

UC Berkeley

UC Berkeley Electronic Theses and Dissertations

Title

Statistical models of learning and using semantic representations

Permalink

<https://escholarship.org/uc/item/4f86w8q8>

Author

Abbott, Joshua Thomas

Publication Date

2016

Peer reviewed|Thesis/dissertation

Statistical models of learning and using semantic representations

by

Joshua Thomas Abbott

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Psychology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Thomas L. Griffiths, Chair

Tania Lombrozo

Mahesh Srinivasan

Terry Regier

Summer 2016

Statistical models of learning and using semantic representations

Copyright 2016
by
Joshua Thomas Abbott

Abstract

Statistical models of learning and using semantic representations

by

Joshua Thomas Abbott

Doctor of Philosophy in Psychology

University of California, Berkeley

Thomas L. Griffiths, Chair

How does cognition organize sparse and ambiguous input from the environment into useful representations and concepts for understanding the world? The work in this dissertation explores how people learn and reason with abstract knowledge, focusing on the kinds of processes and representations used in semantic memory. In particular, I present three case studies, each investigating different assumptions for the semantic representations and algorithms used to model cognition. The first chapter introduces this work situated in the framework of probabilistic models of cognition and outlines the goals of each case study. The second chapter focuses on the distinction between process and representation in semantic memory search. The simulations and analyses in this chapter show that behavioral results on a semantic fluency task previously explained as a strategic search process can also be produced by a non-strategic search process, depending on the structure of representation used for semantic memory. The third chapter investigates semantic representations as a means to explore universals and variation in cognition across cultures. In this chapter, I present a simple computational model operating over an irregularly shaped perceptual color space which accounts both for universal tendencies and for variation in focal colors, or best examples of color terms, across the world's languages. The fourth chapter explores the challenges of developing models that can learn to appropriately apply new labels to concepts from only a few example observations, like people. Building upon a successful Bayesian word learning model, I propose adapting large-scale knowledge representations, typically used in machine learning and computer vision, to automatically construct hypothesis spaces for generalization models that account for these challenges. Finally, in the fifth chapter, I discuss the theoretical and practical implications for this body of work as a whole, and suggest a variety of future directions. Taken together, this research suggests general principles of computation over structured knowledge representations illuminates how people make sense of the world around them, and may lead to developing machines that think more like people do.

Acknowledgments

I would like to foremost thank my advisor Tom Griffiths for his constant guidance and support. I don't know if I'll ever be able to express how grateful I am for the opportunity to study at Berkeley and learn from him. I would also like to extend my gratitude to the rest of my dissertation committee, Tania Lombrozo, Mahesh Srinivasan, and Terry Regier. Their comments and suggestions have helped make the work in this dissertation far surpass anything I could have done alone. I thank the friends, gurus, and labmates I've met along the way. Finally, none of this would have been possible without the unconditional love from my parents.

Table of Contents

List of Figures	iv
List of Tables	viii
1 Introduction	1
1.1 General introduction	1
1.2 Goals of the present dissertation	3
2 Process and representation in semantic memory search	7
2.1 Introduction	7
2.2 Optimal foraging as an account of semantic fluency	9
2.3 Initial comparison of optimal foraging and random walks	10
2.4 Exploring a different semantic representation	12
2.5 The importance of clustering	16
2.6 Discussion	20
3 Universals and variation in color categories	23
3.1 Introduction	23
3.2 Predicting best examples of color categories	24
3.3 Results	28
3.4 Language level analysis	31
3.5 Color categories with unusual extensions	33
3.6 Discussion	37
4 Large-scale word learning	38
4.1 Introduction	38
4.2 The Bayesian generalization framework	39
4.3 Constructing a hypothesis space for Bayesian word learning	42
4.4 Behavioral experiments to validate our approach	44
4.5 Large-scale word learning	47
4.6 Discussion	50
5 Conclusions	51

5.1	Remaining questions and future directions	51
5.2	Broader theoretical and practical implications	53
5.3	Concluding remarks	56
Bibliography		57
A Supplementary analyses of semantic memory search		67
A.1	Effects of degree distributions and edge direction	67
A.2	Effects of connectivity structure	69
B Supplementary analyses of the World Color Survey		74
B.1	Language-level analyses	74
B.2	Category Unusualness	96

