Semantic Universals in Bayesian Learning of Quantifiers

Devin Johnson

A thesis
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2021

Committee:

Shane Steinert Threlkeld

Jakub Szymanik

Program Authorized to Offer Degree:
Linguistics

University of Washington

Abstract

Semantic Universals in Bayesian Learning of Quantifiers

Devin Johnson

Chair of the Supervisory Committee:
Assistant Professor Shane Steinert Threlkeld
Linguistics

Languages undoubtedly exhibit many surface differences; However, past works such as Goddard and Wierzbicka [1994] and von Fintel and Matthewson [2008] have identified semantic properties that are evident in a vast number of languages, i.e. semantic universals. This thesis concerns itself with universal properties of quantifiers (words such as "some", "few", etc.). Although there are many possible explanations for universal properties of quantifiers, I work off of the claim that quantifier universals are explained by ease of learning, i.e. that universal properties are universal in natural language quantifiers precisely because they result in quantifiers which are more easily learnable.

In this thesis, I investigate the claim that representation length in a language of thought Fodor [1975], together with a degree of universality of a quantifier's meaning, can serve as a predictor for ease of learning and thus provide more explanation for quantifier universals. Through use of an artificial quantifier *gleeb* whose meaning is expressed in a language of thought and varied over separate experiments, I study the ease with which both human participants and a Bayesian model learn its meaning through observation of usage contexts. In the end, this method of Bayesian learning far outperforms humans at the same tasks. Results nevertheless exhibit a human and model preference for non-order-based quantifiers. In addition, although models tend toward shorter meaning representations, a degree of universality seems not to significantly affect model learning.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I owe great thanks to my adviser, Shane Steinert-Threlkeld, for his help at every step of the way in both my master's program and this master's thesis. Over the past approximately two years of working together, I have been able to see my own immense progress thanks to his support. I am also grateful for the many colleagues at the CLMBR lab and in the University of Washington CLMS program who have contributed to a friendly and supportive learning and research environment, without which this work would not have been possible. I am thankful to collaborators Jakub Szymanik, Sonia Ramotowska, and Leendert van Maanen at University of Amsterdam and Utrecht University, whose work with human experiments was vital for this thesis. Lastly, I thank Steven Piantadosi at University of California - Berkeley for his helpful comments in regard to his LOTLib3 Python tools and various Bayesian modeling details.

Chapter 1

INTRODUCTION AND BACKGROUND

This thesis examines a cognitive explanation for why certain properties in quantifiers (words such as "some", "few", etc.) are universal in natural language. That is, among all the properties quantifiers could have, why is it that only *some* properties have become universal and others have not? Steinert-Threlkeld and Szymanik [2020] suggest that because such universals become established across numerous languages, they likely result in part from more general features of human cognition. Based upon this notion, this work is rooted in the claim that universals are explained by ease of learning. More precisely put: universal properties are universal in quantifiers precisely because they result in quantifiers which are easier to learn. This claim suggests that some causal link exists between ease of learning and presence in language.

To look further into this claim, this thesis uses a Bayesian concept learning model with a modified prior which allows implementation of degrees of monotonicity and conservativity such at those seen in Posdijk [2019] and Carcassi et al. [2019]. This is used to study to what degree a quantifier meaning's minimum description length in a language of thought (LOT) along with its measured "universality" can be used to predict its ease of learning. In addition, this Bayesian model's learning capabilities are compared with those of human participants for a more complete picture. It is predicted that quantifiers with a lower LOT length and higher degree of universality are more easily learned, with model performance closer to human performance as measured by coefficient of determination (R^2); in the affirmative case, more evidence can be provided for ease of learning's connection to quantifier universals, namely, that there is evidence for a learning bias in humans for monotonic and conservative quantifiers, and not merely a bias for shorter quantifier descriptions. Though there has

been previous work in quantifier universals and learnability, this is among the first to take a Bayesian, concept-learning approach with such a modified prior. Before continuing on, I will first elucidate some of the most important background necessary for this study, as well as highlight previous work in this area.

1.1 Generalized Quantifiers

Quantifiers can be defined simply as semantic objects denoting number or, as the name implies, quantity. In surface form, these objects are expressed through determiners like "some", "few", etc.; however, quantifiers can also sometimes be nouns themselves, such as proper nouns like "John". A more flexible definition for quantifiers is that of a generalized quantifier (Barwise and Cooper [1981]), wherein a quantifier simply denotes[1] a set of sets. This is commonly used as the semantic definition of quantified noun phrases like "every student", which itself denotes the set of sets seen in the expression below.

$$\text{"Every student"} = \{X \mid \{x \mid x \textit{ is a student}\} \subseteq X\}$$

Interpreted in words, "every student" is the set of all sets X such that each X contains all the students in the universe. Given this notion of the generalized quantifier "every student", we can apply this logic to the sentence "every student is happy" and evaluate the truth of this sentence as in the expression below.

$$\begin{aligned} \text{"Every student is happy"} &\iff \{h \mid h \textit{ is happy}\} \in \{X \mid \{x \mid x \textit{ is a student}\} \subseteq X\} \\ &\iff \{x \mid x \textit{ is a student }\} \subseteq \{h \mid h \textit{ is happy}\} \end{aligned}$$

