in domains such as formal patterns (e.g., Ferrigno et al., 2020), music (e.g., Lerdahl and Jackendoff, 1983) and social cognition (e.g., Jackendoff, 2007). Independent lines of work has shown that learners can extract distributional regularities to learn abstract rules in such domains (e.g., Marcus, 2001; Dawson and Gerken, 2009; Baillargeon et al., 2010; Liberman et al., 2013). Therefore, studies that test the current approach in non-linguistic domains could provide novel insights into the language learning mechanism: What processes may be specific to language, and what processes can be accomplished by domain-general cognitive functions? Finally, another important question is whether human learners can indeed make use of such distributional information during language learning, which I will investigate in the next chapter.

Chapter 4

ACQUIRING RECURSIVE STRUCTURES IN ARTIFICIAL LANGUAGES

4.1 Introduction

Chapter 2 and 3 have demonstrated the availability of reliable cues in child-directed speech for learners to acquire syntactic generalizations. However, it is necessary to test the proposal not only against corpus data, but also against human learning behavior, to determine whether it can indeed be a useful mechanism during language acquisition. In this chapter, I investigate whether human learners can use distributional cues as predicted in an artificial language learning paradigm (Esper, 1925; Braine, 1963; Reber, 1967). Artificial language learning experiments can provide important insights into the language learning mechanism since they allow precise control over the language input that learners are exposed to, which is virtually impossible using only natural language data. In particular, in the case of learning generalizations, corpus analyses can only provide an estimate of a child's vocabulary, whereas in an artificial language learning experiment we can precisely control how many regular and irregular words the participant acquires. Therefore, the artificial language paradigm is a particularly helpful methodology to examine the hypotheses in this dissertation.

Previous studies using this paradigm have demonstrated human learners' ability to acquire a range of linguistic knowledge from distributional information alone, such as word segmentation (e.g., Saffran et al., 1996), phonology (e.g., Maye and Gerken, 2001; Maye et al., 2002), categories (e.g., Mintz et al., 2002; Gerken et al., 2005; Reeder et al., 2013; Schuler et al., 2017), phrase structure (e.g., Gomez, 1997, 2002; Gomez and Gerken, 1999; Saffran, 2001, 2002; Thompson and Newport, 2007), and morphological rules (e.g., Schuler et al., 2016; Schuler, 2017). In this chapter, I examine what learners can learn about rules of recursive embedding given purely distributional information with a series of artificial language learning experiments.

There are three experiments in this chapter. In Experiment 1, I asked whether adults can use cues on substitutability in non-embedded examples to acquire recursive structures. I designed two languages that differed in the productivity of structural substitutability: There were sufficient words attested in both structural positions only in one language. It was found that as predicted, adults who received sufficient evidence for substitutabilty were significantly more likely to allow recursive embedding. This finding indicates that adults can indeed use distributional cues to acquire recursive structures. Experiment 2 examined the role of the structural representation. In particular, an open question from Experiment 1 was whether the participants treated the substitutable elements as the head of the structure, which is a requirement of the distributional learning proposal in Chapter 3. Therefore, Experiment 2 provided participants with cues on both structural substitutability and headedness to examine whether they could integrate these cues to acquire recursive structures. I designed two artificial languages: Both languages have sufficient cues for structural substitutability, but only in one language the substitutable elements are the head of the structure. It was found that although participants in both conditions learned substitutability in non-embedded data, only participants who were in the head substitutability condition were willing to accept recursively embedded strings. This suggests that learners can integrate knowledge of headedness and structural substitutability to acquire recursive structures. Finally, Experiment 3 investigated whether children can also use distributional cues to acquire recursive structures. This is an important question given the observed differences between children and adults in first and second language acquisition (e.g., Johnson and Newport, 1989; Mayberry and Kluender, 2018) and also in artificial language learning

experiments (e.g., Weir, 1964; Hudson Kam and Newport, 2005). Through a modified version of Experiment 1, I demonstrated that 6-8 year-old children can acquire recursive structures through distributional learning, although they also exhibited subtle differences from adults. Overall, this chapter demonstrates that both adults and children can acquire recursive structures from purely distributional information.

4.2 Experiment 1

In this section I report an experiment that tests whether learners indeed use distributional information on the productivity of structural substitutability to acquire recursive structures. To preview the experiment, adult participants were exposed to X_1-ka-X_2 strings in an artificial language. The language was designed such that $X_1 \mapsto X_2$ is productive in one condition but not in the other, and we found that participants from the Productive condition were indeed more willing to accept recursively embedded strings (X_1-ka-X_2-ka-X_3) with words unattested in X_2 and X_3 positions than participants from the Unproductive condition.

4.2.1 Methods

4.2.1.1 Participants

Participants were 50 adult native English speakers with typical hearing and vision (or corrected vision). All participants were recruited and run online via Prolific Academic (www.prolific.com) and paid 9 dollars/hour as compensation. The 50 participants were assigned to one of two language conditions, Productive or Unproductive, though 2 participants in the Unproductive condition did not complete the experiment and were excluded from analysis. The final sample of participants includes 48 adults, with 23 in the Unproductive condition (age = 30.48, range = 19-47) and 25 in the Productive condition (age = 27.42, range = 19-40).

4.2.1.2 Stimuli

The exposure stimuli in both conditions consisted of 44 strings generated from an artificial grammar of the form X_1-ka-X_2, where X_1 and X_2 denote the position in the structure (pre- or post-*ka*, respectively). In addition to the functional morpheme *ka*, the artificial language contained 12 nonsense words adapted from Ruskin (2014), all of which were mono- or bi-syllabic words that conformed to English phonotactics.

In both conditions, all 12 words were attested in the X_1 position during language exposure (Table 4.1). Crucially, I manipulated the number of words that were also attested in the X_2 position, ensuring there was sufficient evidence for structural substitutability $X_1 \mapsto X_2$ in the Productive condition (10 of the 12 words attested in X_2) but not in the Unproductive condition (6 of the 12 words attested in X_2). I selected 10 of 12 in the Productive condition and 6 of 12 in the Unproductive condition because these values are consistent with productivity (or lack of productivity in the Unproductive condition) according to the TSP, which permits at most 4 exceptions ($12/ln12 = 4.83$), meaning at least 8 of our 12 words must also be attested in X_2 position for the rule of substitutability to be generalizable.

The exposure set was also constructed such that some words were more frequent than others in order to imitate word frequency in natural language input (e.g., Zipf, 1949). To keep the two conditions balanced, I kept the total token frequency of each word the same across the two conditions, and ensured the most frequent word was attested in both the X_1 and X_2 positions in both conditions. I also ensured that the words that did not occur in the X_2 position included both high and low token frequency words. The distribution of the words and their frequencies across conditions and positions in the exposure set are shown in Table 4.1.

The test strings were generated to include either one (X_1-ka-X_2) or two levels (X_1-ka-X_2-ka-X_3) of embedding. At each level, there were three types of test strings: attested, unattested, and ungrammatical, where attested/unattested means whether the words have been attested at the specific position during exposure. Attested strings were strings or combinations of two strings that had been heard during exposure (i.e., were part of the exposure set). For example, as shown in Table 4.2, for a one-level string, it means the exact string (e.g., 'waso-ka-mito') has been heard during

Word	Frequency	Unproductive		Productive	
		X_1	X_2	X_1	X_2
nogi	36	6	30	12	24
sane	10	10	0	10	0
tesa	6	6	0	3	3
waso	6	6	0	3	3
sito	6	2	4	3	3
kosi	6	2	4	3	3
mito	4	2	2	2	2
kewa	4	2	2	2	2
bila	4	2	2	2	2
seta	2	2	0	1	1
sasa	2	2	0	1	1
tana	2	2	0	2	0
Total	88	44	44	44	44

Table 4.1: The distribution of words in the exposure corpus and word frequency in X_1/X_2 position in Exp 1.

exposure phase; for a two-level string, it means both components (e.g., 'sane-ka-kewa' and 'kewa-ka-nogi') have been heard. Therefore, all the words have been attested in those positions in relation to *ka* during exposure. In unattested strings, the post-*ka* positions (X_2 or X_3) were occupied by a word that never appeared in X_2 position during exposure. Thus, in the unattested strings in Table 4.2, 'sane', 'tesa' and 'tana' have never been attested after *ka*. Finally, ungrammatical strings were strings with wrong word order, such as ka-X_1-X_2 or ka-X_1-X_2-X_3-ka. There were six test strings of each type at each level, leading to 36 test strings in total. I designed the test strings such that in each string type, there were both words of higher frequency and words of lower frequency, in order to avoid the influence of token frequency in the test.

Type	One-level	Two-level
attested	*waso-ka-mito*	*sane-ka-kewa-ka-nogi*
unattested	*nogi-ka-sane*	*waso-ka-tesa-ka-tana*
ungrammatical	*ka-bila-kosi*	*ka-waso-kosi-sito-ka*

Table 4.2: Sample test strings in Unproductive condition in Exp 1.

All exposure and test strings were generated by a female voice using an online speech synthesizer, NaturalReader. Each unique string was generated separately such that all strings were generated with the same speed, volume, and pitch.

4.2.1.3 Procedure

The experiment consisted of two phases: exposure, in which participants were exposed to the artificial language, and test, in which participants were tested on how well they learned the language and whether they formed a productive generalization. In the exposure phase, participants were told they would hear strings from a new language, and that they need to pay careful attention to the strings, because they would be tested on their knowledge of the language later. During exposure, participants heard two repetitions of the exposure corpus (44 X_1-ka-X_2 strings) presented in random order as they viewed a still, unrelated nature scene (i.e., there was no accompanying referential world). There was 1.5s of silence between each string, and participants were offered a break after each repetition of the 44 strings to prevent task fatigue. In order to make sure that the participants were paying attention, other sounds were randomly dispersed among the linguistic strings, such as bird chirping sounds, and participants were later asked how many such sounds they heard. The random sounds occurred only rarely so as not to interfere with the learning of the language (i.e., 2 or 3 times per block). All participants answered those questions correctly.

Once the exposure phase was completed, the test phase began. On each test trial, participants heard a test string, and were asked to rate the acceptability of the string on a scale of 1 to 5. Participants were told to decide if those strings came from the language they had just heard (e.g., whether they think a native speaker of the language would have said that particular string). 1 meant the string was definitely not from the language; 2 meant the string may not have come from the language; 3 meant the string may or may not have come from the language; 4 meant the string may have come from the language; 5 meant the string definitely came from the language. The test strings were delivered in random order.

In both conditions, participants are expected to rate attested strings higher than ungrammatical strings at both levels. Of particular interest are the unattested strings. According to the proposal, only participants in the Productive condition would learn that $X_1 \mapsto X_2$ is productive in the X_1-ka-X_2 structure, and would thus generalize this pattern to unattested words: If a word appeared in position X_1 during exposure, it would be able to appear in position X_2 as well, even though it was never attested there in the input. On the other hand, $X_1 \mapsto X_2$ is not productive in the Unproductive

condition: for words that only appeared in position X_1, participants would be more likely to assume that those words cannot appear in position X_2. Therefore, it is predicted that participants in the Productive condition would rate one-level unattested strings higher than participants in the Unproductive condition. Furthermore, given the productivity at level-1, participants in the Productive condition would acquire the generalization that $X_1 \mapsto X_2$ holds for any level, so all of the 12 words can be used in both X_1 and X_2 positions to create recursive embedding; but for participants in the Unproductive condition, the words unattested in X_2 position are not supposed to be able to appear after *ka* at any level. Thus, participants in the Productive condition are predicted to rate two-level unattested strings higher than participants in the Unproductive condition as well.

4.2.2 Results

4.2.2.1 Learning

To capture how well participants learned their input language, I calculated a learning index for each participant. I took the difference score of a participant's mean response on Attested test strings minus their mean response on Ungrammatical test strings (see (40)). I calculated this index separately for one-level and two-level test strings. For one-level test strings, a positive learning index would suggest that a participant rated X_1-ka-X_2 strings they heard during exposure (Attested) as more consistent with the language than ka-X_1-X_2 strings, which violated the structure of the input grammar (Ungrammatical). For two-level strings, a positive learning index would suggest that a participant rated two-level strings whose post-*ka* positions (X_2 and X_3) were occupied by words attested in X_2 position during exposure (Attested) as more consistent with the input language than two-level strings with the *ka* morpheme in the wrong position (Ungrammatical, e.g., ka-X_1-X_2-X_3-ka).

(40) Learning index = $M_{attested}$ - $M_{ungrammatical}$

Figure 4.1 shows the mean learning index by input condition and embedding level. As shown in the figure, participants in both conditions not only learned the grammar (had a positive learning index on one-level sentences), but also endorsed two-level embedding for words attested in both

X_1 and X_2 position during exposure. We analyzed the results using mixed effects regression: The dependent variable was the learning index; fixed effects included Condition (Productive vs. Unproductive) and Embedding Level (1-level vs. 2-level) (in a two-way interaction). All the categorical predictors were simple coded. Participant was included as random intercept to account for by-participant variance. The model showed no significant main effect of Condition ($\chi^2(1) = 0.49$, $p = 0.48$) or Level ($\chi^2(1) = 0.51$, $p = 0.48$), indicating participants in both conditions learn the grammar equally well. However, there is a significant interaction between Condition and Level ($\chi^2(1) = 9.50$, $p = 0.002$): Participants in the Unproductive condition rated two-level strings significantly higher ($\beta = 0.74$, $SE = 0.23$, $t = 3.17$, $p = 0.003$). Post-hoc analyses confirmed that while there is no significant difference between two conditions at level-one ($\beta = 0.252$, $SE = 0.21$, $t = 1.20$, $p = 0.23$), the learning index in the Productive condition was significantly lower than that in the Unproductive condition at level-two ($\beta = -0.49$, $SE = 0.21$, $t = -2.34$, $p = 0.02$), suggesting that participants were more willing to endorse two-level recursion for attested sentences in the Unproductive condition. This may not be surprising, given that fewer X words are allowed in both positions in that condition (6 of 12) compared to the Productive condition (10 of 12), making the pattern easier to memorize and learn. Therefore, overall, the results on the learning index indicate the participants in both conditions have learned the basic pattern of the artificial grammar.

4.2.2.2 Generalization

To determine whether participants formed a productive generalization permitting words attested in X_1 position to also appear in X_2 position, I also calculated a generalization index for each participant. Here, I took the difference score of a participant's mean response on unattested test strings minus their mean response on ungrammatical test strings (see (41)). As with the learning index, I calculated the generalization index separately for one- and two-level test strings. At both levels of embedding, a positive generalization index would suggest that a participant rated unattested strings (whose post-*ka* positions, X_2 and X_3, were occupied by words never attested in X_2 position during exposure) as more consistent with the language than ungrammatical strings that violated the structure of the input grammar.

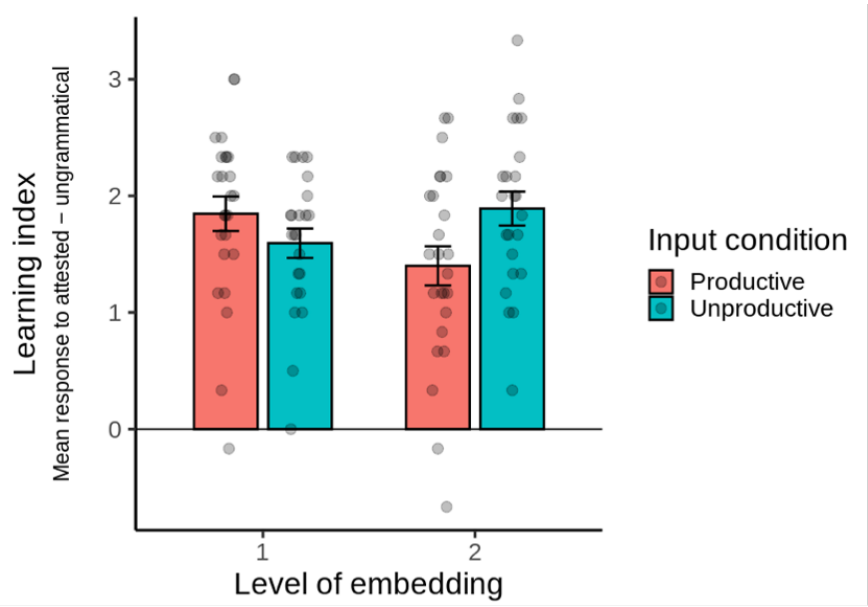

Figure 4.1: Effects of input condition on learning at each embedding level in Exp 1. Learning index is the difference score of each participant's mean response to attested - ungrammatical test sentences. Dots are individual participants and error bars are standard error.