List of Figures

- 2.1 Experimental results of a semantic fluency experiment (free recall from the category of animals) reproduced from Hills et al. (2012). (a) The mean ratio between the inter-item response time (IRT) for an item and the participant’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the participant’s average IRT for the entire task. (b) The relationship between a participant’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a participant produced. 10
- 2.2 Results from 141 simulations of the random walk model from Hills et al. (2012), submitted to the same analyses as their human data. (a) The mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the participant’s average IRT for the entire task. (b) The relationship between a participant’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a participant produced. 13
- 2.3 Results after 141 simulations for the four random walk models: (a) the uniform transition model with no jumps, (b) the weighted transition model with no jumps, (c) the uniform transition model with a jump probability of 0.05, and (d) the weighted transition model with a jump probability of 0.05. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced. 17

2.4	(Top row) A visualization of the similarity between pairs of animals in the semantic network (left panel) and an additive clustering model (right panel), where darker colors represent stronger similarities. (Bottom row) A visualization of the BEAGLE animal similarity space (left panel) and an additive clustering model (right panel). The rows and columns of the each matrix were reordered to display animals in the clusters with largest weight first.	19
2.5	Results for the minimal model on our semantic network with p estimated from the uniform transition word association matrix. (a) The mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a patch and the last word in the preceding patch). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. (b) The relationship between a walker’s deviation from the marginal value theorem policy for patch departures (horizontal-axis) and the total number of words a walker produced.	21
3.1	(a) Color naming stimulus array. The rows correspond to 10 levels of Munsell value (lightness), and the columns correspond to 40 equally spaced Munsell hues. The color in each cell corresponds approximately to the maximum available Munsell chroma for that hue-value combination. (b) The chips of the stimulus array as plotted in CIELAB color space. The irregularity of the outer surface of the color solid can be seen, most notably in the yellow region.	25
3.2	Contour plots of the focus distributions in (a) the WCS, and as predicted by (b) the representativeness model, (c) the likelihood model, (d) the prototype model, (e) the exemplar model, and (f) the chroma model. Each contour line corresponds to 100 focal choices.	29
3.3	Naming data for the Dani language, overlaid with the empirical focus distributions for <i>mili</i> and <i>mola</i>	32
3.4	Naming data for the Berinmo language, overlaid with the empirical focus distribution.	34
3.5	Effect of category unusualness. Left panels (scatterplots): Each dot represents a color category in the WCS, and the dot’s color represents the best-performing model for that category. The horizontal axis represents category unusualness, and the vertical axis represents the model performance: rank position (top panel) and QF distance (bottom panel) of the best-performing model for that category. Right panels (bar charts): The horizontal axis again represents category unusualness, this time partitioned into 10 bins with the same number of categories per bin. The stacked bars show, for each level of unusualness, the proportion of categories at that level of unusualness that were best predicted by each model.	36
4.1	Example word learning experiment trial using ImageNet as a source of stimuli.	45

4.2	Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the three domains in Experiment 1. The generalizations for non-matching items are omitted for brevity (neither the participants chose nor the Bayesian model predicted non-matching objects, while the prototype and exemplar models predicted non-matches less than 4% of the time for each condition).	45
4.3	Participant generalization judgments and the predictions of the Bayesian model for Experiment 2. From left to right, the columns present the results for the three taxonomies (clothing, containers, and seats) and average results. Non-matching items are omitted for brevity (participants only chose non-matches twice, both in the containers domain).	47
4.4	Concepts constructed from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Training set images sampled from the four levels for the leaf node <i>blueberry</i> , with the levels L_0, \dots, L_3 corresponding to the concepts <i>blueberry</i> , <i>berry</i> , <i>edible fruit</i> , and <i>natural object</i> , respectively.	49
4.5	Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the 4,000 concepts in Experiment 3. The horizontal axis for each presents four levels at which training set examples were provided (L_0 to L_3). At each level, five bars show the proportion of test set images from levels L_0 to L_4 that were selected as instances of the concept (where L_4 denotes non-matching items), with the results averaged over all domains.	50
A.1	Results for the uniform non-jumping model and uniform jumping model on the reverse-directed network. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced.	70
A.2	Results for (a) the uniform non-jumping model and (b) uniform jumping model on the undirected network. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced.	71

