6-2020

Doing Away With Defaults: Motivation for a Gradient Parameter Space

Katherine Howitt
The Graduate Center, City University of New York

DOING AWAY WITH DEFAULTS:
MOTIVATION FOR A GRADIENT PARAMETER SPACE

by

KATHERINE HOWITT

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2020

Doing Away With Defaults:
Motivation for a Gradient Parameter Space

by

Katherine Howitt

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	William Sakas Thesis Advisor
Date	Gita Martohardjono Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

Doing Away With Defaults:
Motivation for a Gradient Parameter Space

by

Katherine Howitt

Advisor: William Sakas

In this thesis, I propose a reconceptualization of the traditional syntactic parameter space of the principles and parameters framework (Chomsky, 1981). In lieu of binary parameter settings, parameter values exist on a gradient plane where a learner's knowledge of their language is encoded in their confidence that a particular parametric target value, and thus grammatical construction of an encountered sentence, is likely to be licensed by their target grammar. First, I discuss other learnability models in the classic parameter space which lack either psychological plausibility, theoretical consistency, or some combination of the two. Then, I argue for the Gradient Parameter Space as an alternative to discrete binary parameters. Finally, I present findings from a preliminary implementation of a learner that operates in a gradient space, the Non-Defaults Learner (NDL). The findings suggest the Gradient Parameter Space is a viable alternative to the traditional, discrete binary parameter space, and at least one learner in a gradient space a viable alternative to default learners and classical triggering learners, which makes better use of the linguistic input available to the learner

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance and unending patience of my advisor William Sakas.

To the many people who contributed to the work presented here, especially Paul Feitzinger, Meredith Lancaster, Soumik Dey, and the CoLAG group at Hunter, thank you. Soumik, I owe you a particular note of gratitude for frequently reminding me to "just write". To Janet Fodor, Introduction to Learnability set me down this line of thinking; your guidance and encouragement helped me refine it. Thank you. To the faculty and staff of The Graduate Center Linguistics department and Hunter College Computer Science department, thank you for answering my many questions and offering frequent and genuine encouragement. To Zuzanna Fuchs, thank you for being both a friend and a mentor and for always anticipating which I needed in any given moment. To Callie Robinson, you lived with me through many incarnations of this work and edited much of my code. I am so grateful. To Evan Casper-Futterman, thank you for being the best partner I could ask for. Thank you Catstopher for giving us both comfort and cuddles while we write our long academic works. Dad, Patrick, Francis: I'm done. We can talk about it now. Thank you for the space.

Finally to my mother Karen Kelly and my grandmother Mildred Howitt, I am so grateful for your love and encouragement, especially the tough kind. Thank you for never doubting I'd get here.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Preface

The findings presented in this thesis are preliminary findings that inform work recently submitted (Howitt, Dey, & Sakas, submitted). In this thesis, I explore the feasibility of a non-default learner in the CUNY CoLAG Language Domain, a domain of 3,072 artificial languages generated by 13 syntactic parameters. The purpose of this work is to show that a gradient parameter space for parametric learning could address significant challenges to typical learners in a standard discrete parameter space. The findings of this work suggest that not only is learning feasible in a gradient space, but preferable on many fronts. This thesis explores 10 of the 13 parameters in the CoLAG Domain. Howitt, Dey, & Sakas (*submitted*) expands on these findings both by applying the gradient space to the three additional parameters in CoLAG, modifying some of the e-triggers discussed in this work, and exploring the efficacy of modified learning rates.

1 Introduction

Language acquisition poses a unique modeling problem because the learning mechanism must allow the learner to arrive on a grammar that can parse and generate the set of sentences encountered during learning, but the infinite set of sentences of their language, and the learner must do so in a manner that is relatively quick, efficient, and consistent with observable behavior. Additionally, language learning must in some way depend on the actual language in the learner's environment. Children grow up to speak the language of the adults around them. Children accomplish this with relatively low exposure to surface forms (i.e., sentence level) and no way to see the underlying mechanism that produces these surface forms (cf., Fodor, 1998b).

This lack of exposure, coupled with the potential for infinite grammars in an unconfined grammar space, greatly complicates the problem of learnability.

The principles and parameters (P&P) framework was proposed to simplify the learning process by constraining the possible space of grammars and providing a finite set of innate principles that "sharply restrict the class of attainable grammars and narrowly constrain their form, but with parameters that have to be fixed by experience" (Chomsky 1993, 3). In other words, principles (e.g. languages have subjects) are common to all natural languages, while parameters (e.g. subject-initial or subject-final) would be arrived upon, or set, based on the learner's linguistic environment. Models of acquisition in the P&P framework have required a discrete, generally binary, choice for any parameter in question.

In this thesis, I propose that a parameter is not a discrete binary choice, but rather a point on a gradient scale between these two discrete choices. In other words, a parameter is not "all or nothing". Throughout acquisition, I view the value of the parameter as shifting dynamically on a continuous scale. While the parameter setting itself is a point on the gradient, the *target value* for acquisition would be the discrete choice (a treelet or structure, Fodor, 1998b) on one end of the scale or the other that licenses the sentences in the learner's input. In other words, the parameter setting will change throughout learning asymptotically approaching a target value.

This section will outline a number of concerns relevant to modeling language acquisition. In the following section, I discuss several existing models and how they approach these issues. In section 3, I propose the parametric gradient space as an alternative to modeling language learning and propose a model that can learn in the space. In section 4, I detail an artificial learning domain and summarize the implementation of the learner in that space. In section 6, I

discuss the findings from the implementation. While this thesis provides proof of concept for the gradient parameter space for a monolingual learning environment without noise, I propose that the space itself provides the foundation for exploring a wide range of phenomena including language change, and variation, all of which are briefly discussed in the final section.

1.1 Learnability

Linguistic learnability deals with the logical study of language acquisition and what can, in principle, be learned. The study of learnability is complicated greatly by the challenges outlined in the argument from the poverty of stimulus (Chomsky 1955, 1975). A learner's limited exposure to positive exemplars in the linguistic environment known as *direct positive evidence*, is further complicated by the lack of exposure to what is disallowed in a language, *direct negative evidence* (Brown and Hanlon, 1970, Marcus 1993). The dearth of evidence available to the learner must be accounted for while still *converging* correctly on the target grammar. In other words, the learner must in fact learn the correct grammar.

While learnability encompasses a number of areas, the work in this thesis was in response to the considerations and challenges outlined here.

The Subset Principle (Gold 1977, Berwick 1985, Manzini & Wexler 1987, see extensive discussion in Fodor & Sakas 2005) states that a learner must adopt the subset hypothesis before arriving on the superset hypothesis. Otherwise, a learner would be unable to arrive at a subset hypothesis given only positive exemplars. For example, the interaction of Wh-Movement and Topicalization could create a subset/superset interaction: grammars with obligatory Wh-Movement would generate a subset language of grammars that differ only by obligatory

Wh-movement (the Wh can remain in-situ), but whose Wh elements can move due to either obligatory or optional topicalization. In this case, a learner would have to assume obligatory movement and only arrive on non-obligatory movement given evidence that the Wh is not fronted (i.e., obey the Subset Principle). If the learner were to assume the grammar did not have obligatory movement, it would never be given evidence to the contrary; the hypothesized grammar would over generate sentences.

The P&P framework relies on acceptable triggers (for now I take a trigger to mean a sentence that reveals some fact about the learner's target grammar, see Section 2.1 for expanded discussion), but given no triggers for a subset grammar, models must rely on artificial or psychologically implausible solutions. Some models accept the superset grammar as sufficient for learning (Sakas, Berwick, & Yang 2018), which as stated above, predicts over generation in a hypothesized production system. Other models rely heavily on memory, e.g. The Size Principle in Bayesian learning (see Tenenbaum & Griffiths, 2001; Xu & Tenenbaum, 2007). In these cases, the learner is expected to "remember" past input and hypothesize the subset grammar after some threshold has been met. While these models may arrive on the correct grammar hypothesis, they do so by putting an implausible burden on the learner.

Defaults have been proposed as a less cognitively burdensome solution (Hyams, 1986, Manzini & Wexler, 1987, Fodor & Sakas 2005, Sakas & Fodor 2012). The default hypothesis would be the subset grammar, and the superset grammar hypothesis is only entertained in the presence of direct positive evidence. In other words, default learners adhere to the Subset Principle. However, default models deal with a number of unfavorable consequences including proposing theoretically inconsistent defaults, e.g. movement as a default value (Sakas and Fodor,

2012), or proposing defaults inconsistent with observed psycholinguistic phenomena, e.g. immediate acquisition of obligatory subjects (Hyams & Wexler 1993, Yang 2002). Additionally, default models are unable to account for noisy input.[1] Default models rely on evidence that unambiguously confirms the non-default value. Once a model receives such evidence, it is unable to return to the default value.[2]

Convergence is a learner's arrival at a grammar hypothesis. A model must have a way to measure whether a language has been learned, and, if so, at what point during the course of learning that final grammar was hypothesized. For linguistic learnability, convergence is standardly defined as arriving at a static grammar within a finite amount of time after entertaining a series of grammar hypotheses (e.g., Gold 1967).