(41) Generalization index = $M_{unattested}$ - $M_{ungrammatical}$

Figure 4.2 shows the mean generalization index by input condition and embedding level. A mixed effects regression model showed there is a significant main effect of Condition ($\chi^2(1) = 10.07$, $p = 0.002$), which indicates that participants in the Unproductive condition generalized significantly less (β = -0.56, $SE = 0.17$, t = -3.28, $p = 0.002$). Post-hoc analyses further showed that at both levels, the generalization index in the Productive condition is significantly higher than that in the Unproductive condition (level-one: $\beta = 0.58$, $SE = 0.22$, $t = 2.60$, $p = 0.01$; level-two: $\beta = 0.53$, $SE = 0.22$, $t = 0.38$, $p = 0.02$). There is also a significant main effect of Level ($\chi^2(1) = 7.41$, $p = 0.006$), indicating that participants were significantly less likely to generalize at two levels of embedding (β = -0.40, $SE = 0.15$, t = -2.76, $p = 0.008$). There is no significant interaction between Condition and Level ($\chi^2(1) = 0.03$, $p = 0.86$). Therefore, the results suggest that as predicted, participants generalized more in the Productive condition than in the Unproductive condition at both levels of embedding. This supports the proposal that speakers can use one-level distributional information to learn about recursive structures. However, in both conditions, they were less likely

to generalize for two-level strings. In the next section, I will discuss this pattern of results in more detail and explore how it relates to findings from natural language.

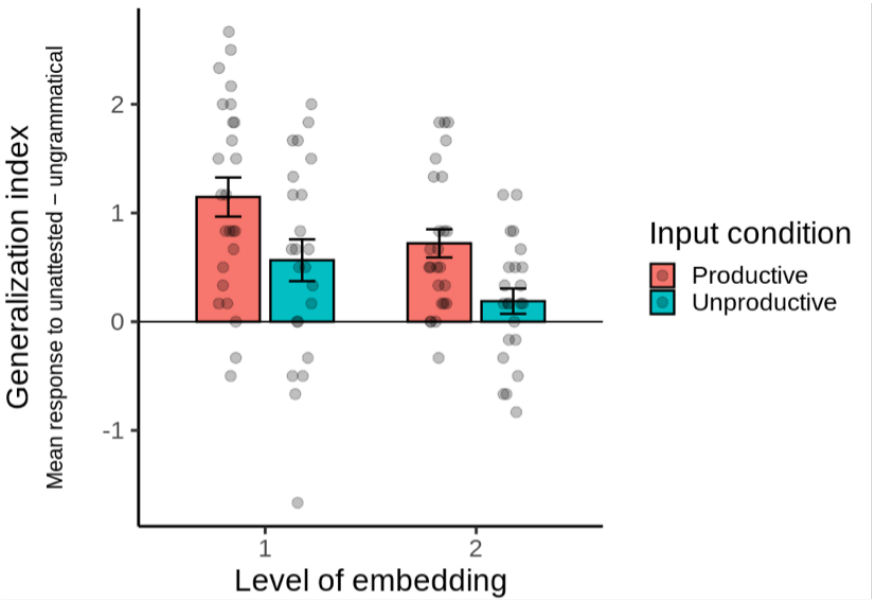

Figure 4.2: Effects of input condition on generalization at each embedding level in Exp 1. Generalization index is the difference score of each participant's mean response to unattested - ungrammatical test sentences. Dots are individual participants and error bars are standard error.

4.2.3 Discussion

This experiment investigated whether speakers can learn recursive structures purely based on the productivity of structural substitutability in simple one-level embedding data. The proposal suggests that, for a structure such as X_1-ka-X_2, if a large enough proportion of words are attested in both the X_1 and X_2 positions in one-level input, then speakers can acquire the generalization that the two positions are productively substitutable. This means if a word is attested in one position, then it is able to appear in the other position as well, even though it has never been attested in the other position in the input. Furthermore, if productivity can be maintained at level one, speakers will learn that the structure can be embedded to any level. In contrast, if the number of words attested in both positions in the input does not reach the productivity threshold, speakers will assume the positions are not productively substitutable and thus the structure cannot be embedded further, except for specific items that have been attested in both positions in the input. It is found

that as predicted, participants exposed to productive input were significantly more willing to generalize to unattested strings at both one- and two-levels than participants exposed to unproductive input. Therefore, the results suggest that learners can indeed access the distributional information and perform the necessary computations as the proposal predicts. Together with previous corpus studies which demonstrated the availability and reliability of distributional information on structural productivity in naturalistic data (Grohe et al., 2021; Li et al., 2021), the findings support the proposal that the recursivity of a structure can be learned distributionally from language-specific one-level experience.

The findings highlights the role of formal properties in language acquisition: Participants were able to learn important rules for recursive embedding through distributional information alone. With this claim, we do not intend to suggest that distributional cues are the only kind of information that is helpful during language acquisition or that distributional learning can give learners everything. In natural language acquisition, we agree that other factors such as semantic, pragmatic and phonetic cues will also play a role in the acquisition of recursive structures (e.g., Rosenbach, 2014), and that future research should investigate how these different cues are exploited and coordinated. On the other hand, our results do suggest that purely distributional information can be very helpful, even when other cues may not be accessible or completely reliable (Maratsos and Chalkley, 1980; Braine, 1987).

One apparent difference between the proposal and the current results is that while the proposal predicts that learners will learn that infinite embedding is allowed once there is sufficient evidence for substitutability in one-level input, our participants were less likely to generalize at two-level even in the Productive condition. Indeed, we agree that in principle, the account would predict a categorical difference in linguistic knowledge: The unattested strings of both embedding levels in the Productive condition should be completely good, while the unattested strings of both embedding levels in the Unproductive condition should be completely bad. However, while the proposal predicts perfect linguistic ability, participants' judgement in experiments are naturally imperfect, and influenced by processing factors. Indeed, even experiments with natural language have found that native speakers experience difficulty processing grammatical but recursively embedded structures, and

their ratings for the structures get lower with increasing levels of embeddings. For instance, in Christianson and MacDonald (2009)'s study, participants rated different recursive structures, such as PP's, possessives and central embeddings, and for all the structures, deeper embeddings were rated significantly worse. Further, the pattern to be learned in the study is complex, and the duration of the exposure phase is brief. That is, our participants are new learners of the artificial language. As such we did not expect our learners to be perfect generalizers, even in the Productive condition. Instead, the crucial finding is that as predicted by the distribution learning proposal, participants in the productive condition do generalize to both one- and two-level sentences, and they do so significantly more strongly than those in the Unproductive condition. Future studies can try different tasks such as production or forced alternative choice tasks to further investigate the nature of learners' linguistic knowledge.

An remaining open question from this experiment concerns the role of structural representation. Recall that an important prerequisite of the proposal is that the substitutable element must be the head of the structure, because only in that case will the structure involve self-embedding, which is the definition of recursion. For example, in English NP-V-NP structures like 'dogs chase cats', the two NPs could be substitutable, but that will not lead to recursion such as '*dogs chase cats chase rats', because the substitutability has nothing to do with the head. So can learners identify the head through distributional learning, and integrate knowledge of the head and substitutability to learn recursive structures? That is, will they only learn recursive structures when the substitutable element is the head, but refrain from recursion when it is not even in the presence of productive substitutability? To examine these questions, I conducted Experiment 2.

4.3 Experiment 2

Experiment 2 tests the role of structural representation in the distributional learning of recursive structures. To preview, I design two new artificial languages that can form A_1-B-A_2 strings, and A_1 and A_2 positions are productively substitutable in both languages; however, the head of the structure is A_2 in one language and is B in the other. Participants were exposed to input where the distributional information indicates both the head and the substitutability, and it is found that

as predicted, participants who learned the A-head language were more willing to accept recursively embedded strings (A_1-B-A_2-B-A_3) although participants from both conditions have learned substitutability at level-one.

4.3.1 Methods

4.3.1.1 Participants

Participants were 50 adult native English speakers with typical hearing and vision (or corrected vision). All participants were recruited and run online via Prolific Academic (www.prolific.com) and paid 9 dollars/hour as compensation. The 50 participants were evenly assigned to two language conditions, A-head language (age = 31.2, range = 20-46) and B-head language (age = 29.5, range = 20-45).

4.3.1.2 Stimuli

The A-head language and B-head language were constructed such that they can both form one-level A_1-B-A_2 strings, but the head of the A_1-B-A_2 string is different in the two languages: it is A, in particular A_2, in the A-head language, and is B in the other language (Figure 4.3). As in other artificial language experiments (e.g., Fetch, 2020), I approximated the distributional character of heads by implementing the rules that the head of the phrase obligatorily appears whenever the phrase is present, and that non-head elements are optional.[1] Therefore, the two languages allow different linear strings as shown in Table 4.3: For one-word strings, the single word must be the head; for two-word strings, the hierarchy and the head determined that AA and BA but not AB are possible in the A-head language, whereas the B-head language allows BA and AB but not AA.

There are 12 category-A words and 1 category-B word in each language. To help participants learn the distinction between head and non-head elements, I used bi-syllabic words for the head and mono-syllabic words for the non-head. The nonsense words were also adapted from Ruskin

[1]By implementing these rules, I do not mean any non-head element in any language must be omittable. There could be language specific rules that complicates those fundamental rules. Neither do I mean those are the only cues in natural languages for learners to identify the head. I choose those rules because they are key features that define heads in theoretical work on natural languages, and they have been proven useful for learners to identify the head in distributional learning studies.

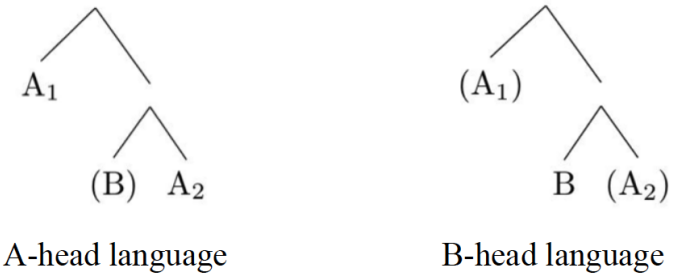

Figure 4.3: Structural representation of two languages in Exp 2.

	A-head language	B-head language
one-word	A, *B	*A, B
two-word	AA, BA, *AB	*AA, BA, AB

Table 4.3: Grammaticality of one-word and two-word strings in two languages in Exp 2.

(2014) and are provided in Table 4.4 and 4.5. Similar to the Productive condition in Experiment 1, $A_1 \mapsto A_2$ is productive in the A_1-B-A_2 structure in both languages in the current experiment: All of the 12 different words were attested in A_1 position, and 9 of them were attested in A_2 position, clearing the productivity threshold predicted by the TSP ($12/ln12 = 4.83$). If learners indeed use distributional information to learn recursive structures as predicted by the proposal, i.e., they learn a structure can be recursively embedded if the head positions are substitutable, then they should license recursion (A_1-B-A_2-B-A_3) in the A-head language but not in the B-head language.

For both languages, I constructed a 144-string exposure corpus (Table 4.4-4.5). Specifically, there were three different types of exposure strings: one-word strings, two-word strings, and A_1-B-A_2 strings. The one-word string for the B-head language was the single category-B word, which was repeated 36 times; for the A-head language, that included each of the 12 category-A words repeated 3 times, thus also making 36 strings in total. The two-word strings for the A-head language included AA strings and BA strings, and for the B-head language included AB strings and BA strings. For both languages, the BA strings consisted of 3 repetitions of each of the 12 category-A words following the category-B word. The AB strings in the B-head languages consisted of 3 repetitions of each of the 12 category-A words preceding the category-B word. The 36 AA strings in the A-head language were selected from all the possible AA combinations such that each category-A word appeared in A_1 position three times and appeared in A_2 position three times and

A_1 and A_2 were not occupied by the same word. There were also 36 A_1-B-A_2 strings for each language. They were selected from all possible A_1-B-A_2 combinations such that all 12 category-A words were attested in A_1 position 3 times; 9 category-A words were attested in A_2 position 4 times; A_1 and A_2 were occupied by two different words. The 144 exposure strings were divided into three blocks: each block contained 12 one-word strings, 12 AA or AB strings, 12 BA strings, and 12 A_1-B-A_2 strings. The frequency of each word was balanced across three blocks.

Word	Frequency	One-word	Two-word			A_1-B-A_2	
		A	A_1 in AA	A_2 in AA	A in BA	A_1	A_2
nogi	19	3	3	3	3	3	4
tesa	19	3	3	3	3	3	4
waso	19	3	3	3	3	3	4
mito	19	3	3	3	3	3	4
bila	19	3	3	3	3	3	4
sane	19	3	3	3	3	3	4
sito	19	3	3	3	3	3	4
kosi	19	3	3	3	3	3	4
kewa	19	3	3	3	3	3	4
seta	15	3	3	3	3	3	0
sasa	15	3	3	3	3	3	0
tana	15	3	3	3	3	3	0

Table 4.4: The distribution of category-A words in the exposure corpus and word frequency in each position in the A-head language from Exp 2. The category-B word in the A-head language is *ka*.

Word	Frequency	One-word	Two-word		A_1-B-A_2	
		B	A in AB	A in BA	A_1	A_2
ka	13	0	3	3	3	4
bo	13	0	3	3	3	4
ru	13	0	3	3	3	4
ni	13	0	3	3	3	4
fei	13	0	3	3	3	4
pao	13	0	3	3	3	4
sa	13	0	3	3	3	4
mo	13	0	3	3	3	4
gu	13	0	3	3	3	4
di	9	0	3	3	3	0
tei	9	0	3	3	3	0
lao	9	0	3	3	3	0

Table 4.5: The distribution of category-A words in the exposure corpus and word frequency in each position in the B-head language from Exp 2. The category-B word in the B-head language is *nogi*.

The test strings included two-word strings to test participants' knowledge of the head (hereby zero-level strings), one-level strings (A_1-B-A_2) to test their knowledge of substitutability, and two-level strings (A_1-B-A_2-B-A_3) to test their knowledge of recursion (Table 4.6). For zero-level strings, participants in both conditions would be tested on 6 AB strings, 6 BA strings, and 6 AA strings. All the category-A words used for those two-word test strings were among the 9 category-A words that were attested in both A_1 and A_2 positions in A_1-B-A_2. For participants in each condition, two types of zero-level strings would be attested, one type would be ungrammatical. For example, for participants in the A-head language condition, as shown in Table 4.3, AA strings (e.g., 'sito-mito') and BA strings (e.g., 'ka-kewa') were attested, while AB strings (e.g., 'tesa-ka') would be ungrammatical. The design of one- and two-level test strings were similar to that in Experiment 1. For both, there were attested strings, unattested strings, and ungrammatical strings. For one-level strings, the attested strings were strings that had been heard during exposure phase. For example, in Table 4.6, 'nogi-ka-mito' is a string selected from the exposure corpus. Unattested strings were strings where the A_2 position was occupied a word that was never attested in A_2 position during exposure. For instance, 'tana' has never appeared after *ka* in the A_1-B-A_2 structure in the exposure corpus. Ungrammatical strings had the word order A_1-A_2-B, which was not allowed in either of the two languages. For two-level strings, the attested strings were combinations of two one-levels strings that were attested during exposure phase: e.g., 'nogi-ka-mito' and 'mito-ka-tesa' are both selected from the exposure corpus. In unattested strings, the A_2 and A_3 positions were filled by words that never appeared after the category-B word during exposure, such as 'seta' and 'sasa' in Table 4.6. The ungrammatical strings were A_1-A_2-A_3-B-B strings, which were impossible in both grammars. There were 6 strings of each type at each level, leading to 54 test strings in total.