A.3	Results for (a) the uniform non-jumping model and (b) uniform jumping model on our semantic network with random node relabelling. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced. .	73
B.1	The 30 most unusual WCS categories (presented in descending order of average Hausdorff distance)	97
B.2	The 30 least unusual WCS categories (presented in ascending ranked order of average Hausdorff distance)	98
B.3	WCS categories in the 25th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)	99
B.4	WCS categories in the 50th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)	99
B.5	WCS categories in the 75th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)	99

List of Tables

3.1	Quantitative assessment of each model against WCS focus distribution (parentheses give number of languages for which this is the best performing model).	30
3.2	Quantitative assessment of each model against Dani focus distribution.	33
3.3	Quantitative assessment of each model against Berinmo focus distribution.	34
4.1	Training domains and example stimuli at various taxonomic levels for Experiment 2. .	46

Chapter 1

Introduction

1.1 General introduction

How does the mind make sense of the world? We receive noisy and ambiguous information from the environment, yet we develop abstract knowledge that generalizes over concepts and categories of “things” like *animals* and *color*. The ways in which people learn language and classify objects into groups of similar kinds require forms of complex reasoning and decision-making that seem to go far beyond the limited data available. These are problems of *induction*, where the evidence constrains, but does not determine, the solution to a problem. Understanding how cognition solves these problems has challenged cognitive scientists and philosophers of the mind since Plato.

Traditional approaches to address these questions assume that if the mind makes inferences beyond the available data, then to make up the difference either we have strong domain-specific learning constraints over structured, *innate* knowledge (Carey, 2000; Spelke & Newport, 1998), or we have strong domain-general associative learning mechanisms over simple, *emergent* structures of connectionist weights (McClelland et al., 2010; Rogers & McClelland, 2004). A recent proposal offers a compromise between the traditional positions with an alternative top-down explanation of *rational analysis* (Anderson, 1990; Marr, 1982): capturing human intelligence requires combining probabilistic inference over flexibly structured knowledge representations (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Framed in the context of this proposal, the present dissertation investigates how inference and representation interact to both guide and constrain how people reason with abstract knowledge. In particular, I present three case studies, each with different assumptions for the semantic representations and algorithms used to model cognition. This chapter provides a brief introduction of using probabilistic models of cognition as a method to study how people solve challenging inductive problems, and outlines the goals and different kinds of semantic representations the present dissertation will explore for each the the three case studies considered.

Probabilistic models of cognition

Viewing the mind as an information processing system that follows principles of computation has been a foundational assumption in cognitive science since the field began (Miller, 1956; Newell, Shaw, & Simon, 1958; Turing, 1950). Associating thought as a computational process provides a formal framework to analyze behavioral phenomena (Pylyshyn, 1984), and various formalisms have been proposed ranging from describing thought through innate rules and grammars (Chomsky, 1957; Grice, Cole, & Morgan, 1975) to general-purpose learning algorithms which update weighted connections in brain-like networks (McClelland et al., 2010; Rogers & McClelland, 2004). Probabilistic models of cognition have been used to explain the complex inferences that people make in their everyday lives (Griffiths & Tenenbaum, 2006), leveraging formalisms and data representations from recent advances in statistics, computer science, and artificial intelligence (Griffiths et al., 2010). These models address the types of solutions an ideal, or “rational” agent would extract from the available information in its environment (Oaksford & Chater, 1998). In contrast to traditional models in cognitive psychology that address *how* the mind leads to particular behaviors, rational models answer questions as to *why* people behave the way they do in light of limited evidence (Anderson, 1990; Marr, 1982). They provide a framework to evaluate the kinds of representations that people may use for solving problems of induction, and demonstrate how statistical evidence from the environment can be combined with prior knowledge (Tenenbaum et al., 2011).