Imposing a criterion of finiteness is motivated on both psychological and mathematical grounds. Research in natural language acquisition has established a *critical period* [3] in which language learning occurs rapidly and effortlessly. At some point, the adult grammar is thought to be relatively static. From a mathematical perspective, finiteness is preferable in terms of learnability (Gold 1967). In other words, a model should have a metric by which it can be measured if and when something has been learned.

[1] Noise is an inevitable aspect of natural language learning and while children are inconsistently given direct negative evidence, they are often given unreliable positive evidence through misspeaks, stutters, and other variations in language use. Models that are unable to withstand noisy input are less desirable than those that can (Pinker 1979).
[2] C.f. Sakas and Fodor 2012, see discussion on the interaction between Optional Topic and Null Topic. While their proposal does allow for "toggling" between the default and non-default values, it was added as a special case. Howitt, Dey, & Sakas (submitted) offer a response to this specific case that makes use of the gradient parameter space proposed in this thesis..
[3] In language acquisition literature the critical period specifies that a language is learned without issue by humans within what is considered a typical learning environment. The model proposed in Section 3.2 assumes learning proceeds in a typical environment. I.e., a two year old exposed to one language over the course of the critical period, measured in this work by 500,000 sentences.

Convergence on a *weakly equivalent* grammar is typically considered sufficient for successful learning (Yang 2002, Sakas & Fodor 2012, Sakas, Berwick, & Yang 2018, and others). A grammar hypothesis $G_{HYPOTHESIS}$ is weakly equivalent to the target grammar G_{TARGET} if $G_{HYPOTHESIS} \neq G_{TARGET}$, but both grammars produce the same set of sentences. In other words, the language generated by either $G_{HYPOTHESIS}$ or G_{TARGET} is the same. In the P&P framework, weak equivalency is created when one or more parameters is irrelevant to the set of sentences generated.

A **Learning Path** is related to, but differs from, convergence. The learning path is thought to describe the sequence of grammar hypotheses (or parameter setting) as the learner acquires more information (Dresher 1999). A model of learnability that is psychologically plausible should be able to mirror intermediate stages of development (Pinker 1979). Learning paths are particularly troublesome for statistical models (e.g. Yang 2002, Gould 2015 and Bayesian model) that rely on occasional unlikely guesses. In these cases, a learner would randomly hypothesize grammars that are not predicted by current or past input. Child acquisition data suggests a fairly predictable learning path (e.g., Yang 2002, Sugisaki & Snyder 2003, 2006, Dresher 1999), which, in many cases, makes random grammar hypothesizing a less desirable model. Similarly default models have difficulty with predicting a learning path appropriately, most obviously when the target grammar is the same as the all default grammar (no learning path) or the target grammar lacks any defaults.[4]

[4] There has been some research on the psycholinguistic viability of defaults (Sugisaki & Snyder 2003, 2006, Sugisaki 2007), but from a computational perspective ease and time course of learning would vary significantly, something not observed in natural language acquisition. Additionally, Gould (2017), offers some evidence that defaults could not account for observed behavior in Korean.

Negative Evidence is evidence about what is disallowed in the learner's language. Learners pay little or no attention to direct negative evidence about what is ungrammatical in their language (Brown and Hanlon, 1970, Marcus, 1993 and references there) and thus must construct a grammar hypothesis given only positive exemplars (i.e., *direct positive evidence*). There is a body of research that suggests that learners can use *indirect negative evidence*, i.e., the lack of a linguistic pattern in the child's linguistic environment from which the child can induce what is disallowed by the grammar being acquired (see Regier & Ghal 2004, Foraker et al. 2009). However a number of concerns have been raised about the ability of a child to identify and utilize indirect negative evidence during language acquisition (e.g., Pinker 1989, Hornstein 2016, Yang 2017).

Ambiguity greatly complicates acquisition. Ambiguous sentences are surface forms that can be generated by different underlying structural properties. Thus, the target grammar cannot be identified using a single ambiguous utterance. Human language contains many utterances that can be either *ambiguous* with respect to a particular grammatical construction, or *irrelevant* to a construction. For example, a sentence containing a preposition adjacent to its object does not unambiguously rule out a grammar that optionally allows preposition stranding. However, a preposition separated from its object *is* unambiguous evidence for a grammar that allows preposition stranding. Note that in this work I am considering cross-linguistic ambiguity, i.e., when different parameter values license the same sentence pattern. This ambiguity can be contrasted with structural ambiguity within a single language, i.e., when a sentence pattern might have multiple structural descriptions given fixed parameter values for a particular language.

Ambiguity stands in contrast to **irrelevance**. In an utterance that does not contain a prepositional phrase, whether or not a grammar allows preposition stranding is irrelevant to that utterance.

Ambiguity and irrelevance are often conflated. In the case of preposition stranding disentangling them is relatively straightforward. However, when multiple constructions interact, for example null topic and null subject constructions, it is difficult to differentiate ambiguity from irrelevance. Most learning models do not make this differentiation.

The gradient learning model discussed in the following sections has the capability to learn in a domain containing subset/superset languages without access to negative evidence. This thesis examines whether simulated learners in a gradient parametric space arrive at the correct, or more accurately, approximately correct, grammar hypothesis after a finite number of inputs. Finally, the model has the ability to distinguish ambiguity from local irrelevance and makes use of both unambiguous and ambiguous utterances.

2 Learning in Discrete Spaces

The classic conception of the principles and parameters framework relies on a discrete parameter space in which there are a finite number of grammars generated by some combination of parameter values. If the parameter space is composed of binary parameters, then the number of possible grammars is at most 2^n where n is the number of parameters in the space. Learners in these spaces hypothesize either individual parameter values (Fodor 1998a,b; Sakas & Fodor 2012, and others) or complete grammar (Yang 2002 and elsewhere, Gould 2017).

The task of a learner is to discard hypotheses that are inconsistent with the language environment the learner is experiencing and to successfully converge on the target grammar.

These learners effectively perform a search of the parameter space where after each utterance, the current hypothesis or hypotheses are tested and either discarded or retained based on one or more search heuristics (Clark 1989, 1992, Yang 2002 and elsewhere, Gibson & Wexler 1994, Fodor & Sakas 2004, Gould 2017 among others).

Notably, each of these learners employ a search of a discrete parameter space where, at any point in the learning process, complete grammars are hypothesized. In contrast, Fodor (1998b) sought to isolate specific parameters, as opposed to *complete grammars* during the course of learning by *decoding* a single input sentence in order to determine some unambiguous structural fact about the target language (i.e., a parameter value or in Fodor's paradigm what she calls a *parametric treelet*).

It is of note that all the learners choose from a discrete set of parameter values (i.e., they operate in a discrete, as opposed to continuous, grammar hypothesis space). What follows is not an exhaustive outline of computational models of learning in a discrete parameter space, but a sample of works that inspired this one.

2.1 Triggering and Learning With Defaults

The earliest conception of learning in P&P involves *triggering*, in which the learner encounters some syntactic fact in their language through an input or utterance. Once, the learner hears this utterance, the relevant parameter value for that learner's linguistic environment is set (see Chomsky 1981b, 1986).

However, the specifics of what triggers a parameter value differ from model to model. The work published in Sakas & Fodor (2012) provides some basis for the work presented here.

Sakas & Fodor define an *e-trigger schema* as an observable pattern in the surface form of a sentence that sheds light on some structural fact about the grammar that generates that sentence. For example, a preposition separated from its object is an e-trigger schema for preposition stranding. Note that Fodor's (1998b) parametric treelets (or *i-triggers* in Sakas & Fodor) provide the structural basis for the surface form e-trigger schemas. This work makes considerable use of Sakas & Fodor's notion of e-triggers, however it should be emphasized that i-triggers are not directly used.

Four things to note about e-triggers:

(i) they are schemas that operate over multiple sentences,

(ii) they do not necessarily depend on an entire sentence,

(iii) they are observable to the child learner in a sentence's surface form, and finally,

(iv) they can serve as the foundation for a classic triggering model.[5]

Sakas & Fodor executed an extensive search for *unambiguous* e-triggers in a complex artificial language domain (described below in Section 4). An unambiguous e-trigger is definitive evidence for a parameter value. If encountered, an unambiguous trigger can be used to permanently adopt that parameter value[6] i.e., the parameter will never be reset by the learner (the *parametric principle* in Sakas and Fodor, 2001). Unambiguous e-triggers existed only in languages with a given parameter value.