Type	Zero-level	One-level	Two-level
attested	*sito-mito, ka-kewa*	*nogi-ka-mito*	*nogi-ka-mito-ka-tesa*
unattested		*bila-ka-tana*	*waso-ka-seta-ka-sasa*
ungrammatical	*tesa-ka*	*nogi-tesa-ka*	*nogi-waso-bila-ka-ka*

Table 4.6: Sample test strings in A-head language condition in Exp 2.

All exposure and test strings were generated by the same speech synthesizer as Experiment 1, using the same voice and speed.

4.3.1.3 Procedure

The procedure was the same as that of Experiment 1, except that there were more blocks in the exposure phase: Participants were offered a break after hearing each block of 36 strings, which were delivered in random order; and after all of the three blocks have been played, they were all repeated one more time since our pilot studies showed that one repetition was not enough for participants to learn the language. Therefore, there were six exposure blocks in total.

4.3.2 Results

This section presents the results for zero-level, one-level and two-level test strings. For one- and two-level strings, as in Experiment 1, I computed a learning index and a generalization index to measure how much the participants learned and generalized. The predictions are as below. First, for zero-level strings, if participants have learned the headedness, then they are predicted to rate attested strings in their language (i.e., AA and BA in A-head language, AB and BA in B-head language) significantly higher than the ungrammatical strings (i.e., AB in A-head language, AA in B-head language). Next, for one-level strings, participants in both conditions should learn the A_1-B-A_2 structure and the substitutability of A_1 and A_2. Therefore, there should be no difference between conditions in either the learning index or the generalization index. In contrast, for two-level strings, participants from the A-head language condition are predicted to rate both attested and unattested strings higher than participants from the B-head language condition: although participants from the B-head language conditions have heard examples analogous to 'dogs chase cats' and 'cats chase rats', they would not be willing to accept 'dogs chase cats chase rats' because of the headednesss; neither would they be willing to allow recursion for unattested words although they have learned substitutability.

4.3.2.1 Zero-level

The zero-level data are shown in Table 4.7. A mixed-effects regression model showed a significant main effect of test string Type (attested vs. ungrammatical) ($\chi^2(1) = 587.42$, $p < 0.001$): Ungrammatical strings were rated significantly lower than attested strings (β = -2.20, $SE = 0.06$, $t =$

-34.19, $p < 0.001$). Post-hoc analyses suggested this holds true in both language conditions (A-head language: $\beta = 1.04$, $SE = 0.09$, $t = 11.41$, $p < 0.001$; B-head language: $\beta = 3.37$, $SE = 0.09$, $t = 36.94$, $p < 0.001$). Therefore, people from both conditions have learned the headedness of the language. The model also revealed a significant main effect of Condition ($\chi^2(1) = 4.02$, $p = 0.045$) and a significant interaction between Condition and Type ($\chi^2(1) = 276.43$, $p < 0.001$), indicating that ungrammatical strings were rated higher in the A-head language condition ($\beta = 2.33$, $SE = 0.13$, $t = 18.05$, $p < 0.001$). Post-hoc analyses confirmed that participants in the B-head language condition rated ungrammatical strings (i.e., AA) lower than those in the A-head language condition (i.e., AB) ($\beta = -1.83$, $SE = 0.16$, $t = -11.26$, $p < 0.001$), which led to the significant main effect of Condition. This result was expected, because the ungrammatical strings in the B-head language were indeed worse than those in the A-head language: The ungrammatical strings in the B-head language condition were AA, which did not contain the head at all; in contrast, the ungrammatical strings in the A-head language condition, AB, did contain a category A-word, so it is reasonable that participants from the B-head language condition would rate their ungrammatical strings lower. Post-hoc analyses also showed that grammatical strings in the B-head language condition (i.e., AB and BA) were rated higher than those in the A-head language condition (i.e., AA and BA) ($\beta = 0.50$, $SE = 0.14$, $t = 3.47$, $p = 0.001$). Further examination showed that this can be attributed to AA strings: while BA strings in both conditions and AB strings in the B-head language condition were rated similarly, AA strings in the A-head language condition were rated lower than them. This can be explained by the fact that while all combinations for AB and BA strings were attested during the exposure phase, there were many possible AA strings (144) and only a small proportion of them (36) were selected for the exposure corpus. Therefore, it is expected that participants would be less certain about the judgments of AA strings and rate them lower than the other grammatical strings. Nevertheless, it is evident that participants in both conditions can distinguish ungrammatical strings from attested strings in their language, exhibiting knowledge of the head.

Condition	AA	AB	BA
A-head	3.88 (0.27)	*3.11 (0.22)*	4.43 (0.20)
B-head	*1.29 (0.15)*	4.75(0.15)	4.56 (0.18)

Table 4.7: Mean rating scores for zero-level test strings in Exp 2. Standard errors are in parentheses. Ungrammatical strings are in italics.

4.3.2.2 One-level

Figure 4.4 shows the results for one-level test strings. For one-level and two-level strings, I analyzed the results using mixed effects regression. The dependent variable was the index. Fixed effects included Condition (A-head vs. B-head) and test Type (Learning vs. Generalization) (in a two-way interaction). All the categorical predictors were simple coded. Participant was included as random intercept to account for by-participant variance. For one-level strings, mixed-effects regression demonstrated that test Type (Learning vs. Generalization) ($\chi^2(1) = 9.38$, $p = 0.002$) but neither Condition ($\chi^2(1) = 0.72$, $p = 0.40$) nor the interaction of Type and Condition ($\chi^2(1) = 1.10$, $p = 0.29$) was a significant predictor of the index. In specific, post-hoc analyses revealed that there is no significant difference between two conditions for either the learning index (β = -0.13, SE = 0.27, t = -0.48, $p = 0.63$) or the generalization index (β = -0.29, $SE = 0.27$, t = -1.11, $p = 0.27$). Therefore, this suggests that participants in both conditions have learned the A_1-B-A_2 structure, and have generalized the rule of substitutability of A_1 and A_2 to similar extent. The significant main effect of test Type showed that the generalization index was generally lower than the learning index, suggesting participants were more willing to accept attested strings than unattested strings.

4.3.2.3 Two-level

Results at level-two are shown in Figure 4.5. Mixed-effects regression showed that both Condition ($\chi^2(1) = 5.04$, $p = 0.025$) and test Type ($\chi^2(1) = 12.46$, $p < 0.001$) but not their interaction ($\chi^2(1)$ = 1.66, $p = 0.20$) were significant predictors of the index. The significant main effect of Condition suggested that the learning indices were higher in the A-head language condition than in the B-head language condition ($\beta = 0.74$, $SE = 0.33$, $t = 2.26$, $p = 0.03$). In particular, post-hoc analyses confirmed that the learning index in the B-head language was marginally significantly lower than that in the A-head language (β = -0.62, $SE = 0.34$, t = -1.81, $p = 0.0765$), and the generalization

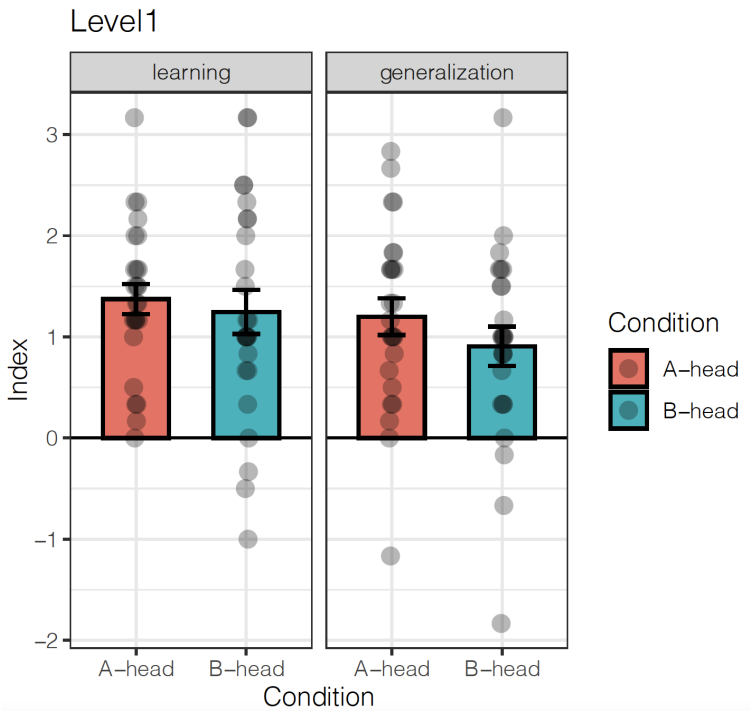

Figure 4.4: Effects of input condition on learning and generalization at level one in Exp 2. Dots are individual participants and error bars are standard error.

index was significantly lower (β = -0.87, SE = 0.34, t = -2.52, p = 0.01). Therefore, although participants from both conditions have learned substitutability in one-level strings, participants from the A-head condition were more willing to accept recursively embedded strings for both attested and unattested words. Finally, similar to one-level data, there is also a significant main effect of Type, suggesting the generalization index was lower than the learning index, though post-hoc analyses reported that the learning and generalization index only differed significantly in the B-head language condition ($\beta = 0.49$, $SE = 0.14$, $t = 3.55$, $p < 0.001$).

4.3.3 Discussion

In summary, Experiment 2 investigated how learners use distributional information about the productivity of structural substitutability to learn recursive structures based on syntactic knowledge of the head of the structure. According to the proposal, a prerequisite for structural substitutability to lead to recursion is that the substitutable element must be the head of the structure, because only in that case will self-embedding be involved, which is the definition of recursion. To test

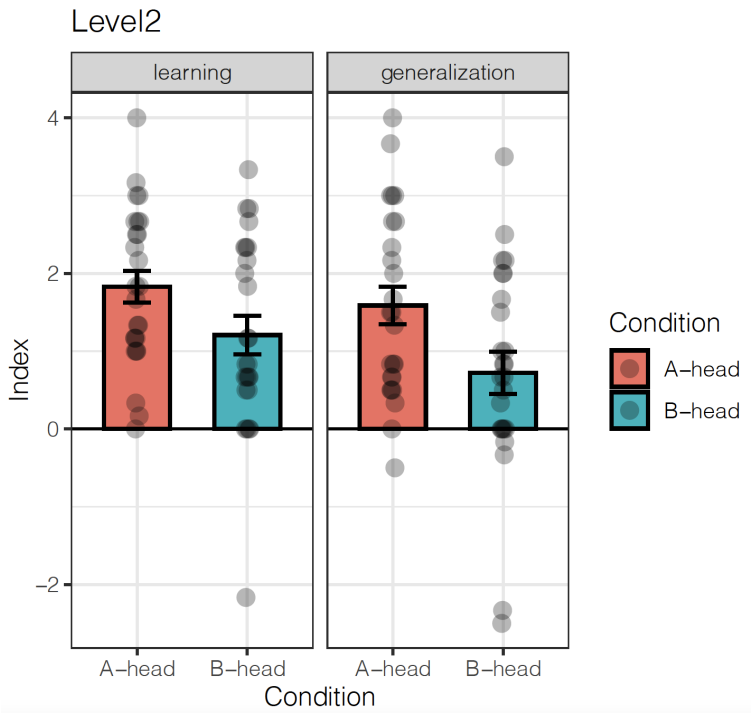

Figure 4.5: Effects of input condition on learning and generalization at level two in Exp 2. Dots are individual participants and error bars are standard error.

the proposal, I exposed participants to two different artificial languages. In both languages, the A_1 and A_2 positions in A_1-B-A_2 are productively substitutable, but participants could also learn from distributional cues that the head of the structure is A in one language but is B in the other language. At test, it is found that as predicted, although participants in both conditions learned substitutability and generalized the rule to unattested words in one-level strings, in the B-head language condition, where recursion was not expected, participants were significantly less likely to accept embedded strings for either attested words or unattested words, thus indicating that learners can integrate knowledge of the syntactic structure to distributionally acquire recursion.

The current results showed that learners can use purely distributional information to learn the head of a linguistic structure, and integrate this knowledge with other distributional information to acquire complex generalizations such as recursion. This finding adds to a body of work that investigates how distributional information can be utilized to acquire higher-order linguistic representations (e.g., Thompson and Newport, 2007; Takahashi and Lidz, 2008; Reeder et al., 2013;

Schuler et al., 2017; Fetch, 2020). By emphasizing the role of formal learning, though, this work does not intend to deny the role of other factors in learning the head and learning recursion, and more generally, in learning linguistic generalizations. And as I pointed out earlier, there may be other distributional cues for the head in natural languages in addition to the ones applied in the current design. The present study only focuses on the role of specific distribution information, and investigates what can be learned from such information alone.

4.4 Experiment 3

Experiment 1 and 2 have shown adult participants' ability to learn recursive structures from distributional information. However, an important open question is whether younger learners can also fully utilize such distributional information. On one hand, many studies have shown that young children and even infants can acquire linguistic knowledge using statistical information in similar ways (e.g., Saffran et al., 1996; Marcus et al., 1999; Shi and Emond, 2023); some have even argued that distributional learning is an ability available from birth (e.g., Gervain et al., 2008; Teinonen et al., 2009; Aslin, 2017). On the other hand, still in the process of development, children are much more limited than adults in a range of cognitive functions such as memory and processing abilities (e.g., Thiessen, 2011; Santolin and Saffran, 2018), and their learning outcomes often differ from adults in first and second language acquisition (Johnson and Newport, 1989; Newport, 1990; Mayberry and Kluender, 2018) as well as in artificial language learning experiments (Weir, 1964; Hudson Kam and Newport, 2005; Austin et al., 2022). Therefore, it is crucial to examine whether young learners exploit the subtle distributional cues in the same way as adults. In Experiment 3, we set to test children's distributional learning of recursive structures in an artificial language.

4.4.1 Methods

4.4.1.1 Participants

Participants were 17 children aged 6 to 8 years.[2] We chose this age range because prior work using similar paradigms has shown that children of this age were able to use purely distributional cues

[2]We plan for 20 children in each condition.

to acquire linguistic rules in artificial language learning experiments (e.g., Schuler et al., 2023). The children were all native English speakers. 10 of these children were assigned to the Productive condition and 7 to the Unproductive condition. An additional 14 children participated but were excluded from analysis based on our exclusion criteria that we will specify in the Results section.[3] In brief, they failed to learn the basics of the artificial grammar or understand the task. We recruited the children through sharing our advisement online. They participated in the experiment over Zoom and received a $10 Amazon gift card as compensation.

4.4.1.2 Stimuli

We generated stimuli from the artificial language in Experiment 1. In the language, strings are formed X_1-ka-X_2, and 9 X-category words were attested in the X_1 position. Crucially, we manipulated whether there was sufficient evidence to form a productive generalization of structural substitutability. In the Productive condition, 6 out of the 9 words were also attested in the X_2 position; in the Unproductive condition, only 4 of the 9 words were. Based on the TSP, 6 out of 9 is productive but 4 out of 9 is not. We decided to use a total of 9 words because our pilot studies showed that this amount of input is feasible for children to learn in a single-day artificial language learning experiment. A difference in the design from Experiment 1 is that the words in Experiment 3 followed a more uniform distribution instead of a Zipfian distribution, because given children's more limited attention span and memory buffer, it could be difficult for them to learn words with low token frequencies; additionally, previous studies have shown a Zipfian distribution is not necessary for the acquisition of productive rules in the lab (e.g., Shi and Emond, 2023). Therefore, there was a total of 54 unique strings in the Productive condition, and 36 in the Unproductive condition. The whole corpus was repeated twice during the exposure phase for the Productive condition and three times for the Unproductive condition so that all children heard 108 strings. The word distribution in each repetition is shown in Table 4.8.