In words, "every student is happy" is true iff the set of all happy things *is one of the sets* X that contains every student in the universe. Note: there could theoretically be many such sets that contain every student. For example, if we had known apriori that every student

[1] More technically, a generalized quantifier is defined using semantic type theory as a function $\langle\langle e,t\rangle, \langle\langle e,t\rangle, t\rangle\rangle$. That is, a function which takes a set as input and returns a set of sets. For the purposes of this thesis; however, I use type theory as minimally as possible.

got an internship, the set of "internship-getting things" would also be one set containing all students in the universe. Thus, we can see that the truth conditions of sentences containing generalized quantifiers can be expressed in terms of binary relations between sets.

1.2 Quantifier Universals

As it turns out, even in a universe of only two objects, there can exist already some $\sim 60,000$ possible binary set operations that could be expressed using quantifiers (Keenan and Stavi [1986]). There are, however, only a comparatively small amount of such operations that are actually expressed in natural language. Furthermore, quantifiers expressing them exhibit universal properties, two of which will be elaborated upon below.

1.2.1 Monotonicity

Monotonicity is one proposed universal property of generalized quantifiers. The monotonicity universal (Barwise and Cooper [1981]) states that all simple determiners are monotone. Monotone quantifiers can take two forms (among others) that are specifically relevant to this thesis: monotonically increasing and monotonically decreasing.

Definition 1 *A generalized quantifier Q[2] is monotonically increasing iff given two sets X and Y, $X \subseteq Y \rightarrow (Q(X) \rightarrow Q(Y))$ (Barwise and Cooper [1981])*

Definition 2 *A generalized quantifier Q is monotonically decreasing iff given two sets X and Y, $X \subseteq Y \rightarrow (Q(Y) \rightarrow Q(X))$ (Barwise and Cooper [1981])*

To illustrate **monotonically increasing quantifiers**, take the following example. Let $X = \{x \mid x \ runs\ fast\}$ and $Y = \{y \mid y\ runs\}$. We can therefore ascertain that $X \subseteq Y$ since all things that run fast do indeed run. Now let a DP (determiner phrase) $Q =$ "every

[2]Here I assume a Q of type $\langle 1 \rangle$ for ease of explanation. This would correspond to a quantifier used in a noun phrase such as "every student", instead of simply "every", which would correspond to type $\langle 1, 1 \rangle$.

cheetah". If we let $Q(X)$ be defined as "every cheetah runs fast" and $Q(Y)$ be defined as "every cheetah runs", then by our definitions above, the expression below must be true.

$$X, Y \in \{Z \mid \{z \mid z \; is \; a \; cheetah\} \subseteq Z\}$$

With X as a subset of Y, we can see that $Q(X)$ entails $Q(Y)$. In other words, if every cheetah runs fast, then every cheetah runs. Extending this logic means that if $Q(X)$ is true, then so is every $Q(I)$ where I is a *superset* of X. This property means that Q = "every cheetah" is upward entailing and therefore a monotonically *increasing* quantifier. Note that the pattern of entailment does *not* go the other way; i.e. when $X \subseteq Y$ with the quantifier Q, $Q(Y)$ does not entail $Q(X)$ nor any other subsets.

To illustrate **monotonically decreasing quantifiers**, take the following similar example. Let $X = \{x \mid x \; runs \; fast\}$ and $Y = \{y \mid y \; runs\}$. We can again ascertain that $X \subseteq Y$ since all things that run fast do indeed run. Now let a DP Q = "no cheetah". If we let $Q(X)$ be defined as "no cheetah runs fast" and $Q(Y)$ be defined as "no cheetah runs", then by a definition similar to those above, the expression shown below must be true.

$$X, Y \in \{Z \mid \{z \mid z \; is \; a \; cheetah\} \cap Z = \emptyset\}$$

With X as a subset of Y, we can see this time that $Q(Y)$ entails $Q(X)$, unlike the opposite case in monotonically increasing quantifiers. In other words, if no cheetah runs, then no cheetah runs fast. Extending this logic means that if $Q(Y)$ is true, then so is every $Q(I)$ where I is a *subset* of Y. This property means that Q = "every cheetah" is downward entailing and therefore a monotonically *decreasing* quantifier. Note again that the pattern of entailment does *not* go the other way; i.e. when $X \subseteq Y$ with the quantifier Q, $Q(X)$ does not entail $Q(Y)$ nor any other supersets.

Finally, there is the case where a quantifier is neither monotonically increasing nor decreasing, i.e. **non-monotone**. In this case, it does not entail upward nor downward. An example of such a quantifier is "exactly one". I will not walk through such a detailed example to show this. Instead, simply recognize that under the same circumstances as the last

examples (X, Y, $Q(X)$, $Q(Y)$), neither $Q(X)$ entails $Q(Y)$ nor does $Q(Y)$ entail $Q(X)$. That is, if exactly one cheetah runs, it is not necessarily true that exactly one cheetah runs fast (no downward entailment). Similarly, if exactly one cheetah runs fast, it is not necessarily true that exactly one cheetah runs (no upward entailment).