[5] Only unambiguous e-triggers can be used to implement a classic triggering model. One addition this work makes to the e-triggers presented in Sakas & Fodor is the development and use of ambiguous e-triggers.

[6] From this point forward *trigger*, *e-trigger* and *e-trigger schema* are used interchangeably. A trigger refers to an observable pattern in the surface-form. The term i- trigger will be used consistently in the sense outlined above.

Sakas & Fodor found that unambiguous triggers existed for both parameter values for 5 of the 13 parameters in the CUNY-CoLAG domain. E-triggers for just one parameter value of an additional 5 parameters. The final three parameters in the domain lacked unambiguous triggers for either value. To learn in this domain, Sakas & Fodor proposed default parameter settings and a small 'toolkit' of disambiguation strategies which would allow a learner to successfully learn every language in the domain. The defaults, disambiguation strategies, and e-triggers were assumed to be provided by UG.

Classic triggering is desirable because it attends to the specific linguistic evidence presented to the learner. Thus, learners in this paradigm only address parameters relevant to a given trigger. Furthermore, defaults can implement the Subset Principle assuming UG posits the subset parameter value as the default (op cit). There is some evidence for the feasibility of defaults based on consistencies in child language acquisition (e.g., Sugisaki & Snyder 2003, 2006, Sugisaki 2007). However, in Sakas & Fodor's work, proposed defaults could contradict canonical syntactic proposals, e.g., movement as the marked value vs. movement as default (Sakas and Fodor, 2012; Sections 2.3.2 and 2.3.3), or contradict known psycholinguistic phenomena, e.g., immediate acquisition of obligatory subjects (see Hyams & Wexler 1993, Yang 2002). Additionally, the burden of language learning varies between the extremes: a learner acquiring a grammar with all non-default values and a learner whose target grammar is entirely made up of the prescribed defaults. No learning would be needed in the latter case; complete grammatical competence would be in place at birth.

2.2 Learning Without Triggers

Other computational P&P models have moved away from triggers and individual parameters, and instead search a grammar space for a complete grammar hypothesis. Such a learner is presented with an input sentence and selects a grammar from a finite grammar space to parse that input sentence. If the parse succeeds the learner proceeds with that grammar, and if the parse fails, the learner discards that hypothesis. The mechanism needed to parse the input sentence need not utilize specific linguistic evidence in the input sentence, nor even have any knowledge of the parameter settings necessary to parse the input.

Yang (2002)'s Variational Learner is such a learner. At any point in time, the Variational Learner has a single grammar hypothesis composed of discrete parameter values: one for each parameter. The learner uses real-valued weights to conduct a non-deterministic search of the discrete parameter space in order to choose the next grammar hypothesis. Given an input sentence, this learner adjusts weights uniformly across all parameters and uses the weights to hypothesize a complete grammar after every utterance. In other words, one value of every parameter in the domain is chosen, which then becomes the VL's current grammar hypothesis. The VL must set values for all parameters on each sentence because it doesn't employ e-triggers for specific parameter values. Additionally, because the learner's parameter values are chosen probabilistically based on the weights, the learner can make a dramatic shift in its grammar hypothesis based on a single input. The model has no way to privilege unambiguous input.

Gould (2017) presents a Bayesian inspired learner that maintains weights similar to Yang's learner. Unlike Yang's learner, Gould's learner is able to identify and learn from ambiguous input. However, he forgoes the use of triggers and in their stead posits that the learner

performs multiple computations over an input sentence. Given a single input, his learner repeatedly samples a discrete space of grammars, either searching for compatible grammars ("simplified version") or for grammatical choices that generate an 'output' that matches the current input ("full version"). The repetition is crucial to the models' success—unambiguous input will cause the learner to positively reinforce the same grammar (or grammatical choices) multiple times given a specific input, whereas ambiguous evidence will distribute the reinforcement over multiple grammars. Gould shows that his model successfully learns from ambiguous evidence, even in the extreme case where the target language is a subset of other languages in the domain. However, it's unclear whether the learning model is computationally tractable given a large and complex domain such as the one described below. While reported successes of the model may be replicable in domains more complex than the restricted domains Gould utilizes, it's unclear if the computational load (and time) required to achieve a result is feasible. In addition, this sort of parallel or multiple parsing/generating is psychologically implausible and theoretically unattractive. See criticism of related models in Sakas & Fodor (2007) and Yang (2017).

3 Triggering and Learning without Defaults

3.1 The Gradient Parameter Hypothesis

I propose a reconceptualization of the parameter space in which the value of a parameter is not a binary choice, but rather exists on a one dimensional gradient between two possible structures of a grammar (or parameter setting). The parameter value is a measure of the learner's confidence in one or the other structure as the correct licenser of the learner's linguistic

environment. In this conception, all learners begin neutral with respect to any given parameter setting. There are no defaults. As mentioned above, the endpoints of the gradient plane can be thought of as structural treelets (Fodor 1998b) that are provided by Universal Grammar. I maintain the standard assumptions that UG provides a finite number of parameters, and cross-linguistic principles. Thus, the grammar space is still greatly constrained.

From a formal learnability perspective, the fact that parameter values are continuous rather than discrete introduces learnability problems (due to infinite variation of target hypotheses) that don't exist in a finite space (Bertollo 2001, Jain et al., 1999). However, from an empirical perspective, the results presented in the simulation below suggest that such problems may be functionally insignificant. While the space of possible grammars is continuous and yields infinite variations, the variation within the space is still constrained by the licensing UG principles. This conception allows a more robust explanation of variational phenomena.

The empirical study presented here demonstrates that learning in this space is not only possible, but preferable on many fronts, including laying the groundwork for modeling language variation (e.g. idiolects, dialects, language change, etc).

The gradient parameter space also marks a significant change in the way a learner can process ambiguous evidence for parameter values. A learner in this space can incrementally encode indirect negative evidence. For example, the Wh-Movement parameter is a subset-superset parameter in the CUNY CoLAG domain. A non-fronted Wh-phrase is an unambiguous e-trigger for Wh-in-situ, and when encountered, the confidence value can be adjusted aggressively. However, unambiguous e-triggers for Obligatory-Wh-Movement (ObWhM) do not exist. When a fronted Wh-phrase is encountered, confidence is conservatively

adjusted toward ObWhM, although the movement could stem from optional topicalization (unrelated to the Wh-Movement parameter). For a language that doesn't contain Wh-in-situ, this conservatism enables the NDL to unfailingly converge (unlike other learners, e.g., Yang's variational learner) on the subset language.

Fodor (1998a) introduced the notion of activation levels in treelets which a learner would maintain during the course of acquisition. Yang (2002) introduced a learner that maintains weights of parameter values which the learner accesses throughout the learning process. Both these models function in a traditional discrete parameter space and were motivation for the work presented here. The key difference is that the activations and weights are not part of the learners' grammars but rather serve as a tool for selecting one parameter target value (in the case of Fodor) or one grammar hypothesis (in the case of Yang) over others. In this conception, the confidence value ***is*** the parameter value and thus a direct outcome of the learner's grammatical knowledge at any point in the learning process. A discussion of the advantages of a gradient parameter space follows the discussion of a specific implementation of a learner in the space.

3.2 The No-Defaults Learner

A No-Default Learner (*NDL*) is a model of parameter setting that makes use of ambiguity in a domain to learn without defaults. The gradient parameter hypothesis removes the assumption that parameter setting is immediate or that there is a finite number of possible grammars. The NDL constructs a grammar hypothesis by adjusting its position on the gradient for each parameter when presented with relevant e-triggers. Confidence changes over time, thus, similar to both Fodor and Yang (op cit.) the point on the gradient encodes past experience. However,

unlike previous parameter setting models, the NDL can both distinguish between and make use of ambiguous and unambiguous e-triggers, and can do so on the basis of single inputs. As a result, the NDL can retreat from non-target superset hypotheses. The NDL proceeds by searching a sentence for an e-trigger for each parameter. When it encounters one, it adjusts its confidence for that parameter accordingly. Each parameter value lies on a real-valued gradient from 0 to 1, where 0 and 1 represent two mutually exclusive parameter values. Following Yang (2002), the NDL uses a modification of Bush & Mosteller(1958)'s L_{R-P} scheme for adjusting the value, where adjusting toward 0 is calculated using Equation (1)

$$P^{v}_{i+1} = P^{v}_{i} - \widehat{R}(P^{v}_{i}) \quad (1)$$

and adjusting toward 1 is calculated using Equation (2)

$$P^{v}_{i+1} = P^{v}_{i} + \widehat{R}(1 - P^{v}_{i}) \quad (2)$$

Where P^{v}_{i} represents the confidence value of parameter P_i, and $\widehat{R}$ can be either R (aggressive rate) or r (conservative rate). The NDL is statistical, but at the same time deterministic: there is no randomness in NDL learning. Given a fixed input corpus an NDL learner will always converge to exactly the same place on the gradient for all parameters. This notion of deterministic is different from the notion of deterministic in parsing and learning where once a deterministic learner or parser makes a choice, the choice cannot be revised. Because the NDL learns on a gradient, it can retreat from one or another endpoint at any point in learning, given evidence against it.