Probabilistic models of cognition assume the mind represents its degree of belief in any particular explanatory hypothesis as a probability distribution, which gets updated as more data becomes available (Griffiths, Kemp, & Tenenbaum, 2008). Beliefs are updated by applying Bayes’ rule, which shows that the posterior probability of a hypothesis, h , given observed data of some phenomena, d , is proportional to the probability of observing d if h were the correct hypothesis (known as the likelihood) multiplied by the prior probability of that hypothesis:

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h' \in H} p(d|h')p(h')} \quad (1.1)$$

where the prior $p(h)$ is the *a priori* degree of belief that a particular hypothesis h is true before observing any data. It encodes the inductive biases of a learner and can act as a constraint when observing unlikely events.

One of the earliest successes under this framework addresses the question of how people can come to learn so much from so little. In particular, the Bayesian model of generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001a; Tenenbaum, Griffiths, & Kemp, 2006) has been successful in modelling human performance in learning from few examples, across numerous domains, e.g., word learning (F. Xu & Tenenbaum, 2007b), concept learning (Tenenbaum, 1999, 2000), sequential rules (Austerweil & Griffiths, 2011), and rule-based categorical concepts (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). It has been used to reconcile traditionally opposing views of how people reason with all-or-none rule-based generalizations or more gradient, similarity based generalizations (Tenenbaum, 2000). In addition, the Bayesian model of generalization has helped reveal previously unknown learning biases. For example, by manipulating assumptions

of the likelihood function, F. Xu and Tenenbaum (2007a) showed that children are sensitive to this sampling procedure, whether it was random or intentional.

Probabilistic models of cognition also address how assumptions of process and representation interact, and how they can support the variety of rich inferences that people make. Kemp and Tenenbaum (2009) showed how different structured representations are needed to account for inferences made across different domains, using a hierarchical Bayesian model over a set of graph-grammar primitives. For example, the voting patterns of U.S. Supreme Court judges can best be explained with a 1-dimensional “left” vs. “right” representation, while reasoning about biological concepts and learning new words from examples can best be explained with tree-structured representations (Kemp & Tenenbaum, 2008, 2009).

1.2 Goals of the present dissertation

The present dissertation works within the framework of probabilistic models of cognition to investigate how people reason with abstract knowledge. In particular, this research focuses on the kinds of knowledge representations that could support the inferences people make. Three case studies are presented which explore a range of inductive tasks and questions. In Chapter 2, I address the benefits in using probabilistic models of cognition to explore possible knowledge representations people use in searching semantic memory, and the constraints imposed from these representational commitments. In Chapter 3, representational constraints from perceptual color space combined with general principles of categorization are shown to account for both similarities and variation in color naming across the world’s languages. The work in Chapter 4 explores how a Bayesian model of generalization, previously used to address how people can learn the extension of words from only a few examples of a concept, can be extended to address similar challenges in machine learning. This additionally provides a framework for comprehensive evaluation of the Bayesian word learning model. In sum, the research presented in these case studies suggests probabilistic inference over structured representations illuminates how the mind could solve the challenges of induction, resulting in a better understanding of the formal principles characterizing human cognition, and machine learning models that behave more like people do. I briefly introduce each of the chapters below.

Chapter 2 investigates different strategies that people might use to search their memories for members of a particular category. Human memory has a vast capacity, storing all the semantic knowledge, facts, and experiences that people accrue over a lifetime. Given this huge repository of data, retrieving any one piece of information from memory is a challenging computational problem. In fact, it is the same problem faced by libraries (Anderson, 1990) and internet search engines (Griffiths, Steyvers, & Firl, 2007; Page, Brin, Motwani, & Winograd, 1999) that need to efficiently organize information to facilitate retrieval of those items most likely to be relevant to a query. In Chapter 2, I investigate proposals for the algorithms and representations used when people search their memory.