Similar to Experiment 1, the test corpus consisted of strings that differed in embedding level

[3]The exclusion rate might seem high. However, a special property of the current experiment is that it depends on purely distributional learning with no semantic world. Therefore, it is more similar to infant studies, where exclusion rates have been similarly high (e.g., Aslin et al., 1998; Shi and Emond, 2023).

Word	Unproductive		Productive	
	X_1	X_2	X_1	X_2
kewa	4	9	6	9
tana	4	9	6	9
sito	4	9	6	9
bila	4	9	6	9
tesa	4	0	6	9
mito	4	0	6	9
nogi	4	0	6	0
seta	4	0	6	0
waso	4	0	6	0
Total	36	36	54	54

Table 4.8: The distribution of words in the exposure corpus and word frequency in X_1/X_2 position in Exp 3.

(One-level and Two-level) and Type (Attested, Unattested, and Ungrammatical). All exposure and test strings were generated by the same speech synthesizer as Experiment 1 and 2, using a slower speed.

Children in both conditions were predicted to rate attested strings high and ungrammatical strings low at both levels of embedding if they have learned the basic structure of the grammar (e.g., they know X_1-ka-X_2 strings are grammatical and ka-X_1-X_2 — strings with a completely wrong word order — are not). Crucially, the unattested test strings allow us to determine whether children have indeed formed the productive generalization that the X_1 and X_2 positions are substitutable (all the words in the X_1 position can also appear in the X_2 position). If children can use distributional cues to acquire recursive structures as predicted, we would expect that only children in the Productive condition would have sufficient evidence to learn that all words used in the X_1 position can also be used in the X_2 position (even though some were never attested in X_2 position in the input). Furthermore, given the productive substitutability in one-level data, only children in the Productive condition would be predicted to acquire the generalization that $X_1 \mapsto X_2$ holds for any embedding level to create recursive embedding. Therefore, unattested strings were predicted to be treated more similarly to attested strings at both embedding levels in the Productive condition, and more similarly to ungrammatical strings at both embedding levels in the Unproductive condition.

4.4.1.3 Procedure

As in Experiment 1 and 2, this experiment consisted of an exposure phase and a test phase; but the experiment was presented as a video game in Experiment 3 so that it would be more attractive to children. The video game paradigm was adapted from Schuler et al. (2023), which successfully worked for children in distributional learning experiments. At the beginning of the game, a cartoon robot explained that an alien, Zooma, was traveling to another planet, and that the goal of the game was to help Zooma learn the new language "Zilly" which is spoken on that planet. Next, there were some practice trials where children were asked to decide how well Zooma was speaking English using a slider scale from "no" to "yes" corresponding to values of 0 to 100, and an experimenter would provide feedback on how to use the slider scale. This phase was to familiarize the children with the task and the rating scale. We used this slider scale instead of a 5-point rating scale as in Experiment 1 and 2 because our pilots studies found that the slider scale worked better for children.

Then, during the exposure phase, children were instructed to listen carefully as Zooma practiced saying sentences in the new language. To keep children attentive, Zooma would get tired and stop practicing at several points in the exposure phase, and children were asked to click on Zooma to wake her up. Every time children clicked on Zooma, they would receive a star; when they collected enough stars, Zooma would progress in her journey to the new planet. A screenshot of the exposure phase is shown in Figure 4.6.

When children heard all the exposure strings, Zooma arrived at the distant planet, and the test phase began. On each test trial, Zooma said a test string, and children were asked to decide whether they heard this sentence during the exposure phase, using the same slider scale that they used in the practice trials. The one-level strings and two-level strings were presented in different blocks, with the one-level strings presented first. The level two instructions were modified slightly, to acknowledge that they would be longer: "Now Zooma will say sentences in Zilly that are longer. Some will be good. Some will be bad. Again, your job is to decide how well Zooma is speaking Zilly." A screenshot of the test phase is shown in Figure 4.7.

Figure 4.6: Screenshot of the exposure phase of Exp 3.

4.4.2 Results

We excluded children whose mean rating for attested strings was not higher than ungrammatical strings at any embedding level, which indicated not learning the fundamentals of the artificial grammar and/or not understanding the task. Results from the remaining children are shown in Figure 4.8. First, as expected, in both conditions and at both embedding levels, attested strings were rated high and ungrammatical strings were rated low. However, crucial for our prediction, the rating for unattested strings differed significantly across conditions. At both levels of embedding, unattested strings were rated higher in the Productive condition than the Unproductive condition.

We analyzed the results using mixed effects regression. The dependent variable was the rating score. Fixed effects included Condition (Productive vs. Unproductive), Type (attested, unattested, ungrammatical) and Level (1-level vs. 2-level) (in a three-way interaction). All the categorical predictors were simple coded. Participant was included as random intercept to account for by-participant variance. Hierarchical modeling showed that Type (attested, unattested, ungrammatical) ($\chi^2(2) = 186.54$, $p < 0.001$), Level ($\chi^2(1) = 7.00$, $p < 0.01$), the interaction between

Figure 4.7: Screenshot of the test phase of Exp 3.

Condition and Type ($\chi^2(2) = 22.68$, $p < 0.001$), and the interaction between Type and Level ($\chi^2(2) = 6.97$, $p = 0.03$) were significant predictors of the rating score, but not Condition or any other possible interactions between Condition, Type and Level. The statistics from the regression model are shown in Table 4.9. Crucially, as predicted, the unattested strings in the Unproductive condition were rated significantly lower, suggesting that children were less willing to allow unattested strings when there was not enough evidence for productive substitutability in non-embedded input.

Post-hoc comparisons further revealed that at one-level, the rating scores for attested, unattested and ungrammatical strings all differed from each other in the Productive condition ($p < 0.01$ for all comparisons), while in the Unproductive condition, the unattested and ungrammatical strings did not differ from each other ($\beta = 3.61$, $SE = 4.86$, $t = 0.74$, $p = 1.00$), and they were both rated lower than the attested strings ($p < 0.001$ for both comparisons); at two-level, in the Productive condition the unattested strings did not differ from the attested strings ($\beta = 4.93$, $SE = 4.60$, $t = 1.07$, $p = 1.00$), both rated higher than the ungrammatical strings ($p < 0.001$ for both comparisons), whereas in the Unproductive condition, the unattested strings patterned with

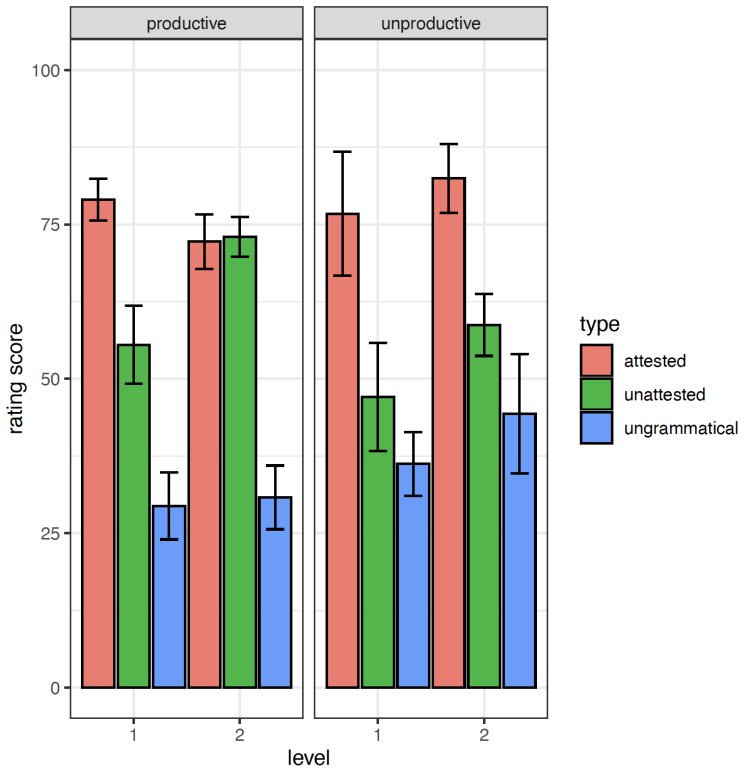

Figure 4.8: Mean rating scores by embedding level and test string type in each condition of Exp 3.

the ungrammatical strings ($\beta = 12.07$, $SE = 4.67$, $t = 2.58$, $p = 0.29$). These results align with what we see in Figure 4.8. Overall, the results showed that the unattested strings — strings that did not occur in the language input — were rated lower in the Unproductive condition than the Productive condition. In particular, although children never heard recursively embedded strings in the learning phase, children in the Productive condition judged two-level unattested strings to be similarly well-formed to attested strings; by contrast, children from the Unproductive condition regarded two-level unattested strings essentially as ungrammatical strings. Therefore, the results suggested that children can indeed use distributional cues to acquire recursive structures

4.4.3 Discussion

In Experiment 3, we asked whether children can use distributional cues from non-embedded examples to learn whether a structure allows recursive embedding. We have shown that adults can acquire recursive structures through distributional learning, but it is necessary to examine whether

Fixed effects	β	SE	t	p
(Intercept)	57.28	2.57	22.31	<0.001***
Condition - unproductive	1.05	5.14	0.20	0.84
Type - unattested	-18.73	2.78	-6.73	<0.001***
Type - ungrammatical	-39.23	2.86	-13.7-	<0.001***
Level - level-2	5.47	2.25	2.43	0.02*
Condition × Type - unproductive × unattested	-13.56	5.57	-2.44	0.02*
Condition × Type - unproductive × ungrammatical	10.73	5.73	1.87	0.06
Condition × Level - unproductive × level-2	4.44	4.50	0.99	0.32
Type × Level - unattested × level-2	12.19	5.57	2.10	0.03*
Type × Level - ungrammatical × level-2	5.36	5.73	0.94	0.35
Condition × Type × Level - unproductive × unattested × level-2	-21.72	11.13	-1.95	0.05
Condition × Type × Level - unproductive × ungrammatical × level-2	-3.31	11.46	-0.29	0.77

Table 4.9: Statistics from the regression model in Exp 3.

children can also do it in order to determine whether such a learning mechanism could be helpful during child language acquisition.

Through an artificial language learning experiment, we demonstrated that children can indeed acquire recursive structures using distributional information: Children who received sufficient input for structural substitutability rated unattested embedded strings higher than children from the other condition, although no recursively embedded strings were attested in their input. Importantly, children who received productive input rated unattested embedded strings as high as attested ones, while for children who received unproductive input, the unattested embedded strings patterned with ungrammatical ones. Overall, the results suggest that children are sensitive to distributional cues on structural substitutability in non-embedded data, and can use these cues to determine whether the structure permits recursive embedding. Taken together, this work demonstrated that the ability to track and utilize sophisticated and subtle distributional information to discover the underlying rules is available from an early age. Therefore, it could be a useful mechanism that helps children with the extremely challenging task of language acquisition.

While it has been found that both children and adults can acquire recursive structures through distributional learning, children's learning behavior in our experiment also showed some differences from adults' behavior in Experiment 1 and 2. Compared to adults, children exhibited categorical generalization, i.e., they rated unattested two-level sentences as high as attested ones in the Productive condition, and as low as ungrammatical ones in the Unproductive condition; whereas

for adults, although they did generalize more in the Productive condition, their rating scores for unattested two-level strings were still significantly lower than attested ones.

Though more work is needed to compare children and adults directly, this observation is perhaps unsurprising given other studies noting differences between children and adults generalization behavior. First, studies on the acquisition of regular rules have found that children are more likely to form categorical generalizations. For example, in Schuler et al. (2016)'s study, when participants learned a noun plural rule in an artificial language, almost all the children applied it to either all or none of the novel words during test depending on the productivity of their input; by contrast, adults from both conditions matched the token frequency of the plural markers from the input. Furthermore, in another line of work on learning from input that contains inconsistent use of grammatical forms, in both natural language acquisition and artificial language learning experiments, children have been observed to regularize, i.e., to produce these forms more consistently, whereas adults tend to reproduce the inconsistencies: For instance, when there is more than one grammatical form in variational use, adults closely reproduce the probabilistic patterns, while young children only use the most consistent form almost all the time (e.g., Singleton and Newport, 2004; Hudson Kam and Newport, 2005, 2009). In general, children seem to be particularly inclined to form categorical rules.

A remaining question from this experiment is children's structural representation of the artificial language: whether they treated the X word as the head. While the current design did not explicitly test this, evidence from our Experiment 2 suggests that it is likely that the children did treat the X word as the head, since the participants in Experiment 2 were more likely to allow recursive embedding in the head substitutability condition; and even if some children learned a linear structure, it is likely that they would also be able to learn a headed recursive structure with the same mechanism provided information for the head. In ongoing work, we are adapting Experiment 2 for children to explicitly test their learning behavior when they are provided with distributional information for both headedness and substitutability.

4.5 General discussion

In this chapter, we asked whether human learners can use distributional cues to learn syntactic generalizations. Previous chapters have shown that there are reliable distributional cues in child-directed speech for learners to acquire language-specific generalizations, but we argued that it is necessary to examine whether human learners can indeed make use of such distributional cues as predicted.

With a series of artificial language learning experiments, we demonstrated that both adults and children can use distributional cues on the productivity of substitutability to determine whether a structure allows recursive embedding; moreover, adults participants can integrate cues for the head and substitutability to avoid wrong generalizations. Overall, we demonstrate that learners of different ages are able to track and utilize sophisticated and subtle distributional information to discover the rules for recursion.

Together with our corpus studies which demonstrated the availability and reliability of distributional information for structural substitutability in naturalistic data, the findings from this chapter provide novel insights into the conceptualization and acquisition of recursion, a crucial concept in linguistics: The recursivity of a structure can be learned as a productive generalization distributionally from language-specific level-one experience. Therefore, this learning mechanism enables speakers to acquire knowledge of infinite embedding from finite input data, addressing an extreme case in learning generalizations.

The findings also deepen our understanding of distributional learning. While this learning mechanism has been demonstrated to work in the acquisition of many other aspects of linguistic knowledge, most of these studies focused on the acquisition of lower-level patterns such as word segmentation, with a few exceptions (e.g., Thompson and Newport, 2007; Wonnacott et al., 2008; Reeder et al., 2013; Schuler et al., 2016). This work shows that distributional learning can enable learners to acquire complex syntactic generalizations as well, adding to a body of work on how distributional information can be utilized to acquire higher-order linguistic representations.

A notable finding from this chapter is that participants are able to learn the generalizations from purely distributional cues with no semantic information at all. We do not intend to deny

the role of other cues during language acquisition, but our results do support an important role of distributional information. In future work, it would be worthwhile to examine whether learners can integrate different types of information, such as in cases comparable to the English possessives studied in Chapter 3, where learners need to learn further semantic rules based on formal analyses of distributional cues. Another direction is to examine whether this learning mechanism also works in other domains than language. As discussed in Chapter 3, the design feature of the current approach does not limit it to language. Since the experiments in this chapter have no referential world, they further suggests that the current learning model is independent of the nature of the hypotheses but has the potential of accounting for learning generalizations in different domains.