The NDL makes use of two types of triggers: (i) unambiguous triggers and (ii) ambiguous triggers. When presented with an unambiguous trigger, the NDL adjusts its confidence aggressively. When presented with an ambiguous trigger, the NDL conservatively adjusts its confidence. Ambiguous triggers are positive surface-level evidence of a syntactic form in the learner's language. For example, the NDL uses the existence of a subject in a declarative sentence as ambiguous evidence that declarative sentences require subjects; it is ambiguous because null subject languages also allow overt subjects. In the absence of contradictory evidence (e.g., a declarative sentence without a subject), the learner will gradually increase its confidence that its language requires overt subjects. It is notable that the NDL only makes use of relevant e-triggers (ambiguous or unambiguous), and thus avoids adjusting the value of a parameter in the absence of evidence for it (e.g., changing the confidence value for the null subject parameter in an imperative sentence). Schematic pseudocode for the NDL upon encountering an utterance (sentence) is outlined in Algorithm 1.

$n \leftarrow$ the number of parameters
NDL encounters a sentence, *s*.
for *Parameter,* P_i, $1 \leq i \leq n$ do
    if *in s, an e-trigger schema,* e_i*, is detected for* P_i **then**
        if e_i *is unambiguous* **then**
            adjust confidence gradient for P_i aggressively.
        else
            e_i is *ambiguous*, adjust confidence gradient for P_i conservatively.
        end
    else
        Nothing in *s* is relevant for learning P_i, do nothing.
    end
end

Algorithm 1: No Defaults Learner (NDL) pseudocode. Note that adjustment to the confidence gradients for each parameter, if any, follows either equation (1) or equation (2) depending on the target value (endpoint, i.e., 0 or 1) that the detected e-trigger is triggering.

The NDL also modifies the notion of conditioning from the disambiguation 'toolkit' proposed by Sakas & Fodor. Conditioning occurs when knowledge of one or more parameters disambiguates ambiguous evidence for another parameter. For example, if a learner has confidence that the target language has obligatory topic marking, and disallows null topics, then lack of a word overtly marked as a topic (in a declarative sentence) is unambiguous evidence for a target grammar that does not require topicalization. The same sentence is ambiguous evidence in languages that don't have topic marking and/or languages that allow null topics.

Sakas & Fodor note that conditioning is dangerous for deterministic learners unless the conditioning parameters' values (e.g, existence of topic marking and disallowed null topics) are set with unambiguous triggers. That is, conditioning values must be correctly set before a conditioned parameter value can be set. However, because the gradient parameter space embodies the learner's confidence in a particular target value at a given point during learning, conditioning for the NDL only requires that the learner's confidence in one value is greater than its confidence for the other value. In other words, the learner need only have been minimally exposed to one e-trigger (ambiguous or unambiguous) in order to make use of conditioning. While it is possible that early in learning the learner might incorrectly condition a given parameter value based on misplaced confidence in a conditioning parameter, the gradient space is such that the NDL can recover not only the conditioning parameter but also the conditioned one. The ability to recover from misplaced hypotheses is dependent on the balance between unambiguous and ambiguous triggers; there must exist enough unambiguous triggers to overcome initial misplaced confidence.

The balance between learning aggressively from unambiguous e-triggers and learning from ambiguity gives the NDL the power to learn in domains with subset/superset relations without resorting to indirect negative evidence. Success of the algorithm is based on the assumption that unambiguous e-triggers exist, even if rare, for superset languages. The NDL proceeds by conservatively gathering evidence for the subset language, which is by definition ambiguous between the subset and superset languages. If the target language is a superset, the NDL will eventually encounter unambiguous triggers. Given unambiguous evidence, the NDL will aggressively adjust its confidence toward the superset language. If successful, this unambiguous evidence will prevail over the ambiguous evidence. Crucially, it is not the absence of an e-trigger for the superset language that allows the NDL to converge on the subset. Rather, it is the existence of positive, albeit ambiguous evidence, for the subset language that the NDL makes use of.

4 The CoLAG Domain

The NDL simulations presented in this thesis make use of the CUNY-CoLAG language domain. The Computational Language Acquisition Group (CoLAG) at the City University of New York (CUNY) created an artificial domain of 3,072 languages for the purpose of exploring computational models of language acquisition within a linguistically rich set of phenomena that have been the focus of language acquisition research more generally.

The language domain contains grammars generated by 13 binary syntactic parameters which have been the focus of language acquisition research in the Government and Binding framework (Chomsky 1981). Exhaustive details of the domain and how it was generated are

available in Sakas (2003) and Sakas & Fodor (2011, 2012). The thirteen parameters express phenomena typical of child speech including head position, verb movement, null subject, and topicalization. There are constraints on certain parameters within the CoLAG UG (such as the incompatibility of Affix Hopping and V-to-I Movement) which decreases the number of possible languages from 2^{13} to 3,072.

The domain contains tokens which encode grammatical roles bound to syntactic categories. The universal CoLAG lexicon consists of the following tokens:

- S [ubject],
- O1 [direct object],
- O2 [indirect object],
- O3 [prepositional object],
- P [reposition],
- Adv [erb],
- Aux [illary Verb],
- Verb,
- Not [Negation, head of NegP],
- Never [Adverb dominated by spec NegP],
- ka [interrogative marker in C],
- -wa [Topic marker].

Each language in the domain contains between 288 and 2,148 sentence patterns containing these tokens, along with fully specified structural tree(s) for each pattern. Examples

(1) and (2) show CoLAG sentence forms found in the CoLAG language 611 (a language in which the parameter values are most closely aligned to those expected of English).

(1)	S	Aux[+FIN]	Verb	Adv	[ILLOC DEC]
	Subject	*Tensed Auxiliary*	*Verb*	*Adverb*	*[Declarative Sentence]*

(2)	Verb[+FIN]	S	P	O3	[ILLOC Q]
	Tensed Verb	*Subject*	*Preposition*	*Prepositional Object*	*[Question]*

In addition to the illocutionary force feature ILLOC, the domain contains other features, including WH, FIN, SLASH, NULL, etc. Relevant to this work is the overt WH feature which marks a Wh word in a question [ILLOC = Q] and FIN which marks a verb or auxiliary with tense. Because the CoLAG-NDL only learns from surface forms that contain e-triggers, non-surface features and full structural trees are not available to the NDL at any time during acquisition and are not discussed further.

A number of motivations exist for using CoLAG in acquisition research instead of natural language data such as that found in CHILDES (MacWhinney, 2000). CoLAG provides the unique opportunity to explore hypotheses related to language variation that are difficult to isolate in natural languages. In addition, while there is much language acquisition data for more studied languages like English, Spanish, German, & Korean, etc. less-studied languages are not represented. Thus unique linguistic phenomena might be ignored. A domain like CoLAG rescues a UG account of language from becoming too English-focused. Additionally, the domain provides the opportunity to compare many different models of parameter setting. The results of these studies can then be compared to child-based acquisition data to reinforce their validity.

CoLAG is not intended to be an exhaustive representation of all natural language phenomena or all possible theories of language, but rather a space to test learning theories on a complex range of specific linguistic phenomena.

5 The CoLAG-NDL E-Triggers

The thirteen parameters that generate the CoLAG domain are listed in Table 1.

Parameter List			
Parameter Name	Abbreviation	Target Value = 0.0	Target Value = 1.0
Subject Position	(SP)	Initial	Final
Headedness in IP	(HIP)	Initial	Final
Headedness in CP	(HCP)	Initial	Final
Optional Topic	*(OpT)*	*Obligatory Topic*	*Optional Topic*
Null Subject	(NS)	No Null Subject	Optional Null Subject
Null Topic	(NT)	No Null Topic	Optional Null Topic
Wh-Movement	(WhM)	Wh-Insitu	Obligatory Wh Movement
Preposition Stranding	(PI)	Obligatory Pied Piping	Prepositional Stranding
Topic Marking	(TM)	No Topic Marking	Obligatory Topic Marking
V to I Movement	(VtoI)	No VtoI Movement	Obligatory VtoI Movement
I to C Movement	*(ItoC)*	*No ItoC Movement*	*Obligatory ItoC Movement*
Affix Hopping	(AH)	No Affix Hopping	Affix Hopping
Question Inversion	*(QInv)*	*No QInversion*	*Obligatory QInversion*

Table 1: CoLAG parameters and corresponding NDL target values. Parameters in italics are not included in the simulations presented in this thesis.