One of the main tasks that has been used to explore memory search is the semantic fluency task, in which people retrieve as many items belonging to a particular category (e.g., animals) as

they can in a limited time period. Both early and recent studies (Bousfield & Sedgewick, 1944; Romney, Brewer, & Batchelder, 1993; Thurstone, 1938; Troyer, Moscovitch, & Winocur, 1997) have consistently found that clusters appear in the sequences of words that people produce, with bursts of semantically related words produced together and noticeable pauses between these bursts. A recent article by Hills, Jones, and Todd (2012) argued that this pattern reflects a process similar to optimal strategies for foraging for food in patchy spatial environments, with an individual making a strategic decision to switch away from a cluster of related information as it becomes depleted. The work in this chapter demonstrates that similar behavioral phenomena also emerge from a simpler process (a random walk) on a richer structured representation (a semantic network derived from human word-association data). Semantic networks are a graph-based representation which provide a way to capture some of the graded and associative aspects of cognition, and are commonly used to explore questions about the structure of semantic memory (Baronchelli, Ferrer-i-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Collins & Loftus, 1975; Griffiths, Steyvers, & Firl, 2007; Steyvers, Shiffrin, & Nelson, 2004). I show that results resembling optimal foraging are produced by random walks when related items are close together in the semantic network. The findings Chapter 2 are reminiscent of arguments from the debate on mental imagery (Kosslyn & Pomerantz, 1977; Pylyshyn, 1973) as Anderson (1978) pointed out: claims with respect to representation cannot be evaluated on behavioral evidence alone without assuming a particular algorithm or process on that representation (due to mimicry).

Recent work has proposed using semantic networks to expose universals and variation in conceptual structure across a diverse set of the world's cultures and languages (Borin, Comrie, & Saxena, 2013; Youn et al., 2016). In Chapter 3, I explore a different approach to exploring semantic representations as universals of cognition, based on a geometric perceptual space rather than an associative network structure. I focus on a particular semantic domain under debate: color cognition. Do patterns of color naming across languages reveal universals of cognition, or culturally varying linguistic convention?

Focal colors, or best examples of color terms, have traditionally been viewed either as the underlying source of cross-language color naming universals, or as derived from category boundaries that vary widely across languages. Current findings present a mixed and empirically complex picture which partially supports and partially challenges each of these views. There are clear universal tendencies of color naming and focal colors across languages (Berlin & Kay, 1969; Kay & Regier, 2003; Lindsey & A. Brown, 2006, 2009), but at the same time there is also substantial cross-language variation (Davidoff, Davies, & Roberson, 1999; Kay & Regier, 2007; Regier, Kay, & Khetarpal, 2009; Roberson, Davidoff, Davies, & Shapiro, 2005).

The work in this chapter advances a position which synthesizes aspects of these two traditionally opposed positions, and accounts for existing data. In this view, color naming may be accounted for in terms of the overall shape of perceptual color space (Jameson & D'Andrade, 1997), which is irregularly shaped such that the most saturated yellow is more saturated than the most saturated blue. It is this irregular shape that constrains what a good color naming system is, which has been confirmed computationally as measured using general principles of categorization (Regier, Kay, & Khetarpal, 2007), and from pressures for informative communication (Regier, Kemp, & Kay, 2015). What this proposal leaves unexplained is the role of focal colors — which lie at the heart

of the debate.

In Chapter 3 I argue that focal colors are well-predicted from category extensions by a statistical model of how representative a sample is of a distribution, independently shown to account for patterns of human inference (Tenenbaum & Griffiths, 2001b). This model accounts both for universal tendencies and for variation in focal colors across languages. The results of this chapter suggest that categorization in the contested semantic domain of color may be governed by principles that apply more broadly in cognition, and that these principles clarify the interplay of universal and language-specific forces in color naming.

Problems of induction have been studied in both cognitive science and machine learning, but these fields approach them with different goals and different methods (Griffiths, 2015). In particular, cognitive scientists aim to build high quality models of human cognition, focusing on how people solve and fail at these problems. Furthermore, working within a laboratory setting provides precision at the cost of ecological validity, thus cognition experiments are typically small scale and use toy or artificial stimuli. On the other hand, researchers in machine learning aim to develop well-engineered solutions to these problems – they are not necessarily concerned with how people solve them. However, given the exponential growth in computing power, evaluating these solutions typically involves large-scale experiments with massive online databases as sources of stimuli. The work in Chapter 4 aims to bridge these two approaches, using resources from machine learning to extend and evaluate a Bayesian model of word learning.