Another interesting finding in this chapter is the difference between adults and children. On one hand, contra some previous studies where adults' generalization behavior did not seem to follow the TSP (e.g., Schuler et al., 2016), the adult participants in our experiments did learn productive generalizations when the TSP predicted they should. This suggests that the TSP is not exclusive to children. Instead, children and adults may have similar learning algorithms. On the other hand, consistent with existing findings in the literature, our experiments showed that children seemed more inclined to learn categorical rules than adults. More work is needed to investigate the differences in children's and adults' generalization behavior.

Chapter 5

CONCLUSION

The goal of this dissertation was to investigate the learning mechanism of syntactic generalizations. I focused on the cases of verb argument structure and recursive structures, which have been known to pose series acquisition challenges. Based on the TSP, a cognitively plausible model that predicts a precise threshold for learning generalizations, I argued that these generalizations are learnable from distributional cues in simple child-directed speech. In this chapter, I summarize the findings of the previous chapters, discuss the implications, and then conclude with some final remarks for future directions.

5.1 Summary of the dissertation

In Chapter 2, I examined the acquisition of verb argument structure. Linguists agree that there are systematic mappings between the syntax and semantics of a verb (e.g., Gruber, 1965; Jackendoff, 1978; Ladusaw and Dowty, 1988), and it is evident that children know these mapping rules from a young age (e.g., Pinker, 1989; Yuan and Fisher, 2009). This knowledge is unlikely to be entirely universal or innate given the considerable variabilities across languages and idiosyncrasies within (e.g., Fisher et al., 1991; Bowerman and Brown, 2008; Bavin and Stoll, 2013). In this work, I argued these systematic mappings are learnable without assuming any prior associations between syntax and semantics. I presented a computational model that automatically learns productive rules between syntax and semantics based on the TSP: Given the syntactic frames where verbs

appear in an input corpus and the observable semantic features of the co-occurring events, the model recursively examines whether the mapping between a syntactic frame and a semantic feature is bidirectionally productive based on the productivity threshold predicted by the TSP; if so, the model will learn a productive syntax-semantics mapping rule; otherwise, it will move on to the next most frequent forms and features and continue this process. A simulation showed that many of the well documented syntax-semantics mapping rules in English are learnable from modest-sized input data. I also conducted model comparisons to demonstrate that the proposed model performs better than another model that learns verb argument structure from existing literature: a Bayesian model which relies on token frequency (Alishali and Stevenson, 2008) is sensitive to statistical patterns that may not be necessarily consistent with intepretable productive rules, and it makes predictions that are at odds with human behavior.

Chapter 3 looked at the acquisition of recursive structures, with less systematic relation between syntax and semantics. I proposed a new conceptualization of recursion: recursion derives from *structural substitutability*, i.e., all words which can appear in one position in a one-level structure can also appear in the other position. I then proposed that this property of structural substitutability (and thus recursion) can be learned as a productive generalization: children can learn that substitutability holds for a structure if a sufficiently large number of words are attested in both positions in the input. Next, I tested the proposals through corpus analyses of possessive structures and VV constructions in typologically different languages which form natural contrasts in their rules for recursion. I demonstrated that as predicted, the rules for recursive embedding can be established through purely formal analyses: for structures that are freely recursive, there are suffient words attested in both positions so that structural substitutability holds in both directions, so learners do not need to learn any further constraints; for structures that allow recursive embedding in certain circumstances, structural substitutability holds only for one direction, so learners need to learn what words allow substitutability and thus trigger recursion - in this way, the core semantic properties such as alienable vs. inalienable possession for English possessives can be identified subsequently; finally, for structures that in general cannot embed, the number of words attested in both positions is too low to meet the productivity threshold. Taken together, the

results suggest there are reliable distributional cues in one-level input data for the acquisition of recursive structures.

To determine whether such distributional cues are indeed helpful during language acquisition, in Chapter 4, I conducted a series of artificial language learning experiments to examine whether human learners can really acquire recursive structures through distributional learning. In Experiment 1, I exposed adult participants to one-level X_1-ka-X_2 strings in an artificial grammar, with 12 different pseudo-words attested in the X_1 position. In the Productive condition, 10 out of the 12 words were also attested in the X_2 position; in the Unproductive condition, only 6 were. According to the TSP, 10/12 met the productivity threshold but 6/12 did not. At test, it was found that as predicted, participants from the Productive condition were significantly more likely to generalize, i.e., accept both one-level and two-level strings where the post-*ka* position was filled by a word never attested in that position in the exposure phase. Experiment 2 tested the role of structural representation. According to the proposal in Chapter 3, structural substitutability only leads to recursion when the X word is the head. To test whether the head is indeed important for the learning of recursive structures, I presented adult participants with distributional information not just on structural substitutability but also on headedness: In an A_1-B-A_2 grammar, A_1 and A_2 were productively substitutable (12 words were attested in A_1, 9 of them were also attested in A_2), but distributional cues suggested that A_2 was the head in one condition and B was the head in the other condition. Participants were found to integrate the information: They were significantly more willing to allow recursive embedding in the A-head language. Finally, Experiment 3 adapted Experiment 1 for children, showing that 6-to-8-year-old children could also use distributional cues to acquire recursive structures, and that they were more likely to learn categorical generalizations compared to adults. In summary, this chapter indicates that human learners can acquire crucial rules for recursive structures from purely distributional information.

Overall, with the application of the TSP and the integration of different research methods covering a range of linguistic phenomena across languages, this work contributed quantitatively rigorous and psychologically realistic solutions to the well known learnability problem of learning linguistic generalizations, offering new insights into the acquisition mechanism.

5.2 Implications

5.2.1 A mechanistic account for learning generalizations

Research on language acquisition has led to great advances in our understanding of what this process looks like from the outside. We know that although human language is an extremely complex and creative system and children only have access to a finite corpus of linguistic data during language acquisition, all typically developing children acquire their native language with efficiency. We have fine-grained knowledge of their developmental trajectory. However, there are still considerable unanswered questions about the underlying learning mechanism: How do children acquire the linguistic knowledge? The acquisition of syntactic generalizations poses a particularly challenging problem. On one hand, there are regular rules in language that children need to acquire. On the other hand, there are always exceptions to these rules. So how can children strike the balance between the regularity and arbitrariness of language and discover the accurate grammar? It is clear that it cannot be all attributed to universal or innate knowledge. Instead, we need precise, mechanistic accounts for the learning mechanism. It is my hope that this dissertation can make a contribution in this regard.

In this dissertation I applied the TSP, a model that provides a precise threshold for generalization. The moral of such an approach is not limited to the TSP itself, but it allows us to make detailed, testable predictions and advances our understanding of the learning mechanism in a rigorous way. In addition, this work explicitly tested the predictions using different research methods. Through quantitative corpus analyses and computational modeling using input data from different languages, we could evaluate the learning algorithm in real-world learning situations. In artificial language learning experiments, we precisely manipulated the input to test whether human learners' learning outcome would be as predicted to make sure that the model is also psychologically realistic. These methods could complement each other and together provide a more comprehensive picture of language acquisition.

5.2.2 The unit of generalization

One crucial question for the mechanism of learning linguistic generalizations is the unit of generalization: whether learners rely on type frequency or token frequency (e.g., Rumelhart and McClelland, 1986; Baayen, 1989; Bybee, 1995). This dissertation provides novel evidence for the status of type frequency. Of course I do not deny that token frequency plays an important role in many aspects of language acquisition, but I propose that only type frequency matters for the proposal and evaluation of a productive generalization. Across several studies in this dissertation, the proposed account which relies only on type frequency successfully accounts for the learnability of different syntactic generalizations across languages. Moreover, the model comparisons in Chapter 2 explicitly demonstrate that the Bayesian model which relies on token frequency led to learning outcomes that were inconsistent with our linguistic knowledge. Additionally, in Chapter 4, although the experiments did not specifically intend to test the role of type vs. token frequency, it is worth noting that the words in Experiment 1 followed a Zipfian distribution and the word of dominant token frequency was attested in both positions across the two conditions, which means that even in the Unproductive condition, the token frequency of words supporting structural substitutability was high, which was 60 out of 88. Yet the participants did not generalize in the Unproductive condition; nor did our further analyses identify any influence of token frequency on the rating score, e.g., some indirect negative evidence approach would predict that if two words were never attested in X_2 in the input and one of them was attested much more frequently in X_1 position than the other, then when they were used in X_2 position of a test string, the rating score for the string whose X_2 word was of higher token frequency in the input would be lower; however, we did not find this pattern. Therefore, taken together, the findings in this dissertation support a unique role of type frequency in forming generalizations. This is consistent with many observations from previous studies: e.g., rules of higher token frequency but lower type frequency are acquired later (Yang, 2016), and in experiments infants and children only generalize given high type frequency but not token frequency (Schuler et al., 2016; Koulaguina and Shi, 2019).

5.2.3 The status of distributional cues

Taken together, the approach in this dissertation prioritizes formal cues during language acquisition. In Chapter 2, I argued that syntax-semantic mapping rules for verb argument structure do not need to be innate, but are learnable from distributional analyses of the syntactic frames where verbs appear and the easily observable semantic primitives which are known to be salient to young children. Chapter 3 examined the case of recursive structures, where syntax-semantics mappings are less systematic. I showed that the semantic rules can be learned following analyses of the distributional regularities. Finally, Chapter 4 examined learning with no semantic world at all and demonstrated that participants can acquire syntactic generalizations from purely distributional information. I do not intend to deny the role of other cues in language acquisition, but the findings make a strong case for the power of distributional cues. In particular, the findings support a radical separation between form and meaning in language learning and representation. Although semantics has a role to play in the rules for both verb argument structure and recursive structures, I demonstrate that the semantic regularities are learnable based on formal analyses of distributional regularities. It is possible for a structure to be formally productive without meaning, but not the other way around. This connects with Fodor's formulation of "productivity" (for form) and "systematicity" (for meaning), which are related but separable (Fodor and Pylyshyn, 1998).

This idea of prioritizing the form is nothing new in linguistics. The distributional approach to language can be traced back to structuralist linguistics (Harris, 1951), where distributional information is regarded as "the main research of descriptive linguistics" and "the only relation which will be accepted as relevant" (Harris, 1951, p.5). In this tradition, though, distributional analyses are only considered formal methods performed by professional linguists, instead of any psychological mechanisms that children use to acquire language. Distributional methods are also evident in the earliest work of generative grammar. For instance, Chomsky (1975) suggested a distributional approach to linguistic phenomena such as categories and phrase structure, and it was considered as part of human psychology. In this work, I provided a psychological procedure in which children can discover the correct grammar using distributional cues.

5.2.4 An empirically plausible and testable model

In this age of big data and big machines, we are witnessing a surge of interests in computational language models, some of which have exhibited impressive performance and utility. This dissertation, however, highlights the necessity of developing language learning models that make precise cognitive commitments and can be independently verified.

Through the model comparisons in Chapter 2, I revealed problems of the A&S model, which relies on Bayesian inference to learn syntax-semantics mappings in verb argument structure. These problems, however, are not unique to the A&S model, but they pose challenges in general to models which apply machine learning techniques to the representation and learning of language.

The first problem is that human learners may not have the cognitive capacity to perform the algorithms of these models. The A&S model tracks the totality of token frequency information, which human learners do not do according to independent evidence (e.g., Schuler et al., 2016; Koulaguina and Shi, 2019). Other lines of work on word learning have demonstrated that human learners only attend to very local information instead of all co-occurrence information across situations as a Bayesian model (Medina et al., 2011; Trueswell et al., 2013), not to mention the computational intractability of neural network models. Thus, the status of these algorithms as a psychological theory of language learning is at best unclear. A common response to this problem, which can be dated back to Horning (1969) in his pioneering work on the Bayesian framework of grammatical inference, is that these models only concern with the computational level instead of the actual cognitive processes. I agree that the problem of computational complexity is not decisive since we still know little about the precise mechanisms of the mind and the brain. However, first, I believe that as models for the learning of language, they should regard it as their goal to contribute insights into how language is actually learned. Second, even though we abstract away from the mechanism and only focus on the learning outcomes for now, these models can yield incorrect results when we consider the empirical aspects of language and language acquisition.

In our model comparisons, we see that the A&S model may learn strong probabilistic associations even in situations where there is no productive rule. This reflects a core issue with probabilistic language learning models. A design feature of such models is their optimal statistical inference: Sit-

uated in the standard machine learning framework, a model always seeks to find the best grammar for a corpus. However, a fact about natural language is that the best grammar is not necessarily correct: It is possible for human learners to not form any generalization, a phenomenon that the models cannot capture. In the experiments in Chapter 4, although there were still words that supported the generalization in the Unproductive condition, the recursively embedded unattested strings were regarded essentially the same as completely ungrammatical strings. There is also ample evidence in the literature for cases with a lack of generalization in language, such as morphological gaps across languages, i.e., the absence of inflected words for no apparent reason (e.g., Halle, 1973; Baerman et al., 2010), and the fact that children would produce no marker at all in a "wug"-style test when the most frequent marker is not productive in artificial language learning experiments (e.g., Schuler et al., 2016).

Therefore, I argue that the current findings on generalization behavior and the ability of the proposed learning account to explain and predict such behavior present a particular challenge for language learning models from the machine learning framework, which should be subject to examination against the realistic setting of language acquisition. Indeed, computational models can be helpful tools for developing general and mechanical solutions, and it is often necessary to make simplifying assumptions during modelling; however, the central issues of the empirical domain of language and its acquisition should not be compromised. The actual input and output of language acquisition and the psychological procedure that children bring to the task have been recognized as fundamental considerations for the study of language acquisition since the earliest days of modern linguistics (e.g., Chomsky, 1965, 1975). Idealized learning which ignores these core facts can hardly lead to real progress in this field despite its potential utility. Therefore, with the prominence of increasingly powerful machines in recent years, it is still necessary to regard language learning as a psychological theory that is committed to empirically plausible and testable mechanisms.

5.2.5 Children-adults difference in language acquisition

A central question in language acquisition research is why children seem to be so much better at language learning than adults. One possibility is that children and adults have completely different learning algorithms. However, the experiments in this dissertation suggested that this is unlikely the case, since children and adults showed similar learning results in the artificial language learning experiments: They could both use distributional cues to decide whether recursive embedding is possible. One the other hand, though, consistent with previous findings in the literature, children were more likely than adults to form categorical generalizations, which invites further investigations.

In an unreported pilot study for Experiment 3, we had adult participants complete the children-version of the experiment in exactly the same way for our children participants, and we found that their learning outcomes diverged from the predictions: Their rating scores for unattested two-level strings were high in both conditions. While more work is needed to further understand this, one possibility is that there is a task effect. For example, the design might be too simple for adults (e.g., fewer words, slower speech, more children-friendly setup), and therefore they might allocate attention differently or use other cognitive functions available to them to complete the task. It is also possible that given adults' well-established knowledge of their native language and their motivation to do well in an experiment and please the experimenters, adults will be more likely to use their existing linguistic knowledge in the experiment. These influences might exist in some previous studies which observed children-adults difference as well. For instance, in Berko (1958)'s study where participants were asked to produce the past tense form of pseudo-verbs ending with '-ing' such as 'gling', children predominantly produced the regular form 'glinged' while most adults produced an irregular form 'glang', following the irregular pattern in English (e.g., 'sing, sang'; 'ring, rang'). However, in the real-world learning situation, when irregular-like verbs were added to the English lexicon, they were inflected with the regular '-ed', such as 'bing, binged' and 'bling, blinged', indicating that what adults did in the experiment might not reflect their actual language learning algorithm. Overall, together with results from Experiment 1 and 2, this might suggest that similar learning algorithms are available to both adults and children, but whether and how they use it could be influenced by other factors such as the cognitive load and how they understand

the task. If this is the case, future work could tease apart the role of the learning algorithm and other factors by careful manipulation of the task.