Sakas & Fodor (2012) launched an extensive search of the CoLAG domain. The goal of this search was to identify unambiguous triggers found at the sentence level that would trigger every parameter value for all the grammars in the domain. Additionally, They hoped to find e-triggers that were available to all languages of the domain.

They found that of the 13 parameters in the domain, five had unambiguous e-triggers for both parameter values: Subject Position, Headedness in IP, Headedness in CP, Topic Marking, and Preposition Stranding. These five parameters are unproblematic as they can be decisively set upon encountering such an unambiguous trigger. Five parameters had unambiguous e-triggers for only one value: Null Subject, Null Topic, Wh Movement, VtoI Movement, and Affix Hopping. The remaining three parameters lacked unambiguous triggers for either value: Optional (as opposed to obligatory) Topicalization, ItoC Movement, and Question Inversion (mandatory ItoC in questions). To address the lack of unambiguous triggers, Sakas & Fodor (2012) proposed default parameter settings and a small 'toolkit' of disambiguation strategies which would allow a learner to successfully learn every language in the domain (p.95, p. 113).

Sakas & Fodor identified unambiguous triggers for at least one value of the parameter for 10 out of the 13 parameters in the domain. In the cases where no ambiguous trigger existed for one value, Sakas & Fodor proposed that that value be the default parameter setting. In cases where both values of the parameter had unambiguous triggers, a default value was randomly assigned. For the non-default values, Sakas & Fodor proposed at least one e-trigger.

The NDL employs all of the e-triggers presented by Sakas & Fodor. By design, the NDL does not make use of defaults, so additional e-triggers were developed for their default values. For the first five unambiguous parameters, these e-triggers were alluded to but not explicit in

Sakas & Fodor. For the following five parameters, ambiguous triggers, which rely on a gradient parameter space, were developed.

What follows is an outline of the e-triggers the NDL employs for the ten parameters for which unambiguous e-triggers existed for at least one parameter value. The three remaining parameters (ItoC Movement, Question Inversion (ItoC Movement in Questions only), and Optional/Obligatory Topic) are not included in the simulations presented in this thesis. Question Inversion is discussed briefly in Section 7 and the others are mentioned in Section 8 when their interactions affect the observed outcomes.[7]

5.1 Parameters with Unambiguous Triggers for Both Values

Parameters with unambiguous triggers for both settings are generally unproblematic for learners as they can be set quickly and straightforwardly. The NDL has the potential to improve on the capabilities of a default learner by more accurately predicting a time course and learning path of setting these parameters. Because there is no default value both settings require input and, as expected, learners converge on the expected value for these parameters relatively quickly. Subject Position, Headedness in IP, Headedness in CP, Topic Marking and Preposition Stranding are such parameters in the CoLAG domain.

The e-triggers described in this section are not exhaustive of available e-triggers in the CoLAG domain. The triggers below are sufficient for correctly setting the parameter values. The decision to restrict the NDL to only one or two of many valid e-triggers was made in part for simplicity and in part to limit the computational load placed on the learner. Because the

[7] See Howitt, Dey, & Sakas (submitted) for simulations and discussion of the NDL including all thirteen CoLAG parameters.

parameters described in this section rely on unambiguous triggers, providing the NDL additional triggers for any parameter would in theory speed up the time course of learning that parameter, but not affect the overall outcome of the learner.

Subject Position

The Subject Position (SP) parameter stipulates whether the subject branches in initial or final position. In COLAG, the SP parameter can be unambiguously set given two mutually exclusive e-triggers. Sakas and Fodor use the term *globally valid* to describe triggers that are unambiguous for all languages in the domain. The globally valid trigger for subject initial is a subject preceding a direct object in a declarative when the subject is not the first overt word in the sentence. This restriction rules out the possibility that the subject was topicalized, and thus the underlying structure is subject final. The globally valid trigger for subject final is a subject following the direct object when the direct object is not in the initial position. Again, this restriction rules out the possibility that the O1 was topicalized. Subject position is one of the most quickly set parameters in the domain.

Headedness in IP

The Headedness in IP (HIP) parameter reflects headedness in IP, Negative Phrase (NegP), Prepositional Phrase (PP), and Verb Phrase (VP).[8] CoLAG contains many globally valid triggers for the IP parameter, but the NDL uses only the Prepositional Phrase and Imperative Verb Phrase.

[8] While natural languages are not always consistently initial or final for every phrase type, the consistency is a fact of the CoLAG domain and so the triggers described here suffice. However, the CoLAG domain does contain unambiguous triggers for each IP, NegP, VP, and PP so one could imagine a different domain which specifies a parameter for headedness for each phrase.

In PP, the learner looks for an adjacent P and O3. If the P precedes the O3, the learner adjusts its confidence towards the head-initial value (0), if the P follows the O3, the learner adjusts its confidence towards the head-final value (1). As with SP, restrictions on initial position prevent the learner from drawing incorrect conclusions due to topicalization.

The NDL makes use of an additional e-trigger in imperative sentences. If the Verb precedes the O1 in an imperative sentence, the languages in IP initial, if the Verb follows the O1 it is IP final.

Headedness in CP

The Headedness in CP (HCP) parameter reflects headedness in Complementizer Phrase (CP). Because COLAG lacks an overt complementizer in declarative sentences, the NDL looks only at questions when considering the HCP parameter. Headedness can be learned through the position of *-ka* in languages with an overt polar question morpheme or with Aux in languages that lack *-ka*. In non WH-questions, if the *-ka* or Aux is sentence initial, the language is C-head initial. If the *-ka* or Aux is sentence final, the language is C-head final.[9]

Topic Marking

The Topic Marking (TM) parameter reflects whether a language marks when a word has been topicalized. In COLAG, topic marking is either obligatory or unneeded. The topic marking morpheme is *-WA*. The NDL will adjust towards topic marking given the presence of *-WA*. It will adjust toward no topic marking when a word has clearly been topicalized (i.e., if an O1 and an O2 are non-adjacent) and no *-WA* is present.

[9] There are non-overt complementizers in CoLAG, but because the NDL makes use of only overt morphological or lexical items, deleted or non-phonologically realized items are not discussed further in this work.

Preposition Stranding

The Preposition Stranding (PI) parameter expresses whether the preposition and its object (O3) can be separated. If the preposition and object can be separated, the grammar allows preposition stranding. The NDL looks specifically for the presence of a preposition and object in a given sentence. If they are adjacent, the NDL conservatively adjusts its confidence toward pied piping. When presented with a preposition separated from its object, the NDL aggressively assumes the grammar licenses preposition stranding and adjusts its confidence accordingly.

5.2 Parameters with Ambiguous Triggers for One Parameter Value

Parameters for which one value has unambiguous triggers, but for which the other value lacks unambiguous triggers can be addressed efficiently by the NDL in gradient parameter space without resorting to defaults. The NDL learns from ambiguity by employing a conservative learning rate. While the aggressive learning rate is used when the NDL is presented with unambiguous evidence, the learner can also conservatively make a hypothesis (i.e., adjust its point on the parametric gradient) when presented with ambiguous information. These conservative hypotheses allow the learner to arrive at the correct hypothesis even when only ambiguous evidence is available. In the absence of unambiguous evidence, conservative learning suffices. However, when presented with unambiguous evidence, the aggressive learning rate can "rescue" an incorrect hypothesis. Crucially, learning from ambiguous evidence allows the learner to learn the correct parameter value even when two values from a subset/superset relationship.

Null Subject

The Null Subject (NS) parameter specifies whether a language allows null subjects. The NDL looks for null subjects in declarative sentences that have evidence that something other than the subject is topicalized. This stipulation removes the possibility that the subject was topicalized and then deleted. When the NDL finds evidence of a non-topicalized null subject, it aggressively adjusts its confidence interval towards null subject. However, evidence that a subject is not null (i.e., an S in the sentence) is ambiguous evidence for no null subject because null subject languages allow overt subjects in addition to null subjects. In these cases, the NDL can conservatively adjust its confidence towards no null subject. The NDL only adjusts its confidence toward no null subject when presented with a sentence that contains evidence of a topicalized element other than S, consistent with the trigger for null subject.