1.2.2 Conservativity

Conservativity is another proposed universal property of generalized quantifiers. The conservativity universal (Barwise and Cooper [1981]) states all simple determiners are conservative. In terms relevant to the experiments in this thesis, given a model of the universe consisting of two sets X and Y, the conservation of this model is simply the intersection of those two sets, $X \cap Y$. We can then ascertain the definition for conservativity below. The definition of conservativity implies that the semantic denotation of a conservative quantifier Q only depends on X and $X \cap Y$ and not $X - Y$.

Definition 3 *A generalized quantifier* Q[3] *is conservative iff given two sets* X *and* Y*,*

$$Q(X, Y) \iff Q(X, X \cap Y)$$

To exemplify a conservative quantifier, take the case of the quantifier "every". If we assume X is the set of dogs and Y is the set of things that eat, then note that the denotations of the following sentences are equivalent by conservativity:

1. Every dog eats

2. Every dog is a dog who eats

In other words, when applying "every" to "dogs" and "things that eat", we receive the same truth values as applying "every" to "dogs" and "dogs who eat".

---

[3]Here Q denotes a type $\langle 1, 1 \rangle$ function, as opposed to a type $\langle 1 \rangle$ function as assumed above.

1.3 Quantifier Learning

There is a history[4] of work regarding the learning of novel quantifiers using experiments with artificial quantifiers such as *gleeb*; however, for this thesis, I focus on extending approaches which attempt to establish a connection between ease of learning and universality with both machine and human learning experiments. In addition, I base my learning model upon previous work in Bayesian concept learning.

1.3.1 Bayesian Concept Learning in a Language of Thought

In cognitive science literature, concept learning can be defined as "...the search for and listing of attributes that can be used to distinguish exemplars from non exemplars of various categories." (Bruner et al. [1956]). From this definition and several subsequent works in cognitive science, we see that concepts are marked by a few defining features; according to Goodman et al. [2008a], concepts are firstly "...mental representations that are used to discriminate between objects, events, relationships, or other states of affairs". Secondly, concepts are "learned inductively from the sparse and noisy data of an uncertain world". Lastly, concepts are "formed by combining simpler concepts, and the meanings of complex concepts are derived in systematic ways from the meanings of their constituents". Based on this definition, examples of concepts may therefore include things as common as our notion of number, given that it is compositional, used to discriminate between objects, and is able to be learned among sparse and noisy data. Given that these concepts form an integral basis for our functioning in the world, we may naturally ask how they can be acquired. For example, how is it that we may learn to use the concepts of color and number to distinguish "two red circles" from "three yellow circles"? Studies which investigate the acquisition of these capabilities fall under the umbrella of concept learning.

Bayesian concept learning, as introduced in Goodman et al. [2008a], is a method for describing the acquisition of concepts, which has shown success in several areas of cognitive

[4]See: Hunter and Lidz [2012], Spenader and Villiers [2019], Chemla et al. [2019]

science through its combination of both logical representations of concepts along with statistical inference. At its core, this method of learning constructs representations using logical primitives and infers the most likely meanings given sample data. An example of such a study is Piantadosi et al. [2012], where the authors apply their formulation to the acquisition of quantifier meanings in natural language, i.e. the concept of quantity/number. This work argues that listeners, when determining the meaning of a quantifier in speech, perform Bayesian inference over the hypothesis space of possible meanings, which are represented by strings consisting of logical primitives. This notion of Bayesian learning over a space of logically-constructed hypotheses has been expanded upon in recent years to lead to a framework for Bayesian concept learning in an LOT (Fodor [1975]). This formulation seeks to combine Bayesian inference techniques with more specified conceptual representations based on earlier notions of conceptual thought in cognitive science, using such languages of thought as "...psychological theories that are subject to empirical evaluation". An example of this can be seen in Piantadosi et al. [2016], which features an array of concept learning experiments with usage of this framework. Importantly for this thesis, this framework provides reliable modeling of key phenomena in human concept learning, being able to predict more fine-grained aspects of human learning curves. In addition, through its usage of a modifiable, language-of-thought prior, it is possible to experiment with priors (possible quantifier meanings) which satisfy certain semantic universals; this proves key for the work shown in this thesis.

1.3.2 Ease of Learning and Semantic Universals

The study of linguistic universals is the study of properties found in a vast number natural languages. Such studies[5], along with those which focus on semantic universals, are not novel.

In recent years, several advancements have been made not only in identifying semantic universals, but also in studying explanations for their existence. Most notably for this work

[5]Examples include Goddard and Wierzbicka [1994] and von Fintel and Matthewson [2008]

is Steinert-Threlkeld and Szymanik [2020], where, by use of an RNN learning model, preliminary evidence is shown that quantifier universals can be explained by ease of learning. Importantly for this thesis, it is noted that there are multiple other possible explanations for quantifier universals — not just the ease with which an RNN model learns a quantifier's meaning. In particular, one of such methods mentioned is the minimum description length in an LOT, as seen in Piantadosi et al. [2016]. Steinert-Threlkeld and Szymanik note: "...while neural network learnability provides a unified explanation of semantic universals in many different domains, whether or not other general notions of complexity from the literature also can remains open...It is, for instance, possible that minimal description length in an appropriate LOT can explain both the presence of the universals and their ease of learning." This direction is precisely where this thesis continues toward. It has been shown in earlier work that minimum description length in an LOT is a reliable predictor of learnability; however, in addition to utilizing description length and its relation to learnability, consideration is also taken to include measures of monotonicity and conservativity in order to test whether there exists a human learning bias for quantifiers with these universal properties.