Beyond the mechanisms, the current approach predicts that difference in vocabulary size will also make learning linguistic generalizations easier for children. Given the design feature of the TSP, the proportion of regular items needed for a productive generalizations to hold increases dramatically as the number of items to which the generalization may apply (N) increases. As a result, in all the cases studied in this dissertation, it is easier to learn a productive generalization with a smaller vocabulary. This may be a reason why children enjoy an advance in natural language acquisition. However, this cannot be the whole story, since the adults in our experiments also had a small vocabulary yet their behavior still differed from children in subtle ways. Perhaps some combination of all of the accounts above together makes childhood an optimal time for acquiring productive generalizations.

5.3 Future directions

Finally, there are a number of different directions in which to pursue this work in the future. First, it would be desirable to test the predictions of the learning account proposed in this dissertation against a broader range of linguistic phenomena across languages. For instance, it is known that languages have different verb argument structure rules, e.g., some languages allow argument omission, so it would be an important question whether the model proposed in Chapter 2 can capture these language specific patterns. Beyond syntax-semantics mappings, the model can also be extended to other phenomena in verb argument structure, such as syntax-syntax mappings to investigate the acquisition of construction alternations. While Chapter 2 looked at the causative-inchoative alternation, there are many other well-documented alternations, such as the dative alternation (e.g., 'Mary sold a car to Tom.' 'Mary sold Tom a car.'), the unexpressed object alternation (e.g., 'Mike ate the cake.' 'Mike ate.') and the passive alternation (e.g., 'Jane chased Ann.' 'Ann was chased by Jane.'). Next, for recursive structures, as discussed in Chapter 3, there are many cross-linguistic differences regarding the rules for recursive embedding that have not been tested yet against the current proposal. There are also structures where the crucial positions are not in a selectional

relation, which are not a focus of this current work but are worth more investigation. While preliminary data suggested that these recursive structures can also be learned from distributional cues, futures research can study these in greater depth. The examination of a broader range of linguistic phenomena against the proposed accounts does not only contribute to our understanding of the acquisition mechanism but also has theoretical implications: There are extensive debates in theoretical linguistic regarding these generalizations, and learnability data have the potential of informing theoretical discussions in meaningful ways. For example, for verbs that can be used in multiple constructions, one debated question is which variant of the verb should be considered derived from the other (e.g., Levin and Rappaport Hovav, 1995; Davis and Demirdache, 2000; Chierchia, 2004; Reinhart and Siloni, 2004; Ramchand, 2008). Given the current learning account, we would expect to see one-way productivity from the base variant to the derived variant in input data. Some suggested cross-linguistic differences in this phenomena (e.g., Haspelmath, 1993). To test this, we can easily examine whether the direction of productivity differs across languages. If so, then it will suggest that the answer to this debate need not be universal knowledge but is learnable from language specific experience. Another example where learnability can be helpful for linguistic theories is that the arbitrariness of language has been questioned by some authors using data in the domain of verb argument structure. Coppock (2009), for instance, argued that if we consider more detailed semantic and morphophonological constraints such as internal vs. external causation, prosodic weight, and morphological complexity, then whether a verb can participate in the causative alternation and the double object construction will not be arbitrary; instead, productive generalizations will hold within the finer-grained subcategories of words. However, if we can show that these rules are learnable by recursively splitting the vocabulary and searching for productivity, then the rules do not need to be structural constraints but simply consequences of distributional learning.

Second, it will also be important to empirically test some other predictions of the proposed account using experiments: For example, when there is no cross-the-board productive generalization, will learners indeed split the vocabulary to search for local productive generalizations? Another important prediction is that learners can adjust their decision on productive generalization based

on changing numbers of regular and irregular items in the input. To test this prediction, one could conduct an artificial language learning experiment that consists of two phases. In the first phase, the number of words following the generalization exceeds the productivity threshold. In the second phase, participants will learn more irregular words such that at the end of the second phase, the generalization falls out of productivity. The current approach will predict that participants will overgeneralize and then retreat, as often observed in various domains of natural language acquisition.

Future research should also investigate the difference in children's and adults' learning and generalization behavior in greater details. Finally, the proposed approaches are not limited to the topics studied in this dissertation: The problem of learning generalizations is prominent in many different aspects of language. Both the proposal and the research paradigm can be adapted to investigate a range of other acquisition phenomena. In fact, they may not even be limited to language. Yang (2016) suggested that the TSP may be applied to other types of productive generalizations beyond language. The experiments in Chapter 4 suggested this might be the case since the participants' behavior were as predicted even when they did not have access to meaning at all. Therefore, another interesting direction is to ask whether the mechanisms proposed in this dissertation is specific to language or whether it is some domain general mechanism.

Appendix A

SYNTACTIC FRAMES WHERE EACH VERB WAS ATTESTED IN CHAPTER 2

bite: V NP

break: V NP, V

bring: V NP PP, V NP

build: V NP, V NP for NP, V NP NP, V

call: V NP

carry: V NP

catch: V NP

clean: V NP, V

climb: V PP, V

close: V, V NP

come: V PP, V

cry: V

cut: V NP, V, V NP for NP

draw: V, V NP, V NP NP

drink: V NP

drive: V

drop: V, V NP

eat: V NP, V

fall: V, V PP

fly: V

give: V NP to NP, V NP NP, V NP

go: V PP, V

hide: V

hit: V NP, V, V NP with NP

hold: V NP

hug: V

jump: V, V PP

kick: V NP

kiss: V, V NP

move: V NP PP, V NP, V PP, V

open: V NP, V

paint: V NP, V

pick: V NP

play: V with NP, V NP, V

press: V NP, V

pull: V, V NP

push: V NP, V NP PP, V

put: V NP PP, V NP, V PP

reach: V NP, V

read: V NP, V NP to NP, V, V to NP

ride: V

run: V PP

say: V to NP, V CP, V NP

sing: V, V NP, V for NP

sit: V, V PP

stand: V PP

step: V PP, V NP

take: V NP, V NP PP

talk: V to NP, V

tell: V NP, V NP CP, V NP to do, V NP NP

throw: V NP, V NP PP, V, V NP to NP

tickle: V NP

touch: V NP

turn: V NP, V NP PP, V

walk: V, V PP

wash: V NP

wipe: V NP

write: V NP, V, V NP NP

Bibliography

Abbot-Smith, K., Lieven, E., and Tomasello, M. (2004). Training 2;6-year-olds to produce the transitive construction: The role of frequency, semantic similarity and shared syntactic distribution. *Developmental Science*, 7(1):48–55.

Alishali, A. and Stevenson, S. (2008). A computational model of early argument structure acquisition. *Cognitive Science*, 32:789–834.

Ambridge, B. and Lieven, E. V. M. (2011). *Child language acquisition: Contrasting theoretical approaches*. Cambridge University Press.

Ambridge, B., Pine, J. M., Rowland, C. F., Jones, R. L., and Clark, V. (2009). A semantics-based approach to the "no negative evidence" problem. *Cognitive Science*, 33(7):1301–1316.

Ambridge, B., Pine, J. M., Rowland, C. F., and Young, C. R. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgements of argument-structure overgeneralization errors. *Cognition*, 106(1):87–129.

Ammon, M. S. H. (1980). *Development in the linguistic expression of causal relations: Comprehension of features of lexical and periphrastic causatives*. PhD thesis, University of California, Berkeley.

Anderson, S. R. (1969). *West Scandinavian vowel systems and the ordering of phonological rules*. PhD thesis, MIT, Cambridge, MA.

Aronoff, M. (1976). *Word formation in generative grammar*. MIT Press.

Arunachalam, S. (2017). Preschoolers' acquisition of novel verbs in the double object dative. *Cognitive Science*, 41:831–854.

Arunachalam, S., Escovar, E., Hansen, M. A., and Waxman, S. R. (2013a). Out of sight but not out of mind: 21-month-olds use syntactic information to learn verbs even in the absence of a corresponding event. *Language and Cognitive Processes*, 28:417–425.

Arunachalam, S., Leddon, E. M., Song, H., Lee, Y., and Waxman, S. R. (2013b). Doing more with less: Verb learning in korean-acquiring 24-month-olds. *Language Acquisition*, 20:292–304.

Arunachalam, S. and Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. *Cognition*, 114(3):442–446.

Arunachalam, S. and Waxman, S. R. (2011). Grammatical form and semantic context in verb learning. *Language Learning and Development*, 7(3):169–184.

Aslin, R. N. (2017). Statistical learning: A powerful mechanism that operates by mere exposure. *WIREs Cognitive Science*, 8:e1373.

Aslin, R. N., Saffran, J. R., and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, 9(4):321–324.

Aslin, R. N. and Shea, S. L. (1990). Velocity thresholds in human infants: Implications for the perception of motion. *Developmental Psychology*, (4):589–598.

Austin, A. C., Schuler, K. D., Furlong, S., and Newport, E. L. (2022). Learning a language from inconsistent input: Regularization in child and adult learners. *Language Learning and Development*, 18(3):249–277.

Baayen, H. (1989). *A corpus-based approach to morphological productivity: Statistical analysis and psycholinguistic interpretation.* PhD thesis, Vrije Universiteit.

Baayen, R. H. and Renouf, A. (1996). Chronicling the Times: Productive lexical innovations in an English newspaper. *Language*, 72(1):69–96.

Baerman, M., Corbett, G. G., and Brown, D., editors (2010). *Defective paradigms: Missing forms and what they tell us*. Oxford University Press.

Baillargeon, R., Li, J., Gertner, Y., and Wu, D. (2010). How do infants reason about physical events? In Goswami, U., editor, *The Wiley-Blackwell handbook of childhood cognitive development. 2nd ed*. John Wiley & Sons, Inc.

Baker, C. L. (1979). Syntactic theory and the projection problem. *Linguistic Inquiry*, 10(4):533–581.

Bates, E. and MacWhinney, B. (1987). Competition, variation, and language learning. In MacWhinney, B., editor, *Mechanisms of language aquisition*, page 157–193. Lawrence Erlbaum Associates.

Bauer, L. (1978). *The grammar of nominal compounds, with special reference to Danish, English, and French*. Odense University Press.

Bavin, E. L. and Stoll, S., editors (2013). *The acquisition of ergativity*. John Benjamins.

Belth, C., Payne, S., Beser, D., Kodner, J., and Yang, C. (2021). The greedy and recursive search for morphological productivity. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*, pages 2869–2875.

Belvin, R. (1993). The two causative haves are the two possessive haves. *MIT Working Papers in Linguistics*, 20:19–34.

Berko, J. (1958). The child's learning of English morphology. *WORD*, 14(2-3):150–177.

Berman, R. A. (1982). Child language as evidence for grammatical description: Preschoolers' construal of transitivity in the verb system of Hebrew. Unpublished manuscript, Tel Aviv University (cited by Aronoff 1982).

Bertenthal, B. I., Proffitt, D. R., Kramer, S. J., and Spetner, N. B. (1987). Infants' encoding of kinetic displays varying in relative coherence. *Developmental Psychology*, (2):171–178.

Berwick, R. and Chomsky, N. (2017). *Why only us*. MIT Press.

Bisang, W. (2009). Serial verb constructions. *Language and Linguistics Compass*, 3(3):792–814.

Björnsdóttir, S. M. (2021). Productivity and the acquisition of gender. *Journal of Child Language First Review*, 48:1–26.

Bohnemeyer, J. (2008). The pitfalls of getting from here to there: Bootstrapping the syntax and semantics of motion event coding in Yukatek Maya. In Bowerman, M. and Brown, P., editors, *Crosslinguistic perspectives on argument structure: Implications for learnability*, pages 49–68. Lawrence Erlbaum Associates.

Bowerman, M. (1982). Reorganizational process in lexical and syntactic development. In Wanner, E. and Gleitman, L. R., editors, *Language acquisition: The state of the art*, page 319–346. Cambridge University Press.

Bowerman, M. (1987). The "no negative evidence" problem: How do children avoid constructing an overly general grammar? In Hawkins, J. A., editor, *Explaining language universals*. Basil Blackwell.

Bowerman, M. and Brown, P., editors (2008). *Crosslinguistic perspectives on argument structure: Implications for learnability*. Erlbaum.

Bowerman, M. and Croft, W. (2008). The acquisition of the English causative alternation. In Bowerman, M. and Brown, P., editors, *Crosslinguistic perspectives on argument structure: Implications for learnability*, page 279–306. Erlbaum.

Braine, M. D. S. (1963). On learning the grammatical order of words. *Psychological Review*, 70:323–348.

Braine, M. D. S. (1971). On two types of models of the internalization of grammars. In Slobin, D. I., editor, *The ontogenesis of grammar: A theoretical symposium*, page 153–186. Academic Press.

Braine, M. D. S. (1987). What is learned in acquiring word classes – A step toward an acquisition

theory. In MacWhinney, B., editor, *Mechanisms of language acquisition*. Lawrence Erlbaum Associates.

Braine, M. D. S. and Brooks, P. J. (1995). Verb argument structure and the problem of avoiding an overgeneral grammar. In Tomasello, M. and Merriman, W. E., editors, *Beyond names for things: Young children's acquisition of verbs*, page 353–376. Lawrence Erlbaum Associates.

Bresnan, J. (1979). *Theories of complementation in English syntax*. Garland.

Brooks, P. J. and Tomasello, M. (1999). How children constrain their argument structure constructions. *Language*, 75(4):720–738.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Brown, R. and Hanlon, C. (1970). Derivational complexity and the order of acquisition in child speech. In Hayes, J. R., editor, *Cognition and the development of language*, pages 11–53. Wiley.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, 10(5):425–455.

Campbell, A. and Tomasello, M. (2001). The acquisition of English dative constructions. *Applied Psycholinguistics*, 22(2):253–267.

Carlson, M., Sonderegger, M., and Bane, M. (2014). How children explore the phonological network in child-directed speech. *Journal of Memory and Language*, 75:159–180.

Cartmill, E. A., Armstrong, B. F. r., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., and Trueswell, J. C. (2011). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, (28):11278–11283.

Chao, Y. (1980). *A grammar of spoken Chinese*. The Chinese University Press.

Chen, J. (2008). *The acquisition of verb compounding in Mandarin Chinese*. PhD thesis, Vrije Universiteit Amsterdam.

Chierchia, G. (2004). A semantics for unaccusatives and its syntactic consequences. In Alexiadou, A., Anagnostopoulou, E., and Everaert, M., editors, *The unaccusativity puzzle: Explorations of the syntax-lexicon interface, Oxford studies in theoretical linguistics*, page 22–59. Oxford University Press.

Childers, J. B. and Tomasello, M. (2001). The role of pronouns in young children's acquisition of the english transitive construction. *Developmental Psychology*, 37(6):739–748.

Chomsky, N. (1955). The logical structure of linguistic theory. Unpublished manuscript, Harvard University.

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1970). Remarks on nominalization. In *Readings in English transformational grammar*, pages 184–121. Ginn & Co.

Chomsky, N. (1975). *Reflections on Language*. Pantheon Books.

Chomsky, N. (1981). *Lectures in government and binding*. Foris Publications.

Christianson, M. H. and MacDonald, M. C. (2009). A usage-based approach to recursion in sentence processing. *Language Learning*, 59:126–161.

Chung, S. (1998). *The design of agreement: Evidence from Chamorro*. University of Chicago Press.

Conwell, E. and Demuth, K. (2007). Early syntactic productivity: Evidence from dative shift. *Cognition*, 103:163–179.

Coppock, E. (2009). *The logical and empirical foundations of Baker's paradox*. PhD thesis, Stanford University.

Crain, S. (1991). Language acquisition in the absence of experience. *Behavioral and Brain Sciences*, 14(4):597–612.

Cullicover, P. (1988). Autonomy, predication and thematic relations. In Wilkins, W., editor, *Syntax and semantics, Volume 21: Thematic relations*. Academic Press.

Davis, H. and Demirdache, H. (2000). On lexical verb meanings: Evidence from Salish. In Tenny, C. and Pustejovsky, J., editors, *Events as grammatical objects: The converging perspectives of lexical semantics and syntax*, page 97–142. CSLI Publications.