Null Topic

The Null Topic (NT) parameter licenses the deletion of an element when it's in topic position. In CoLAG, a Null Topic grammar must also be obligatory topic. Evidence for a null topic element in a CoLag sentence is the presence of an O2 and the absence of O1. In the presence of such a trigger, the NDL unambiguously learns the language has a null topic grammar and additionally that this grammar is obligatory topic. Evidence for no null topic is less common. A sentence with a complete set of VP complements ultimately tells the learner nothing has been deleted from the sentence.[10] This trigger suffices because of the nature of the NDL. With no

[10] Sakas & Fodor reject these so-called "Full House" sentences because in natural languages such sentences do not exist: a sentence can contain a potentially infinite number of VP complements. However, for an NDL in the CoLAG domain, I accept such triggers. In a domain with infinite VP complements the NDL could make use of other sources not present in the CoLAG domain (prosody, ellipsis, gesture, etc.), and therefore, be as effective in learning ambiguous parameters.

alternative evidence, the conservative rate will, even with a small amount of exemplars, converge on the expected parameter value.

Wh-Movement

Wh Movement (WhM) is the tendency of languages to front interrogative elements in questions. This movement is either obligatory in CoLAG or the Wh must remain in-situ. However, because all CoLAG languages have topicalization (either obligatory or optional) the WhM parameter is the canonical example of the subset/superset parameter. Wh-in situ languages will also contain the set of sentences with raised Wh.

The Wh-Superset grammars led to the necessity of proposing movement as a default in Sakas & Fodor. Because in their proposal, a defaults learner would have no way of recovering from an incorrect hypothesis, it assumes the Wh-Movement is obligatory unless presented with a Wh-in-situ.

The NDL however can correctly arrive at the target value without assuming a default. When presented with a question containing a Wh-element, the NDL looks to see whether that element occurs first in the sentence. If it does, the learner conservatively assumes movement. If the Wh-element is in-situ, the learner aggressively adjusts its confidence accordingly.

VtoI Movement

VtoI movement is the movement of the verb into the IP to get tense. In CoLAG, verbs in the verb phrase are adjacent to their direct object, when the verb moves into I, an element can intervene between them. When presented with a V that is not adjacent to O1, the NDL unambiguously learns verb movement. Verbs that do not move into IP can become tensed

through Affix Hopping or through an Auxiliary. An Aux is thus ambiguous evidence toward no VtoI as it shows the verb has remained in the VP.

Verb movement can be a complex interaction in CoLAG with Affix Hopping and ItoC (discussed below).

Affix Hopping

The Affix Hopping (AH) parameter captures the phenomenon of tense in I "hopping" down to the verb to create a finite verb. In CoLAG, grammars can either have VtoI Movement or affix hopping, but not both. Thus, an unambiguous trigger of VtoI is also an unambiguous trigger for no affix hopping.

The unambiguous trigger for affix hopping is the presence of an adverb Never that attaches above the verb phrase in the CoLAG grammar followed directly by a verb and its direct object (or, in head final languages, object verb never) in a sentence that does not contain Aux. Here, it is not the negative nature of "Never" but the position to which it attaches that is relevant. Because there is no Aux in the sentence, the verb must be tensed. The Never adverb intervenes between the I head and the V head, giving unambiguous evidence that the tense has hopped to V. Again, because a grammar cannot have both affix hopping and VtoI, this is similarly unambiguous evidence of no VtoI movement.

The NDL also takes the presence of Aux to be conservative evidence of no Affix Hopping. The Aux unambiguously shows that the tense has not moved to the verb.

6 The NDL Simulations and Results[11]

Now that I have presented the NDL and the CoLAG domain e-triggers the NDL makes use of, I will present two simulation studies to show how the NDL performs in acquiring languages in the domain. The first study simulates learning of the ten parameters on just 4 languages; the second simulates learning of those same parameters on all 3,072 CoLAG languages.

6.1 Preliminary Simulation (Howitt, Lancaster, & Sakas 2017)

To test the efficacy of the NDL in CoLAG, four languages from the domain were chosen for a preliminary simulation. For each language within each iteration, 500,000 sentences[12] were randomly chosen with replacement from a pool of all the sentences licensed by the target grammar were presented to an instance of the learner.

As reported in Howitt, Lancaster, and Sakas (2017), the acquisition process for this study consists of iterations of a simulated learner, or an '*e-child'* attempting to acquire a single CoLAG language. During each iteration, an e-child consumes sentences exclusively from their target language. The e-child searches each sentence for an e-trigger for each parameter, and, if one is found, adjusts the confidence value accordingly. For each e-child, for each parameter, confidence values (*C-value*) were initialized at 0.5.

[11] The following simulations were performed in collaboration with the CoLAG Group at The Graduate Center and Hunter College under the direction of Dr. William Sakas. While the gradient parameter space and NDL learner proposed in this work are my own, the simulations and results presented in this section would not have been possible without the input and work of Paul Feitzinger and Soumik Dey.

[12] A word order pattern in CoLAG.

```
for language as L in COLAG:
    for e-child in range(100):
        for i in range(500000):
            Randomly pick a sentence, s, from L with replacement.
            for parameter, p, in range(10):
                if e-triggers exist for p in the s:
                    Based on e-trigger adjust confidence score using
                    appropriate learning rate for p in accordance with
                    Equation (1) or (2).
```

ALGORITHM 2: NDL Simulation. Algorithm 2 presents pseudocode for the simulation driver for 100 e-children. For the simulations reported here learning rates were set to R = 0.02 and r = 0.001 and confidence values (C-value) were initialized at 0.5

In this simulation, we recorded the number of sentences it took for the NDL to reach a C-value convergence of 0.1 from the target value, learning proceeded to a maximum of 500,000 sentences. Thus, the period of learning in these studies is taken as 500,000 sentences consumed. We recorded the number of sentences to understand the amount of input necessary for the learner to converge given the type of trigger. Table 2 shows the C-value after 500,000 sentences for the four CoLAG languages. In the table, each language is named for the natural language it most resembles. *#Sentences* is the average number of sentences consumed by the NDL before reaching an imposed *C-value* threshold of 0.01 (or 0.99) or reaching the maximum 500,000 sentences.

	Subject Position		Headedness in IP		Headedness in CP		Null Subject		Null Topic	
	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences
CoLAG German	0.01	1,035	0.99	802	0.01	3,645	0.01	293,431	0.94	500,000
CoLAG French	0.01	877	0.01	598	0.01	3,230	0.01	195,525	0.01	97,716
CoLAG English	0.01	753	0.01	646	0.01	4,611	0.01	167,917	0.01	83,899
CoLAG Japanese	0.01	5,634	0.99	849	0.99	418	0.95	500,000	0.01	376,779

	WH-Movement		Pied Piping		Topic Marking		VtoI Movement		Affix Hopping	
	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences	C-Value	# Sentences
CoLAG German	0.99	32,618	0.01	7,611	0.01	7,309	0.97	500,000	0.01	1,055
CoLAG French	0.99	21,759	0.01	6,577	0.01	4,895	0.96	500,000	0.01	1,438
CoLAG English	0.99	31,026	0.99	5,508	0.01	4,180	0.01	3,359	0.99	3,818
CoLAG Japanese	0.03	500,000	0.01	9,353	0.99	458	0.01	12,845	0.01	6,420

TABLE 2: Preliminary Simulation Data. Averages of 100 simulated NDL 'e-children' with a maximum of 500,000 sentences per e-child for each of the four CoLAG languages. In the table, each language is named for the natural language it most resembles.

The NDL converged on all 10 parameters with the target value, or a weakly equivalent grammar with a non-target value. The NDL converged within 0.1 of the target value for 6 out of the 10 parameters. The four parameters on which it did not converge as strongly were Null Subject, Null Topic, Wh-Movement, and VtoI Movement. In each of these cases, the parameter value could only be set using ambiguous triggers (and thus were learned conservatively). Parameter values which were learned through unambiguous evidence arrived at the threshold of within 0.1 of the target value before 500,000 sentences.

6.2 Simulation On All 3,072 CoLAG Languages[13]

Following the success of the preliminary experiment, we simulated 100 e-children for each of CoLAG's 3,072 languages on the ten parameters outlined above. The acquisition process followed Algorithm 2 and we recorded the C-value at the end of a simulation run of 500,000

[13] Reported in Howitt, Lacaster, & Sakas (2017), expanded in Howitt, Dey, & Sakas (submitted).

sentences. For this study we did not record the sentence at which the C-value reached an arbitrary threshold.