Chapter 2

METHODS

To test the claim that that representation length in an LOT together with degrees of monotonicity and conservativity of a quantifier's meaning can serve as a predictor for ease of learning, this thesis introduces a Bayesian concept learning model and human quantifier learning experiments of an artificial quantifier, *gleeb*.

2.1 Human Learning Experiments

2.1.1 Experimental Setup

In work conducted at University of Amsterdam by Jakub Szymanik, Sonia Ramotowska, and Leendert van Maanen, 30 participants were asked to complete 8 artificial quantifier learning experiments; the experiments featured varying meanings of the unknown quantifier *gleeb* including: "first 3", "between 3 and 6", "at most 2", etc. In these experiments, participants were asked to learn the meaning of the quantifier, *gleeb*, through observation of 96 of its use contexts. Participants were shown these contexts in succession, each featuring $n \in [1, 8]$ objects each. Objects shown were of shape $\in$ {circle, triangle}, and varied in color $\in$ {red, blue}. Upon seeing each context, the human participants were asked to identify whether the sentence, "Gleeb triangles are red", is true or false. An example of this is shown in Figure 2.1. After answering true or false, the participant is shown the correct answer; this is, therefore, a form of supervised learning. At the end of each experiment, performance is measured as the percentage of humans getting the correct answer at a given context. For example, at the 45th context (out of 96), there may have been 60% of humans who answered correctly, thus 60% would be the human accuracy measure for this context.

In the Bayesian theoretical framework we first presume that a participant has prior

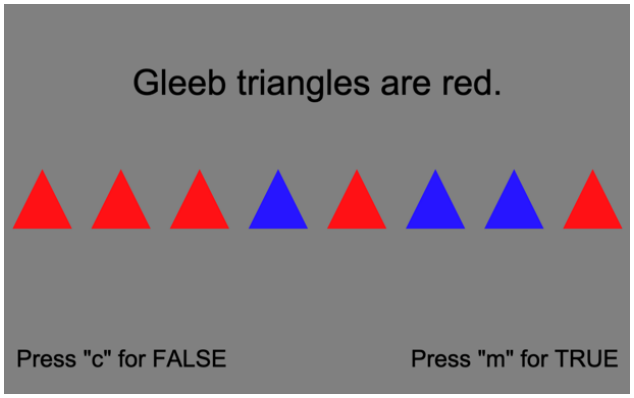


Figure 2.1: Example context (Szymanik, Ramotowska)

assumptions about what the meaning of *gleeb* could be. That is, the participant has a space of possible hypotheses of word meanings and their probabilities *already* constructed in their mind. Then, upon seeing each context, they attempt to use the contexts (i.e. data) seen so far. In other words, the participant is presumed to be performing inference using the data seen to search over this space of possible meanings to search for a suitable meaning for *gleeb*. "Suitable" in this case would, of course, mean most likely when combined with the participant's apriori assumptions. With more data, it is therefore plausible that a participant can gradually learn the meaning of *gleeb* by updating their prior assumptions through trial and error. Of interest for this thesis, is whether a bias for monotonicity and conservativity (besides merely a bias for shorter quantifier meanings) is included in the participant's apriori assumptions.

2.2 A Bayesian Concept Learning Model

2.2.1 Modeling Framework

After completing human experiments, Bayesian concept learning models[1] are trained over the same data as those shown to human participants. For the study, there are as many

[1]Repo can be accessed on Github

models trained as there are humans, and their performance is measured through a posterior predictive probability to approximate percentage of humans correct per data seen, as used in previous work in this learning scenario. The goal of the models is to use given data to learn the probability distribution of hypotheses which are quantifier meanings expressed in a language of thought (i.e. meanings expressed with logical primitives like $\cap$, $\cup$, etc.), or $P(h|d)$. As the nomenclature of the learning style suggests, this can be calculated using Bayes' rule:

$$P(h|d) \propto P(d|h)P(h)$$

First, just as in the human case, a prior probability $P(h)$ is calculated to model the apriori assumptions of the model about possible meanings of *gleeb*. In this case, meanings of *gleeb* are represented by strings in an LOT and weighted by length such that longer meanings are less likely. Since the prior plays a large part in this study, it is explained in more depth in Definition 6. Next, given the same data as one of the humans (one model per human), the likelihood $P(d|h)$ models the probability of hypotheses given the current data seen. In other words, it answers the question: "Given what I see in this context, how likely is it that the meaning of *gleeb* is this one?". Following common practice for Bayesian learning experiments, a single likelihood calculation[2] is performed over each context (data point) with $\alpha \in [0, 1]$ representing noise in the data:

$$\begin{array}{rl} log(\alpha) & \text{If meaning true in context} \\ log(1-\alpha) & \text{Else} \end{array}$$