Dawson, C. and Gerken, L. (2009). From domain-generality to domain-sensitivity: 4-month-olds learn an abstract repetition rule in music that 7-month-olds do not. *Cognition*, 111:378–382.

Demuth, K., Culbertson, J., and Alter, J. (2006). Word-minimality, epenthesis, and coda licensing in the acquisition of English. *Language and Speech*, 49:131–174.

den Dikken, M. and Dékány, E. (2018). A restriction on recursion. *Syntax*, 21(1):37–71.

Deng, X. (2010). *The acquisition of resultative verb compound in Mandarin Chinese*. Master thesis, Chinese University of Hong Kong.

Deng, X. (2019). The acquisition of resultative verb compounds in mandarin chinese. *Journal of Chinese Linguistics*, (1):42–81.

Dol, P. (1999). *A grammar of Maybrat: A language of the Bird's Head, Irian Jaya, Indonesia*. PhD thesis, Leiden University, Leiden, Netherlands.

Elsen, H. (2002). The acquisition of German plurals. In *Morphology 2000: Selected Papers from the 9th Morphology Meeting, Vienna, 25-27 February 2000*, page 117. John Benjamins.

Erbaugh, M. (1992). The acquisition of Mandarin. In Slobin, D., editor, *The crosslinguistic study of language acquisition. Volume 3*, pages 373–455. Lawrence Erlbaum.

Esper, E. A. (1925). A technique for the experimental investigation of associative interference in artificial linguistic material. *Language Monographs*, 1:1–47.

Everett, D. L. (2005). Cultural constraints on grammar and cognition in Pirahã: Another look at the design features of human language. *Current Anthropology*, 46(4):621–634.

Feldman, N. (2011). *Interactions between word and speech sound categorization in language acquisition*. PhD thesis, Brown University.

Feldman, N., Griffiths, T., Goldwater, S., and Morgan, J. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, (4):751–778.

Feldman, N., Griffiths, T., and Morgan, J. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of CogSci 2009*.

Fernández-Dobao, A. and Herschensohn, J. (2021). Acquisition of Spanish verbal morphology by child bilinguals: Overregularization by heritage speakers and second language learners. *Bilingualism: Language and Cognition*, 24:56–68.

Ferrigno, S., Cheyette, S. J., Piantadosi, S. T., and Cantlon, J. F. (2020). Recursive sequence generation in monkeys, children, U.S. adults, and native Amazonians. *Science Advances*, 6(26):eaaz1002.

Fetch, A. (2020). *Does learnability predict syntactic universals? An investigation using artificial languages*. PhD thesis, Georgetown University, DC, Georgetown.

Fillmore, C. J. (1968). Lexical entries for verbs. *Foundations of Language*, 4:373–393.

Fisher, C. (1996). Structural limits on verb mapping: The role of analogy in children's interpretation of sentences. *Cognitive Psychology*, 31(1):41–81.

Fisher, C. (2002a). The role of abstract syntactic knowledge in language acquisition: A reply to Tomasello (2000). *Cognition*, 82:259–278.

Fisher, C. (2002b). Structural limits on verb mapping: The role of abstract structure in 2.5-year-olds' interpretations of novel verbs. *Developmental Science*, 5(1):55–64.

Fisher, C., Gleitman, H., and Gleitman, L. R. (1991). On the semantic content of subcategorization frames. *Cognitive Psychology*, 23:331–392.

Fisher, C., Hall, D. G., Rakowitz, S., and Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, 92:333–375.

Fisher, C., Jin, K. S., and Scott, R. M. (2020). The developmental origins of syntactic bootstrapping. *Topics in Cognitive Science*, (1):48–77.

Fodor, A. (1985). Why learn lexical rules? Paper presented at BUCLD-10.

Fodor, J. A. (1970). Three reasons for not deriving "kill" from "cause to die". *Linguistic Inquiry*, 1(4):429–438.

Fodor, J. A. and Pylyshyn, Z. W. (1998). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Frank, M. C., Braginsky, M., Yurovsky, D., and Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press.

Frank, M. C., Goodman, S., and Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, (5):578–585.

Fujimura, C. (2010). Acquisition of recursive possessives in Japanese. Unpublised manuscript, University of Massachusetts.

Gentile, S. (2003). On the acquisition of left-branching recursive possessives. University of Massachusetts honors thesis.

Gentner, D. and Boroditsky, L. (2001). Individuation, relativity and early word learning. In Bowerman, M. and Levinson, S. C., editors, *Language acquisition and conceptual development*, pages 215–256. Cambridge University Press.

Gerken, L. A., Wilson, R., and Lewis, W. (2005). 17-month-olds can use distributional cues to form syntactic categories. *Journal of Child Language*, 32:249–268.

Gervain, J., Macagno, F., Cogoi, S., Pena, M., and Mehler, J. (2008). The neonate brain detects speech structure. *The Proceedings of the National Academy of Sciences*, 105:14222–14227.

Giblin, I., Zhou, P., Bill, C., Shi, J., and Crain, S. (2019). The spontaneous eMERGEnce of recursion in child language. In *Proceedings of the 43rd annual Boston University Conference on Language Development*, pages 270–286.

Gillette, J., Gleitman, H., Gleitman, L. R., and Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, 73:135–176.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, 1(1):3–55.

Gleitman, L. R. and Trueswell, J. C. (2020). Easy words: Reference resolution in a malevolent referent world. *Topics in Cognitive Science*, 12(1):22–47.

Göksun, T., Küntay, A. C., and Naigles, . (2008). Turkish children use morphosyntactic bootstrapping in interpreting verb meaning. *Journal of Child Language*, 29(3):545–566.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, 10:447–474.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goldberg, A. E., Casenhiser, D. M., and Sethuraman, N. (2004). Learning argument structure generalizations. *Cognitive Linguistics*, 15(3):289–316.

Goldwater, S. (2006). *Nonparametric Bayesian models of lexical acquisition*. PhD thesis, Brown University.

Goldwater, S., Griffiths, T., and Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, (1):21–54.

Gomez, R. L. (1997). Transfer and complexity in artificial grammar learning. *Cognitive Psychology*, 33:154–207.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science*, 13:431–436.

Gomez, R. L. and Gerken, L. A. (1999). Artificial grammar learning by one-year-olds leads to specific and abstract knowledge. *Cognition*, 70:109–135.

Green, G. M. (1974). *Semantics and syntactic regularity*. Indiana University Press.

Grimshaw, J. (1990). *Argument structure*. MIT Press.

Grimshaw, J. and Pinker, S. (1989). Positive and negative evidence in language acquisition. *Behavioral and Brain Sciences*, 12:341–342.

Grohe, L., Schulz, P., and Yang, C. (2021). How to learn recursive rules: Productivity of prenominal adjective stacking in English and German. Paper presented at GALANA-9.

Gropen, J., Pinker, S., Hollander, M., and Goldberg, R. (1991a). Affectedness and direct objects: The role of lexical semantics in the acquisition of verb argument structure. *Cognition*, 41(1):153–195.

Gropen, J., Pinker, S., Hollander, M., and Goldberg, R. (1991b). Syntax and semantics in the acquisition of locative verbs. *Journal of Child Language*, 18(1):115–151.

Gropen, J., Pinker, S., Hollander, M., Goldberg, R., and Wilson, R. (1989). The learnability and acquisition of the dative alternation in English. *Language*, 65:203–27.

Gruber, J. S. (1965). *Studies in lexical relations*. PhD thesis, MIT.

Hall, E. and Pérez-Leroux, A. T. (2022). Children's comprehension of NP embedding. *Glossa: A journal of general linguistics*, 7(1):1–41.

Halle, M. (1973). Prolegomena to a theory of word formation. *Linguistic Inquiry*, 4(1):3–16.

Hao, M., Shu, H., Xing, A., and Li, P. (2008). Early vocabulary inventory for Mandarin Chinese. *Behavior Research Methods*, 40:728–733.

Harley, H. (1998). You're having me on: Aspects of have. In Guéron, J. and Zribi-Hertz, A., editors, *La grammaire de la possession*, pages 195–226. Université Paris X - Nanterre.

Harley, H. (2002). Possession and the double object construction. *Linguistic Variation Yearbook*, 2(1):31–70.

Harley, H. and Miyagawa, S. (2016). Ditransitives. In *Oxford Research Encyclopedia of Linguistics*.

Harris, Z. S. (1951). *Methods in structural linguistics*. University of Chicago Press.

Haspelmath, M. (1993). More on the typology of inchoative/causative verb alternations. In Comrie, B. and Polinsky, M., editors, *Causatives and transitivity*, page 87–120. John Benjamins.

Haspelmath, M. (2016). The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics*, 17(3):291–319.

Henke, R. E. (2022). Rules and exceptions: A Tolerance Principle account of the possessive suffix in Northern East Cree. *Journal of Child Language*, 2022 Jun 14:1–36.

Hiraga, H. (2010). Acquisition of recursive verbal compound nouns. Paper presented at the 32nd DGfS.

Hoekstra, T. (1984). *Transitivity: Grammatical relations in Government-Binding theory*. Foris.

Hollebrandse, B., Hobbs, K., de Villiers, J. G., and Roeper, T. (2008). Second order embedding and second order false belief. In *Proceedings of GALA 2007*, pages 270–280.

Horning, J. J. (1969). *A study of grammatical inference*. PhD thesis, Stanford University.

Hudson Kam, C. L. and Newport, E. L. (2005). Regularizing unpredictable variation: The roles of adult and child learners in language formation and change. *Language Learning and Development*, 1(2):151–195.

Hudson Kam, C. L. and Newport, E. L. (2009). Getting it right by getting it wrong: When learners change languages. *Cognitive Psychology*, 59(1):30–66.

Huijbregts, M. (2019). Infinite generation of language unreachable from a stepwise approach. *Frontiers in Psychology*, 10:425.

Hyman, L. M., Alford, D., and Elizabeth, A. (1970). Inalienable possession in Igbo. *Journal of West African Languages*, 8(2):85–101.

Irani, A. (2019). *Learning from positive evidence: The acquisition of verb argument structure*. PhD thesis, University of Pennsylvania, Philadelphia, PA.

Jackendoff, R. (1978). Grammar as evidence for conceptual structure. In Halle, M., Bresnan, J., and Mille, G., editors, *Linguistic theory and psychological reality*. MIT Press.

Jackendoff, R. (1983). *Semantics and cognition*. MIT Press.

Jackendoff, R. (1987). The status of thematic relations in linguistic theory. *Linguistic Inquiry*, 18(3):369–411.

Jackendoff, R. (1990). *Semantic structures*. MIT Press.

Jackendoff, R. (2007). *Language, consciousness, culture: Essays on mental structure*. MIT Press.

Jelinek, E. and Carnie, A. (2003). Argument hierarchies and the mapping principle. In Carnie, A., Harley, H., and Willie, M., editors, *Formal approaches to function in grammar*, page 265–296. John Benjamins.

Jelinek, F. (1998). *Statistical methods for speech recognition.* MIT Press.

Jin, K.-S. and Fisher, C. (2014). Early evidence for syntactic bootstrapping: 15-month-olds use sentence structure in verb learning. In *Proceedings of BUCLD-38*.

Johnson, J. S. and Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, 21(1):60–99.

Johnson, S. P. and Mason, U. (2002). Perception of kinetic illusory contours by two-month-old infants. *Child Development*, (11):22–34.

Jung, Y. J. and Miyagawa, S. (2004). Decomposing ditransitive verbs. In *Proceedings of SICGG*, pages 101–120.

Kauschke, C., Kurth, A., and Domahs, U. (2011). Acquisition of German noun plurals in typically developing children and children with specific language impairment. *Child Development Research*, page 718925.

Kidd, E., Donnelly, S., and Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in cognitive sciences*, 22(2):154–169.

Kline, M. and Demuth, K. (2014). Syntactic generalization with novel intransitive verbs. *Journal of Child Language*, 41:543–574.

Kline, M., Snedeker, J., and Schulz, L. (2017). Linking language and events: Spatiotemporal cues drive children's expectations about the meanings of novel transitive verbs. *Language Learning and Development*, 13(1):1–23.

Kodner, J. (2019). Estimating child linguistic experience from historical corpora. *Glossa*, 4(1):1–14.

Kodner, J. (2022). Language acquisition guiding theory and diachrony: A case study from Latin morphology. *Natural Language and Linguistic Theory*, 41:733–792.

Koulaguina, E. and Shi, R. (2019). Rule generalization from inconsistent input in early infancy. *Language Acquisition*, 26(4):416–435.

Labov, W. (2012). What is to be learned. *Review of Cognitive Linguistics*, 10(2):265–293.

Ladusaw, W. and Dowty, D. (1988). Toward a nongrammatical account of thematic roles. In Wilkins, W., editor, *Syntax and semantics, Volume 21: Thematic relations*. Academic Press.

Landau, B. and Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child*. Harvard University Press.

Lee, J. N. and Naigles, L. R. (2008). Mandarin learners use syntactic bootstrapping in verb acquisition. *Cognition*, 106(2):1028–1037.

Lerdahl, F. and Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Percept*, 1:229–252.

Leslie, A. M. (1982). The perception of causality in infants. *Perception*, (2):173–186.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception*, (3):287–305.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Levin, B. and Rappaport Hovav, M. (1995). *Unaccusativity: At the syntax–lexical semantics interface*. MIT Press.

Li, C. and Thomson, S. (1981). *Mandarin Chinese: A functional reference grammar*. Berkeley University Press.

Li, D., Grohe, L., Schulz, P., and Yang, C. (2021). The distributional learning of recursive structures. In Dionne, D. and Covas, L.-A. V., editors, *Proceedings of BUCLD45*, pages 471–485. Cascadilla Press.

Liang, K., Marsala, D., and Yang, C. (2022). Distributional learning of syntactic categories. In *Proceedings of BUCLD-46*, pages 1442–455.

Liberman, Z., Kinzler, K. D., and Woodward, A. L. (2013). Friends or foes: Infants use shared evaluations to infer others' social relationships. *Journal of Experimental Psychology: General*, 143:966–971.

Lidz, J., Gleitman, H., and Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, 87(3):151–178.

Lieven, E. V. M., Pine, J. M., and Baldwin, G. (1997). Lexically-based learning and early grammatical development. *Journal of Child Language*, 24(1):187–219.

Limbach, M. and Adone, D. (2010). Language acquisition of recursive possessives in English. In *Proceedings of BUCLD*, page 281–290.

MacWhinney, B. (1985). Hungarian language acquisition as an exemplification of a general model of grammatical development. In Slobin, D., editor, *The crosslinguistic study of language acquisition. Volume 2*. Lawrence Erlbaum.

MacWhinney, B. (1987). The competition model. In MacWhinney, B., editor, *Mechanisms of language acquisition*, page 249–308. Lawrence Erlbaum Associates.

MacWhinney, B. (2000). *The CHILDES project*. Earlbaum.

Maling, J. (2002). Verbs with dative objects in Icelandic. *Íslenskt mál og almenn málfræi*, page 31–106.

Maratsos, M., Gudeman, R., Gerard-Ngo, P., and DeHart, G. (1987). A study in novel word learning: The productivity of the causative. In MacWhinney, B., editor, *Mechanisms of language acquisition*, page 89–113. Lawrence Erlbaum Associates, Inc.

Maratsos, M. P. and Chalkley, M. A. (1980). The internal language of children's syntax: The nature and ontogenesis of syntactic categories. In Nelson, K., editor, *Children's language (Vol. 2)*. Gardner Press.

Marcus, G. F. (2001). *The algebraic mind*. MIT Press.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., and Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, (4):1–182.

Marcus, G. F., Vijayan, S., Rao, S. B., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283(5398):77–80.