Within this simulation we define, for a single e-child, convergence for a given parameter as the value on the gradient arrived at by the end of a simulation run (i.e., 500,000 sentences). We consider successful convergence on a target grammar, for an e-child, as approaching the target value of each of the 10 parameters; either less than 0.5 on the gradient if the target value is 0, or greater than 0.5 if the target value is 1. For grammars which contain irrelevant parameters we accept any value on the gradient as correct convergence. Results from the simulation are presented in Figure 1.

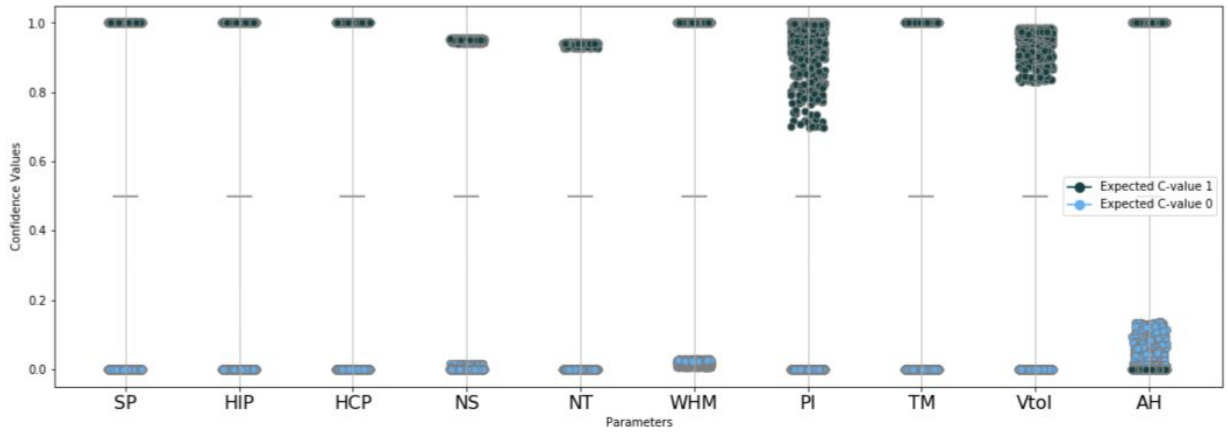

FIGURE 1: Simulation Data on 3,072 Languages. This figure contains ten one dimensional scatterplots where the y-axis shows the average confidence for each of the 10 parameters under consideration, represented on the x-axis. Each dot represents one language in the CoLAG domain. The dot's placement on the gradient is the average confidence value for 100 e-children given 500,000 sentences.
The colors represent the expected parameter value, where green is 1 and blue is 0. All languages converge toward their expected parameter values with the exception of Affix Hopping (AH) discussed below. These results were initially presented at BUCLD 42 (Howitt, Lancaster, & Sakas 2017).

To compare our learner's efficacy to other learners in the field we take convergence towards a value to effectively acquire the parameter setting delineated by the target value of that parameter, i.e., a zero or one. In this simulation each e-child converged on their target grammar or a weakly equivalent. However, it should be noted that our criteria for successful convergence is different from the grammar that is actually learned by an e-child. For example, given the learning rates used in this simulation, two e-children with different input samples (from the same language but with different distributions) differ in their actual grammars (i.e., their C-values for each parameter). However, we consider these e-children to have successfully converged on the target grammar if the C-values were in the direction of the target values. In general, the advantage of this system is the ability to represent individual or group variation based on the specific input given to individuals or groups.

As Figure 1 shows, not all e-children converged as close to the target value for each parameter as they did in the preliminary simulation. For 147 head-final, subject-final grammars all 100 e-children for each converged with an average of 0.86 for the Preposition Stranding (PI) parameter.

Although the learning of movement of a verbal element to the C head (so-called I to C movement) was not simulated for this study, it nonetheless masks the unambiguous trigger for VtoI movement in 125 languages; e-children which had one of these target grammars showed considerable variation shown in Figure 1 (though still in the direction of their target value). 47 language families, a subset of these 125 languages, also did not converge as strongly for the Affix Hopping parameter due to the same interaction.

A number of facts about the CoLAG domain not previously explicitly addressed were observed through the simulation. Obligatory affix hopping and obligatory VtoI movement are disallowed by the CoLAG UG, while obligatory ItoC movement and obligatory affix hopping are not explicitly disallowed despite the fact that obligatory ItoC movement renders the Affix Hopping parameter irrelevant for all languages (e.g. either setting of the parameter will generate the same strings -- and structures.) For all 512 languages for which this fact is relevant, AH converged toward the no affix hopping (0).[14]

As in the preliminary simulation, each of the parameter values set with ambiguous triggers fell further from the target value than those set with unambiguous triggers.

7 Additional Parameter Interactions

In a similar vein to the exploration of the interaction between ItoC movement and obligatory affix hopping discussed in the previous section, exploration of the NDL and the CoLAG domain uncovered significant interaction in other verb movement parameters that created irrelevance in the domain. Although Sakas & Fodor (2012) addressed many issues concerning the verb movement parameters, in this section I present an important finding concerning parametric interaction that Sakas & Fodor did not address. The Question Inversion (QInv) parameter specifies whether a language has ItoC movement in questions. The QInv parameter seems initially challenging because the COLAG super grammar does not prevent grammars with obligatory ItoC movement from having no movement in questions, despite the

[14] In Figure 1 these +AH +ItoC languages in green appear to be "incorrectly" set toward 0. This result is due to the specific implementation of the NDL shown here. However, a different implementation could have all -AH +ItoC languages converge towards 1 for AH. In this case 512 languages would still be "incorrectly" set. However, each of those 512 languages are strongly equivalent to the 512 with the other parameter setting for AH.

fact that ItoC movement would presumably take precedence at the sentence and structural level. All grammars with obligatory ItoC movement are *strongly equivalent* with respect to QInv. In other words, if a grammar has obligatory ItoC movement, the NDL need not learn a parameter value for Q-Inversion: either value will produce the same set of sentences and, crucially, the same set of structural trees – the QInv parameter is irrelevant to learning for any language that has obligatory ItoC movement.

To learn the value of the Qinversion parameter, the NDL need only consider questions which do not contain a WH element, i.e., polar questions. If the polar question contains *-ka* the learner knows there is no Q inversion because *-ka* occupies the C head. In polar questions that do not include *-ka*, the learner knows there is ItoC movement, because in CoLAG the C head must always be filled. In questions, if the CoLAG sentence lacks *-ka,* then either Verb or Aux raises to the C-head. Whether this raising is due to the parameter value for Q-Inversion or ItoC movement is irrelevant because either grammar would produce the same set of sentences, and the same set of corresponding parse trees.

Therefore, when accepting strongly equivalent grammars, the NDL can unambiguously set QInv. Learning in a modified domain that takes this into account would have removed significant barriers to learning for a classic triggering learner such as the one outlined in Sakas & Fodor. While they identified weak equivalency due to the ItoC parameter, they did not point out that the generated languages were strongly equivalent. In CoLAG, some parameter interactions, such as one between obligatory affix hopping and obligatory VtoI movement, are ruled out by the CoLAG UG. One can imagine a version of CoLAG that does not contain irrelevance due to the interaction of obligatory ItoC movement and QInv parameter. In any case, the NDL can

easily handle setting the QInversion parameter using only unambiguous triggers without modification to the CoLAG domain accepting strong equivalence.[15]

8 Implications for the Investigation of Linguistic Variation

The CoLAG-NDL simulations presented in this thesis demonstrate that the parametric gradient hypothesis is a viable paradigm for learning in the P&P framework. One advantage of the NDL is its ability to model variation based on facts about its input. The data presented here serves as proof-of-concept for one implementation of the gradient parameter hypothesis: the CoLAG No Defaults Learner. The fact that the distribution of input affects learning provides motivation for exploring variation and its effects in learning.

Investigation of languages where the parameter values fell further from the target value yielded significant insight into the distribution of sentence patterns in the CoLAG domain. For example, the variance in convergence for the preposition stranding parameter shed light on the availability of unambiguous e-triggers for CoLAG languages that are both head-final and subject-final. Many sentences in these languages exhibit vacuous movement in which the object of the preposition is topicalized (i.e., moved to the specifier of CP), leaving the preposition in its canonical position (dominated by VP). In the surface form the preposition and its object remain adjacent. For example the bracketed CoLAG sentence in Example (3) has a stranded O3, which is undetectable by the learner who encounters the surface form shown in Example (4).

(3) $[_{CP}\, O3\ [_{IP}\, [_{VP}\, t_{O3}\, P\ O1\ Verb]\ S]$

(4) *O3 P O1 Verb S*

[15] Had Sakas & Fodor acknowledged this strong equivalency, the Question Inversion parameter would not have been included in problem parameters Sakas & Fodor (2012; Table 1, p. 104).