Finally, using a Metropolis-Hastings sampler, the prior and likelihood are used to estimate the posterior distribution $P(h|d)$ over k steps of sampling. This process is repeated for each model, corresponding to each human, and is run over successive chunks of data, mimicking how humans are presented the data in successive screens. That is, likelihood calculations and

[2]For a more detailed explanation, see Goodman et al. [2008b]

sampling are performed every time a new context is seen. At each context seen, the top 100 hypotheses (ranked by posterior score) are stored. For example, as the model comes across the 50th successive context, by using data from 0 contexts through the first 49 contexts, it will infer the top 100 hypotheses for the 50th context. After all of the data has been seen in this manner, the top 100 hypotheses from all numbers of contexts seen is used to form a fixed hypothesis space over which a predictive posterior probability for each number of contexts seen can be calculated. This fixed hypothesis space is pruned to eliminate hypotheses which are semantically equivalent before calculating posterior probabilities. For example, if there were a total of 10 contexts, 1000 hypotheses would be in the fixed hypothesis space before pruning, i.e. 100 per each context. By calculating the posterior predictive probabilities over each amount of data seen, a measure of model performance is achieved at each amount of data seen. This can then be plotted and compared with human accuracy. Of course, in this modeling framework, posterior probabilities (and thus learnability) may highly depend on differing definitions of a prior, which must be taken into account. Therefore, defining a prior in more detail will be the task of the next subsection.

2.2.2 A Weighted LOT Prior

As this thesis attempts to establish a connection between learnability and quantifier universals, it is undoubtedly important to specify how quantifier meanings are represented, and how universality in these meanings is to be formulated. In this setup, as mentioned before, hypotheses (possible meanings of the *gleeb*) are represented as strings in a language of thought. This language of thought consists of lambda expressions over two sets (A and B) and set-theoretic primitives such as $\cup$, $\cap$, and more. In this way, hypotheses can be constructed using combinations of these set-theoretic primitives embedded into lambda expressions. Set A is to be used to contain all triangles in a context, and set B is used to contain all red objects. These LOT expressions represent a learner's apriori assumptions about which meanings are possible, as well as a probability distribution $P(h)$ where longer meanings are less probable. The inner workings of this formulation of $P(h)$ depend on a

probabilistic context-free grammar (PCFG) G, which is formalized by $G = (M, T, R, S, P)$:

- M is a set of non-terminal symbols. Non-terminal symbols must be expanded fully until reaching a string with only terminals in order to obtain a well-formed parse (expression in the language of thought).
- T is a set of terminal symbols, which are primitives in the language of thought. Terminal symbols, by definition, cannot be further expanded in a parse tree.
- R is a set of production rules. Production rules define how non-terminals can expand into strings of non-terminals, terminals, or a combination of both.
- S is the start symbol. In other words, S is the non-terminal from which all parses must start.
- P is a set of production rule probabilities. Probabilities in P are defined uniformly, favoring hypotheses which are shorter in length since production rule probabilities are multiplied as a parse tree grows deeper.

The non-terminals M, terminals T, and productions R which were used for this thesis are summarized in Figure 2.2 below. For example, given Figure 2.1 above, we might assume that the meaning of *gleeb* is "more than 2". If that is the case, and a set A is made up of all triangles in the context and a set B is the set of all red things in the context, a possible hypothesis[3] for this quantifier's meaning would thus be $\lambda\, A, B : |A \cap B| > 2$. Similarly, another hypothesis such as $\lambda\, A, B : |(A \cap B) \cap (A \cap B)| > 2$ would be valid — though it may not be the best once we examine other contexts in which *gleeb* is used. Using these two example hypotheses and our definition of a prior, one can ascertain that the first would have a smaller prior probability than the second, since each primitive used brings us further down a PCFG parse tree, in turn multiplying more probabilities together to create a smaller and smaller resultant prior probability. Using this framework, the goal of the Bayesian learner, much like the human learner, is to find which of these possible hypotheses in the language of thought is most fitting for the meaning of *gleeb* across all contexts which form the dataset.

[3] I use slightly different notation from the PCFG above for readability purposes. The interpretation is not affected.

BOOL → subset(SET, SET)
→ card-lt(CARD, NUM)
→ card-gt(CARD, NUM)
→ card-lteq(CARD, NUM)
→ card-gteq(CARD, NUM)
→ and(BOOL, BOOL) | or(BOOL, BOOL) | not(BOOL)
CARD → cardinality(SET)
SET → intersection(SET,SET)
→ union(SET,SET)
→ set-difference(SET,SET)
→ a | b
NUM → 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10

Figure 2.2: PCFG used for this thesis; production probabilities uniform, non-terminals are written in capital letters, terminals are written in lowercase letters. The terminals a and b correspond to sets A and B. Production probabilities are assumed to be uniform. The start symbol is BOOL.