Marcus, G. M. (1993). Negative evidence in language acquisition. *Cognition*, 46:53–85.

Margetts, A. (2008). Learning verbs without boots and straps? The problem of 'give' in Saliba. In Bowerman, M. and Brown, P., editors, *Crosslinguistic perspectives on argument structure: Implications for learnability*, pages 111–140. Lawrence Erlbaum.

Mascalzoni, E., Regolin, L., Vallortigara, G., and Simion, F. (2013). The cradle of causal reasoning: Newborns' preference for physical causality. *Developmental Science*, (3):327–335.

Matsuo, A., Kita, S., Shinya, Y., Wood, G. C., and Naigles, L. (2012). Japanese two-year-olds use morphosyntax to learn novel verb meanings. *Journal of Child Language*, 39(3):637–663.

Matthei, E. H. (1982). The acquisition of prenominal modifier consequences. *Cognition*, 11:301–332.

Mayberry, R. I. and Kluender, R. (2018). Rethinking the critical period for language: New insights into an old question from American Sign Language. *Bilingualism: Language and Cognition*, 21(5):886–905.

Maye, J. and Gerken, L. (2001). Learning phonemes: How far can the input take us? In *Proceedings of the 25th annual Boston University Conference on Language Development*, pages 480–490. Cascadilla Press.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):101–111.

McClure, K., Pine, J. M., and Lieven, E. V. M. (2006). Investigating the abstractness of children's early knowledge of argument structure. *Journal of Child Language*, 33(4):693–720.

Medina, T. N., Snedeker, J., Trueswell, J. C., and Gleitman, L. R. (2011). How words can and cannot be learned by observation. *PNAS*, 108(22):9014–9019.

Merkuur, A. (2021). *Changes in modern Frisian verbal inflection*. PhD thesis, University of Amsterdam, Amsterdam, NL.

Messenger, K., Yuan, S., and Fisher, C. (2015). Learning verb syntax via listening: New evidence from 22-month-olds. *Language Learning and Development*, 11(4):356–368.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, 90(1):91–117.

Mintz, T. H., Newport, E. L., and Bever, T. G. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26:393–425.

Murrary, W. S. and Forster, K. I. (2004). Serial mechanisms in lexical access: The rank hypothesis. *Psychological Review*, 111(3):721–756.

Naigles, L. R. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, 17(2):357–374.

Naigles, L. R. (1996). The use of multiple frames in verb learning via syntactic bootstrapping. *Cognition*, 58(2):221–251.

Naigles, L. R., Gleitman, L. R., and Gleitman, H. (1992). Children acquire word meaning components from syntactic evidence. In Dromi, E., editor, *Language and Cognition: A developmental perspective*, pages 104–140. Ablex.

Naigles, L. R. and Kako, E. T. (1993). First contact in verb acquisition: Defining a role for syntax. *Child Development*, 64(6):1665–1687.

Naigles, L. R. and Lehrer, N. (2002). Language-general and language-specific influences on children's acquisition of argument structure: A comparison of French and English. *Journal of Child Language*, 29(3):545–566.

Naigles, L. R. and Swensen, L. D. (2007). Syntactic supports for word learning. In Hoff, E. and Shatz, M., editors, *The handbook of language development*, page 212–231. Blackwell.

Nancekivell, S., Friedman, O., and Gelman, S. (2019). Ownership matters: People possess a naïve theory of ownership. *Trends in cognitive sciences*, 23(2):102–113.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science*, 14(1):11–28.

Nichols, J. and Bickel, B. (2013). Possessive classification. In Dryer, M. S. & Haspelmath, M., editor, *The world atlas of language structures.*

Nowenstein, I. (2023). *Building yourself a variable case marking system: The acquisition of Icelandic datives.* PhD thesis, University of Iceland.

O'Connor, C., Maling, J., and Skarabela, B. (2013). Nominal categories and the expression of possession: A cross-linguistic study of probabilistic tendencies and categorical constraints. In *Morphosyntactic categories and the expression of possession*, pages 89–122. Benjamins.

Pearl, L. and Sprouse, J. (2013). Syntactic islands and learning biases: Combining experimental

syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, pages 23–68.

Pearl, L. and Sprouse, J. (2021). The acquisition of linking theories: A Tolerance and Sufficiency principle approach to deriving UTAH and rUTAH. *Language Acquisition*, 28(3):294–325.

Perfors, A., Tenenbaum, J., and Regier, T. (2006). Poverty of the stimulus? A rational approach. In *Proceedings of CogSci 2006*, pages 663–668.

Perfors, A., Tenenbaum, J., and Regier, T. (2011). The learnability of abstract syntactic principles. *Cognition*, (3):306–338.

Pesetsky, D. (1995). *Zero syntax: Experiencer and cascades*. MIT Press.

Pierrehumbert, J. (2003). Probabilistic phonology: Discrimination and robustness. In Hay, J. and Jannedy, S., editors, *Probabilistic linguistics*, page 177–228. MIT Press.

Pinker, S. (1984). *Language learnability and language learning*. Harvard University Press.

Pinker, S. (1987). Resolving a learnability paradox in the acquisition of the verb lexicon. In *Lexicon Project Working Papers*. MIT Center for Cognitive Science.

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. MIT Press.

Pinker, S., Lebeaux, D. S., and Frost, L. A. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, 26(3):195–267.

Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception: Implications for child language acquisition. *Cognition*, 38(1):43–102.

Pérez-Leroux, A., Roberge, Y., Schulz, P., and Lowles, A. (2022). Structural diversity does not affect the acquisition of recursion: The case of possession in German. *Language Acquisition*, 29:54–78.

Pérez-Leroux, A. T., Castilla-Earls, A., Bejar, S., and Massam, D. (2012). Elmo's sister's ball: The problem of acquiring nominal recursion. *Language Acquisition*, 19(4):301–311.

Pérez Leroux, A. T., Castilla-Earls, A., Lara-Dáz, M. F., and Pettibone, E. (2018). Recursion follows productivity, not vice versa: The case of Spanish NP recursion. Paper presented at BUCLD 2018.

Pérez-Leroux, A. T., Peterson, T., Castilla-Earls, A., Béjar, S., Massam, D., and Roberge, Y. (2018). The acquisition of recursive modification in NPs. *Language*, 94(2):332–359.

Pérez-Leroux, A. T. and Roberge, Y. (2018). A way into recursion. In *UMOP 41: Thoughts on mind and grammar (T.O.M. and grammar): A Festschrift in honor of Tom Roeper*. University of Massachusetts.

Quirk, R., Greenbaum, S., Leech, G., and Svartvik, J. (1985). *A comprehensive grammar of the English language*. Longman.

Ramchand, G. (2008). *Verb meaning and the lexicon. A first phase syntax*. Cambridge University Press.

Reber, A. S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, 77:317–327.

Reeder, P. A., Newport, E. L., and Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66(1):30–54.

Reinhart, T. and Siloni, T. (2004). Against an unaccusative analysis of reflexives. In Alexiadou, A., Anagnostopoulou, E., and Everaert, M., editors, *The unaccusativity puzzle: Explorations of the syntax-lexicon interface, Oxford studies in theoretical linguistics*, page 159–180. Oxford University Press.

Richards, N. (2006). A distinctness condition on linearization. Unpublished manuscript, MIT.

Ringe, D. and Yang, C. (2022). The threshold of productivity and the irregularization of verbs in Early Modern English. In Los, B., Cowie, C., Honeybone, P., and Trousdale, G., editors, *English historical linguistics: Change in structure and meaning*, pages 91–111. John Benjamins.

Robinson, M. (2022). *Recursive prepositional phrases in child English*. Master thesis, Macquarie University.

Roeper, T. (2007). *The prism of grammar: How child language illuminates humanism*. MIT Press.

Roeper, T. (2011). The acquisition of recursion: How formalism articulates the child's path. *Biolinguistics*, 5(1-2):57–86.

Roeper, T. and Snyder, W. (2004). Recursion as an analytic device in acquisition. *LOT Occasional Series*, 3:401–408.

Roeper, T. and Snyder, W. (2005). Language learnability and the forms of recursion. In DiScullo, A. M., editor, *UG and external systems: Language, brain and computation*, page 155–169. John Benjamins.

Rosenbach, A. (2014). English genitive variation – the state of the art. *English Language and Linguistics*, 18:215–262.

Rowe, M. L. and Goldin-Meadow, S. (2009). Differences in early gesture explain ses disparities in child vocabulary size at school entry. *Science*, 323(5916):951–953.

Rowland, C. F. and Noble, C. L. (2010). The role of syntactic structure in children's sentence comprehension: Evidence from the dative. *Language Learning and Development*, 7(1):55–75.

Rowland, C. F., Noble, C. L., and Chan, A. (2014). Competition all the way down: How children learn word order cues to sentence meaning. In MacWhinney, B., Malchukov, A., and Moravcsik, E., editors, *Competing motivations in grammar and usage*. Oxford University Press.

Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of English verbs. In Rumelhart, D. E. and McClelland, J. L., editors, *Parallel distributed processing: Explorations in the microstructure of cognition*, page 216–271. MIT Press.

Ruskin, D. (2014). *Cognitive influences on the evolution of new languages*. PhD thesis, University of Rochester, Rochester, NY.

Saffran, J. R. (2001). The use of predictive dependencies in language learning. *Journal of Memory and Language*, 44:493–515.

Saffran, J. R. (2002). Constraints on statistical language learning. *Journal of Memory and Language*, 47:172–196.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Santolin, C. and Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, 22(1):52–63.

Saxe, R. and Carey, S. (2006). The perception of causality in infancy. *Acta Psychologica*, (1-2):144–165.

Schuler, K. (2017). *The acquisition of productive rules in child and adult language learners*. PhD thesis, Georgetown University.

Schuler, K., Yang, C., and Newport, E. (2016). Testing the Tolerance Principle: Children form productive rules when it is more computationally efficient to do so. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*, pages 2321–2326.

Schuler, K. D., Lukens, K., Reeder, P. A., Newport, E. L., and Aslin, R. N. (2023). Children can use distributional cues to acquire grammatical categories. Manuscript under review.

Schuler, K. D., Reeder, P. A., Newport, E. L., and Aslin, R. N. (2017). The effect of Zipfian frequency variations on category formation in adult artificial language learning. *Language Learning and Development*, 13:357–374.

Shi, R. and Emond, E. (2023). The threshold of rule productivity in infants. *Frontiers in Psychology*, 14:1251124.

Shi, R. and Melançon, A. (2010). Syntactic categorization in french-learning infants. *Infancy*, 15(5):517–533.

Shi, R., Werker, J. F., and Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy*, 10(2):187–198.

Singleton, J. L. and Newport, E. L. (2004). When learners surpass their models: The acquisition of American sign language from inconsistent input. *Cognitive Psychology*, 49(4):370–407.

Slobin, D. (1985). Cross-linguistic evidence for the language-making capacity. In Slobin, D., editor, *The crosslinguistic study of language acquisition. Volume 2*. Lawrence Erlbaum.

Sneller, B., Fruehwald, J., and Yang, C. (2019). Using the Tolerance Principle to predict phonological change. *Language Variation and Change*, 31(1):1–20.

Subramanian, D. (2019). *The Tolerance Principle: A closer look at the dative shift*. Master thesis, San Francisco State University.

Swingley, D. (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1536):3617–3632.

Takahashi, E. and Lidz, J. (2008). Beyond statistical learning in syntax. In *Proceedings of GALA 2007*, page 444–454.

Talmy, L. (1975). Semantics and syntax of motion. In Kimball, J., editor, *Syntax and Semantics*. Academic Press.

Teinonen, T., Fellman, V., Naatanen, R., Alku, P., and Huotilainen, M. (2009). Statistical language learning in neonates revealed by event-related brain potentials. *BMC Neuroscience*, 10:21.

Thiessen, E. D. (2011). Domain general constraints on statistical learning. *Child Development*, 82(2):462–470.

Thompson, S. P. and Newport, E. L. (2007). Statistical learning of syntax: The role of transitional probability. *Language Learning and Development*, 3(1):1–42.

Tomasello, M. (1992). *First verbs: A case study of early grammatical development*. Cambridge University Press.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, 74(3):209–253.

Tomasello, M., Akhtar, N., Dodson, K., and Rekau, L. (1997). Differential productivity in young children's use of nouns and verbs. *Journal of Child Language*, 24:373–387.

Trueswell, J. C., Lin, Y., Armstrong, B. F. r., Cartmill, E. A., Goldin-Meadow, S., and Gleitman, L. R. (2016). Perceiving referential intent: Dynamics of reference in natural parent-child interactions. *Cognition*, pages 117–135.

Trueswell, J. C., Medina, T. N., Hafri, A., and Gleitman, L. R. (2013). Propose but verify: fast mapping meets cross-situational word learning. *Cognitive Psychology*, 66(1):126–156.

van Riemsdijk, H. C. (1988). The representation of syntactic categories. In *Proceedings of the Conference on the Basque Language, second Basque World Congress*, pages 104–116.

van Riemsdijk, H. C. (1998). Categorial feature magnetism: The endocentricity and distribution of projections. *Journal of Comparative Germanic Linguistics*, 2:1–48.

van Tuijl, R. and Coopmans, P. (2021). The productivity of Dutch diminutives. *Linguistics in the Netherlands*, 38(1):128–143.

Vendler, Z. (1972). *Res cogitans*. Cornell University Press.

von Hofsten, C., Kochukhova, O., and Rosander, K. (2007). Predictive tracking over occlusions by 4-month-old infants. *Developmental Science*, (5):625–640.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review*, 71(6):473–490.

Weiß, H. (2008). The possessor that appears twice? Variation, structure and function of possessive doubling in German. In *Microvariation in syntactic doubling*, pages 381–401. Emerald.

Williams, E. (1981). Argument structure and morphology. *Linguistic Review*, 1(1):81–114.

Wonnacott, E., Newport, E. L., and Tanenhaus, M. K. (2008). Acquiring and processing verb argument structure: Distributional learning in a miniature language. *Cognitive Psychology*, 56(3):165–209.

Xu, F. and Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, (3):531–556.

Xu, F. and Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, 114(2):245.

Xu, Y. (2006). *The acquisition of resultative verb compounds in Mandarin*. Master thesis, Tsinghua University.

Xun, E., Rao, G., Xiao, X., and Zang, J. (2016). Dashuju beijing xia BCC yuliaoku de yanzhi [The development of the BCC corpus in the context of big data]. *Yuliaoku Yuyanxue [Corpus Linguistics]*, 1(3):93–118.

Yang, C. (2005). On productivity. *Language Variation Yearbook*, 5:265–302.

Yang, C. (2013). Ontogeny and phylogeny of language. *PNAS*, 110:6324–6327.

Yang, C. (2015). Negative knowledge from positive evidence. *Language*, 91(4):938–953.

Yang, C. (2016). *The price of linguistic productivity: How children learn to break rules of language*. MIT Press.

Yang, C. (2018). A formalist perspective on language acquisition (target article with commentary and authors' reply). *Linguistic Approaches to Bilingualism*, 8:665–706.

Yang, X. (2006). Syntactic complexity and productivity: A study of early verbs in L1 acquisition of Mandarin Chinese. In *BUCLD-30 Online Proceedings Supplement*.

Yu, C. and Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420.

Yuan, S. and Fisher, C. (2009). "Really? She blicked the baby?" Two-year-olds learn combinatorial facts about verbs by listening. *Psychological Science*, 20(5):619–626.

Yuan, S., Fisher, C., and Snedeker, J. (2012). Counting the nouns: Simple structural cues to verb meaning. *Child Development*, 83(4):1382–1399.

Zhu, D. (1982). *Yufa jiangyi [Lectures on grammar].* The Commercial Press.

Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology.* Addison-Wesley.

Zwicky, A. (1971). In a manner of speaking. *Linguistic Inquiry*, 11(2):223–233.