Phenomena of this kind provide an interesting course of exploration when the interaction of parameters, such as head-final, subject-final, and preposition stranding, generates a specific distribution of sentence patterns. In this case, points on the parametric gradient capture that specific distribution and, with an appropriate production model, might lead to predictions of frequency in a speaker's output (idiolect), group's output (dialect), or generational differences (language drift).

Finally, as hypothesized, the NDL was also able to capture a more nuanced variation with respect to ambiguous parameter values over unambiguous parameter values. For example, Null Subject and Null Topic are not obligatory and thus do not converge as closely to 1 as obligatory elements such as headedness in phrases.

Of course, modifying the aggressive and conservative learning rates or providing a different distribution of sentences would lead to different learning outcomes. Later simulations of the NDL (Howitt, Dey, & Sakas, submitted) show the ability of the NDL to arrive within 0.1 of each parameter value whether ambiguous evidence exists for that value or not. The ideal learning rates to model child language acquisition is an empirical question and depends on the distribution of input sentences presented to the learner. Ideally, a distribution that mirrors actual child-directed speech (c.f., Sakas, Berwick & Yang, 2018) could shed some light on the efficacy of the NDL in light of a learning path. I leave this question for future study.

The model succeeds in capturing both statistical information about the input in addition to variation in input. While in this case the variation was language to language, one can see how the model could be used to explore inter speaker variation and inter generational variation. The

C-value reflected the rarity of input and thus could be thought of as a meaningful way of modeling variation in syntax.

Syntactic variation is a source of considerable academic inquiry that has recently been empirically studied, as in experimental syntax, (see, Sprouse and Hornstein, 2013 and references there). For example, Sprouse (2007) has quantified the variation in acceptability judgments by speakers of the same language. While these variations in adult language are clearly observable, their source is less clear (Lau et al., 2014, 2015, 2017; c.f., Sprouse et al., 2018). A gradient space might be one way to embody these empirical variations within a generative model of parameters, thus shedding light on the source of variation.

9 References

Bertolo, S. (2001). A brief overview of learnability. In S. Bertolo (Ed.), *Language Acquisition and Learnability* (pp. 1–14). Cambridge University Press.

Berwick, R. C. (1985). *The acquisition of Syntactic Knowledge*. MIT press.

Brown, R., & Hanlon, C. (1970). Derivational complexity and order of acquisition in child speech. In J. Hayes (Ed.), *Cognition and the Development of Language* (pp. 11–53). Wiley.

Bush, R. R., & Mosteller, F. (1955). *Stochastic Models for Learning*. Wiley.

Chomsky, N. (1981). A naturalistic approach to language and cognition. *Cognition and Brain Theory*, *4*(1), 3–22.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin, and Use*. Praeger.

Chomsky, N. (1981). *Lectures on Government and Binding*. Foris.

Chomsky, N. (1993). Lectures on government and binding: The Pisa lectures. Walter de Gruyter.

Chomsky, N. (1955/1975). Published 1955, republished 1975. *The Logical Structure of Linguistic Theory*. New York, NY & Chicago, IL: Plenum Press & University of Chicago Press.

Clark, R. (1992). The selection of syntactic knowledge. *Language Acquisition*, *2*(2), 83–149.

Clark, R. (1989). On the relationship between the input data and parameter setting. In J. Carter & R.-M. Déchaine (Eds.), *Proceedings of The 19th Annual Meeting of the North East Linguistic Society (NELS 19)* (pp. 48–62). GSLA, University of Massachusetts.

Dresher, B. (1999). Charting the Learning Path: Cues to Parameter Setting. *Linguistic Inquiry, 30*(1), 27-67.

Fodor, J. D. (1998a). Parsing to learn. *Journal of Psycholinguistic Research*, *27*(3), 339–374.

Fodor, J. D. (1998b). Unambiguous triggers. *Linguistic Inquiry*, *29*(1), 1–36.

Fodor, J. D., & Sakas, W. G. (2004). Evaluating models of parameter setting. In A. Brugos, L. Micciulla, & C. E. Smith (Eds.), *Proceedings of The 28th Annual Boston University Conference on Language Development (BUCLD 28)* (pp. 1–27). Cascadilla Press.

Fodor, J. D., & Sakas, W. G. (2005). The Subset Principle in syntax: Costs of compliance. *Journal of Linguistics*, *41*(3), 513–569.

Foraker, S., Regier, T., Khetarpal, N., Perfors, A., & Tenenbaum, J. (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric one. *Cognitive Science*, *33*(2), 287–300.

Gibson, E., & Wexler, K. (1994). Triggers. *Linguistic Inquiry*, *25*(3), 407–454.

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, *10*(5), 447–474.

Gould, I. (2017). *Choosing a Grammar: Learning paths and ambiguous evidence in the acquisition of syntax*. John Benjamins Publishing Company.

Hornstein, N. (2016). *Indirect negative evidence. Faculty of Language Blog*.

Howitt, K., Dey, S., & Sakas, W. G. (submitted). *Gradual syntactic triggering: the gradient parameter hypothesis*. Language Acquisition.

Howitt, K., Lancaster, M., & Sakas, W. G. (2017). *Doing away with Defaults: The Parametric Gradient Hypothesis*. Poster presentation. Boston University Conference on Language Development (BUCLD 42).

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*. Reidel.

Hyams, N., & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry*, *24*(3), 421–459.

Jain, S., Osherson, D., Royer, J. S., & Sharma, A. (1999). *Systems that Learn : An Introduction to Learning Theory*. MIT Press.

Lau, J. H., Clark, A., & Lappin, S. (2015). Unsupervised Prediction of Acceptability Judgements. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1618–1628.

Lau, J. H., Clark, A., & Lappin, S. (2014). Measuring gradience in speakers' grammaticality judgements. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 821–826.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: a probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk. Third Edition*. Lawrence Erlbaum Associates.

Manzini, M. R., & Wexler, K. (1987). Parameters, Binding Theory, and Learnability. *Linguistic Inquiry*, *18*(3), 413–444.

Marcus, G. F. (1993). Negative evidence in language acquisition. *Cognition*, *46*(1), 53–85.

Pinker, S. (1979). Formal models of language learning. *Cognition*, *7*(3), 217–283.

Pinker, S. (1989). *Learnability and Cognition*. Harvard University Press.

Regier, T., & Gahl, S. (2004). Learning the Unlearnable: The role of missing evidence. *Cognition*, *93*(2), 147–155.

Sakas, W. G. (2003). A Word-Order Database for Testing Computational Models of Language Acquisition. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 415–422. https://doi.org/10.3115/1075096.1075149

Sakas, W. G., Berwick, R., & Yang, C. (2018). Parameter setting is feasible. *Linguistic Analysis*, *41*(3--4), 391–408.

Sakas, W. G., & Fodor, J. D. (2007). *"Ideal" Language Learning and the Psychological Resource Problem*.

Sakas, W. G., & Fodor, J. D. (2011). *Generating CoLAG languages using the "supergrammar."* http://www.colag.cs.hunter.cuny.edu/pub/COLAG_2011_supergrammar.pdf

Sakas, W. G., & Fodor, J. D. (2012). Disambiguating syntactic triggers. *Language Acquisition*, *19*(2), 83–143.

Sakas, W. G., & Fodor, J. D. (2001). The Structural Triggers Learner. In S. Bertolo (Ed.), *Language Acquisition and Learnability* (pp. 172–233). Cambridge University Press.

Sprouse, J. (2007). *A Program for Experimental Syntax: Finding the Relationship between Acceptability and Grammatical Knowlege*. University of Maryland, College Park.

Sprouse, J., & Hornstein, N. (2013). *Experimental Syntax and Island Effects*. Cambridge University Press.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R. C. (2018). Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575–599.

Sugisaki, K. (2007). A note on the default values of Parameters. *Biolinguistics*, *1*(4), 114–117.

Sugisaki, K., & Snyder, W. (2006). The Parameter of Preposition Stranding: A View From Child English. *Language Acquisition*, *13*(4), 349–361.

Sugisaki, K., & Snyder, W. (2003). Do Parameters Have Default Values?: Evidence from the Acquisition of English and Spanish. *In The Proceedings of the Fourth Tokyo Conference on Psycholinguistics, Ed. Yukio Otsu, 215-237, Hituzi Syobo*.

Tenenbaum, J. B., & Griffiths, T. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Xu, F., & Tenenbaum, J. B. (2007). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, *10*(3), 288–297.

Yang, C. (2017). Rage against the machine: Evaluation metrics in the 21st century. *Language Acquisition*, *24*(2), 100–125. https://doi.org/10.1080/10489223.2016.1274318

Yang, C. (2002). *Knowledge and Learning in Natural Language*. Oxford University Press.