In addition to this weighting based upon length, also included are degrees of monotonicity and conservativity, as defined below and adapted from Carcassi et al. [2019] and Posdijk [2019]. For the degree of monotonicity in Definition 4, I employ random variables $\mathbb{1}_Q$, $\mathbb{1}_{\overrightarrow{Q}}$ and the space of possible models as follows: $\mathbb{1}_Q$ is the value that the quantifier Q assigns to a given context. $\mathbb{1}_{\overrightarrow{Q}}$ is whether a context has a subcontext that the quantifier considers true. $H(\mathbb{1}_Q)$, quantifies the uncertainty about what truth value Q will assign to a context. The conditional entropy $H(\mathbb{1}_Q|\mathbb{1}_{\overrightarrow{Q}})$ quantifies the uncertainty about what Q will assign to a context, given that one knows whether the model has a subcontext that Q considers true. For example, $H(\mathbb{1}_Q|\mathbb{1}_{\overrightarrow{Q}})$ is minimized (0) for

an upward monotone quantifier since if it is known that a context has a true subcontext, and the quantifier is upward monotone, the quantifier must be true in that context. The mutual information, $I(\mathbb{1}_Q; \mathbb{1}_{\overleftarrow{Q}}) := H(\mathbb{1}_Q) - H(\mathbb{1}_Q|\mathbb{1}_{\overleftarrow{Q}})$, measures how much information $\mathbb{1}_{\overleftarrow{Q}}$ provides about $\mathbb{1}_Q$. For example, given an upward monotone quantifier, the mutual information is simply equal to $H(\mathbb{1}_Q)$. This can then be used to define a degree of (upward) monotonicity as $\frac{I(\mathbb{1}_Q;\mathbb{1}_{\overleftarrow{Q}})}{H(\mathbb{1}_Q)}$, which with some manipulation, is equivalent to the form seen in Definition 4. As an extension of Carcassi et al. [2019], the max of both upward and downward monotonicity measures is taken, where our downward measure is calculated in the same manner, but searches in supercontexts instead of subcontexts in order to calculate the distribution of $\mathbb{1}_{\overrightarrow{Q}}$.

Definition 4 $Mon_Q := \max\{1 - \frac{H(\mathbb{1}_Q|\mathbb{1}_{\overleftarrow{Q}})}{H(\mathbb{1}_Q)}, 1 - \frac{H(\mathbb{1}_Q|\mathbb{1}_{\overrightarrow{Q}})}{H(\mathbb{1}_Q)}\}$

To calculate a degree of conservativity, the exact same formulations as above are used, but with another random variable $\mathbb{1}_Q^{con}$). This random variable maps to either true or false depending on whether the quantifier is true in the conservation of a given context. The conservation of a context consisiting of two sets, A and B, is simply the intersection of these two sets, $A \cap B$.

Definition 5 $Con_Q := 1 - \frac{H(\mathbb{1}_Q|\mathbb{1}_Q^{con})}{H(\mathbb{1}_Q)}$

Thus, given a quantifier Q's prior (log) probability p_Q (as determined by its length in the LOT), these measures are combined with hyperparameters λ_1, $\lambda_2 \in [0, 1]$ to obtain a final weighted prior probability for a proposed quantifier's meaning:

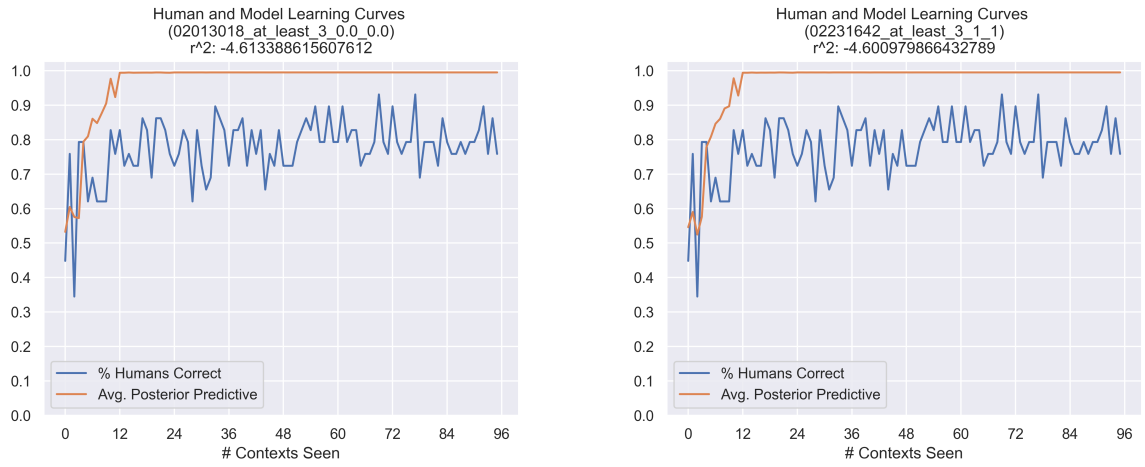
Definition 6 $P_W := p_Q + \lambda_1(log_2(Mon_Q))) + \lambda_2(log_2(Con_Q))$

Chapter 3

RESULTS

Nearly all models were quickly able to learn the meaning of *gleeb*, with most achieving near 100% posterior predictive probabilities within seeing 10 use contexts of the quantifier and sample steps $k = 1000$. This number of sample steps is, in comparison to previous studies with $k = 100,000$, far lower, which would normally cause worse performance. Nevertheless, the models achieve high performance on *both* sample steps settings, as evidenced by learning curves averaging their posterior predictive probabilities over each number of contexts seen. High performance also holds even as λ_1 and λ_2 are raised in value to signify more weight toward degree of monotonicity and degree of conservativity. Finally, among all models, it was clear to see a bias for shorter hypotheses, which was expected and is consistent with previous studies. The pattern of high performance, which held for all non-order-dependent quantifiers, is exemplified in Figure 3.1 and Figure 3.2. Such a performance pattern may be due to the order in which data was presented to humans and models and/or difficulty of the concepts being learned.

However, notably, this performance pattern excludes models for quantifiers whose meanings are order-dependent. The case of order-dependent quantifier meanings can be seen in Figure 3.3. In this case, model performance better fits human performance as measured by an r^2 closer to zero (less negative), though this is might due to the fact that a majority of the human subjects did not posit order-dependent quantifier meanings (or at least did not report doing so) and that the grammar used in these models to generate hypotheses did not posses the logical primitives to construct such meanings, rather than an indication of fit due to other reasons.

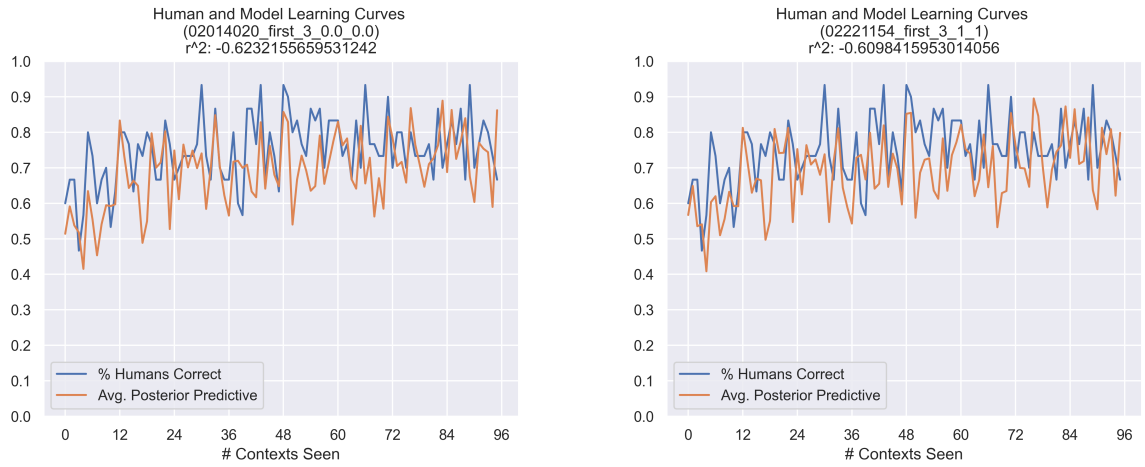


(a) (b)

Figure 3.1: Learning curves for *gleeb* meaning of "at least 3 ". The y axis represents probabilities from $[0, 1]$. The average posterior predictive is plotted in yellow, which is intended to fit to the average human learning curve in blue. $n = 30$, $k = 1000$. r^2 is provided for curve similarity. This curve shows the high performance of the Bayesian learner. This pattern of learning appeared in all non-order-dependent quantifiers tested, not just "at least 3". **(a)** $\lambda_1 = \lambda_2 = 0$, **(b)** $\lambda_1 = \lambda_2 = 1$

(a) (b)

Figure 3.2: Learning curves for *gleeb* meaning of "between 3 and 6". The y axis represents probabilities from $[0, 1]$. The average posterior predictive is plotted in yellow, which is intended to fit to the average human learning curve in blue. $n = 30$, $k = 1000$. r^2 is provided for curve similarity. This curve shows the high performance of the Bayesian learner. This pattern of learning appeared in all non-order-dependent quantifiers tested, not just "between 3 and 6". **(a)** $\lambda_1 = \lambda_2 = 0$,**(b)** $\lambda_1 = \lambda_2 = 1$



(a) (b)

Figure 3.3: Learning curves for *gleeb* meaning of "first 3". The y axis represents probabilities from $[0, 1]$. The average posterior predictive is plotted in yellow, which is intended to fit to the average human learning curve in blue. $\lambda_1 = 0$, $\lambda_2 = 0$, $n = 30$, $k = 1000$. r^2 is provided for curve similarity.**(a)** $\lambda_1 = \lambda_2 = 0$,**(b)** $\lambda_1 = \lambda_2 = 1$

Chapter 4

DISCUSSION

To the question of whether a bias for monotonicity and conservativity serve to better model human learning patterns of quantifiers, there still remains no clear answer. Overall at fault is the fact that the learning setup formulated in this thesis does not allow proper investigation into this question. Certainly, model posterior predictive probabilities could have been adjusted to be lower to better match human performance, though the quality and the quantity of such changes (e.g. reducing sample steps to 1, forcefully reducing model confidence by adjusting α noise level, etc.) would have amounted to severely underpowering the model, which would make the results far less credible.

These results differ from results seen in previous quantifier concept learning experiments such as Piantadosi et al. [2016]. Though there are a multitude of reasons that this could occur, barring more obvious model errors, one reason may lie in the manner in which data was presented to both models and humans. More specifically, in previous studies, the *same data* was presented to all humans and in the *same order*, rather than IID random samples for each human. Posterior predictive probability of a context given all previous contexts was therefore intended to capture the percentage of people correctly guessing the context. In this experimental setup, this percentage is not explicitly represented, since sets of data shown to humans were random. Given this, estimates of this percentage are undoubtedly noisier. Another possible explanation for high performance in models is simply that the target concepts were easily learnable.

What can be confidently stated from the results seen is that, given this data, Bayesian concept learning can easily and reliably predict the meaning of a novel quantifier *gleeb*. This conclusion does not ultimately differ much from previous studies, however, it is useful in that it now excludes the question of whether Bayesian concept learning is fruitful for this research question. In future iterations of this work, clear next steps will be to test different modeling architectures, and perhaps reiterate Bayesian methods with altered human experimentation/data setup. Other modeling

architectures could include simple neural networks, linear classifiers, etc. If it is indeed the case that target concepts (i.e. quantifier meanings like "between 3 and 6", etc.) are simply too easily learnable, then it should be apparent that other models present similar over-performance problems.

Chapter 5

CONCLUSION

In this thesis, I've introduced novel methods for investigating explanations for semantic quantifier universals by using a Bayesian concept learning model over hypotheses constructed in a language of thought. Results showed that this method of learning does not serve as a fruitful direction for investigating this problem, as model performance highly exceeds human performance, making direct comparisons of learning curves of little use to the question. In future work, it will undoubtedly be important to evaluate varied model learning setups, even those such as simple linear classifiers. In addition, future iterations of this work ought to test Bayesian methods on varying human experimental/data setups. Though the results in this thesis do not directly answer the original question at hand, they serve as a useful and necessary step in leading the broader work of explanations for quantifier universals toward a more insightful destination.

BIBLIOGRAPHY

Jon Barwise and Robin Cooper. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219, 1981. ISSN 01650157, 15730549. URL http://www.jstor.org/stable/25001052.

J. Bruner, J. Goodnow, and G. Austin. *A Study of Thinking*. 1956.

Fausto Carcassi, Shane Steinert-Threlkeld, and Jakub Szymanik. The emergence of monotone quantifiers via iterated learning, May 2019. URL psyarxiv.com/8swtd.

Emmanuel Chemla, Brian Buccola, and Isabelle Dautriche. Connecting Content and Logical Words. *Journal of Semantics*, 36(3):531–547, 03 2019. ISSN 0167-5133. doi: 10.1093/jos/ffz001. URL https://doi.org/10.1093/jos/ffz001.

Jerry A. Fodor. *The Language of Thought*. Harvard University Press, 1975.

Cliff Goddard and Wierzbicka. *Semantic and Lexical Universals: Theory and empirical findings*. John Benjamins, 1994. URL https://www.jbe-platform.com/content/books/9789027285782.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008a. doi: https://doi.org/10.1080/03640210701802071. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210701802071.

Noah D. Goodman, Joshua B. Tenenbaum, Jacob Feldman, and Thomas L. Griffiths. A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1):108–154, 2008b. doi: 10.1080/03640210701802071. URL https://onlinelibrary.wiley.com/doi/abs/10.1080/03640210701802071.

Tim Hunter and Jeffrey Lidz. Conservativity and Learnability of Determiners. *Journal of Semantics*, 30(3):315–334, 08 2012. ISSN 0167-5133. doi: 10.1093/jos/ffs014. URL https://doi.org/10.1093/jos/ffs014.

Edward Keenan and Jonathan Stavi. A semantic characterization of natural language determiners. *Linguistics and Philosophy*, pages 253–326, 1986. URL https://doi.org/10.1007/BF00630273.

Steven Piantadosi, Joshua Tenenbaum, and Noah Goodman. The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, 123, 04 2016. doi: 10.1037/a0039980.

Steven T. Piantadosi, Joshua B. Tenenbaum, and Noah D. Goodman. Modeling the acquisition of quantifier semantics : a case study in function word learnability. 2012.

Wouter Posdijk. The influence of the simplicity/informativeness trade-off on the sematic typology of quantifiers, 2019.

J. Spenader and J. Villiers. Are conservative quantifiers easier to learn ? : Evidence from novel quantifier experiments. 2019.

Shane Steinert-Threlkeld and Jakub Szymanik. Ease of learning explains semantic universals. *Cognition*, 195:104076, 2020. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2019.104076. URL http://www.sciencedirect.com/science/article/pii/S0010027719302495.

Kai von Fintel and Lisa Matthewson. Universals in semantics. *The Linguistic Review*, 25(1-2):139 – 201, 2008. doi: https://doi.org/10.1515/TLIR.2008.004. URL https://www.degruyter.com/view/journals/tlir/25/1-2/article-p139.xml.