Language learning is a classic problem solved better by the human mind than any computer. A four-year old child knows the meanings of thousands of words and can learn new words accurately from just a handful of labelled examples (Carey, 1978; Waxman & Markow, 1995). Given the infinite number of possible referents a word might have, this is a difficult problem of induction (Quine, 1975). Developing machine learning algorithms that approximate human performance on even one simple aspect of this problem – learning the meaning of a novel noun – is thus a significant challenge. Bayesian word learning models (F. Xu & Tenenbaum, 2007b) are a step towards answering this inductive challenge, using Bayesian inference over a tree-structured representation to identify the intended referent of a novel noun (e.g., does the word refer to Dalmatians, dogs, or all mammals?) in a similar manner to human word learning. However, to construct the hypothesis space of their Bayesian model, F. Xu and Tenenbaum (2007b) elicited approximately 400 similarity judgments from their participants. Clearly this is not practical to extend into every domain where people learn words.

The work in this chapter attempts to address this concern. I propose an automated method of constructing a hypothesis space and prior for the Bayesian word learning model using WordNet, a large online database that encodes the semantic relationships between words as a network (Miller, 1995). As a source of naturalistic stimuli, I explore another common online database, ImageNet, which provides at least 500 quality images per word in WordNet (Deng et al., 2009). This approach is first validated by replicating a previous word learning study (F. Xu & Tenenbaum, 2007b), and an additional experiment featuring three additional taxonomic domains (clothing, containers, and seats). In both experiments, I show that the same automatically constructed hypothesis space explains the complex pattern of generalization behavior, producing accurate predictions across each of the six different domains. Using the ImageNet Large Scale Vision Challenge dataset (Rus-

sakovsky et al., 2015), a third experiment was conducted spanning 4,000 concepts automatically generated from the structure of WordNet. This bridges the work with recent developments in Computer Vision and Machine Learning, providing new challenges and opportunities for developing machines that think more like people do.

Chapter 2

Process and representation in semantic memory search

2.1 Introduction

How do people search their memory for information related to a given cue? One classic method for exploring this question, the semantic fluency task, asks people to retrieve as many members of a category as possible in a limited amount of time (Bousfield & Sedgewick, 1944; Thurstone, 1938). This simple task has been used to explore the representations and processes that support semantic memory, and has even been used in clinical settings to study memory deficits in patients with different forms of dementia (Lezak, 1995; Tröster, Salmon, McCullough, & Butters, 1989; Troyer, Moscovitch, Winocur, Leach, & Freedman, 1998). Previous work has found that retrieval from semantic memory in fluency tasks tends to be produced in bursts of semantically related words with large pauses between bursts (Bousfield & Sedgewick, 1944; Romney et al., 1993; Troyer et al., 1997). For example, Troyer et al. (1997) asked participants to “name as many animals as you can” and observed that the retrieved animals tended to group into clusters (“pets”, “African animals”, etc.). The pauses between pairs of retrieved words in the same cluster (e.g., “dog-cat”) were very small when compared to the large pauses between pairs of retrieved words that do not belong to any of the same clusters (e.g., “cat-giraffe”). This pattern of patchy responses led Troyer et al. (1997) to posit that search through semantic memory is comprised of two processes, one process that jumps between clusters related to the given cue and another process that retrieves words within the current cluster.

Inspired by this pattern of bursts in retrieval from semantic memory, recent work by Hills et al. (2012) compared search through semantic memory to how animals forage for food. When animals search for food, they must consider the costs and benefits of further depleting their current food source as opposed to searching for a new patch of food. A large literature in biology called optimal foraging theory has compared animal foraging to ideal strategies (Stephens & Krebs, 1986). In particular, the *marginal value theorem* shows that an animal’s expected rate of food retrieval is optimal if they stop exploiting the current patch of food when the instantaneous rate (the marginal

value) of food being acquired from the current patch is lower than their overall expected rate of food retrieval (Charnov, 1976).

Human search through semantic memory could be considered analogous to how animals search for food, with semantically-related clusters playing the role of patches. If this were the case, then the pattern of pauses between pairs of retrieved words could be consistent with optimal foraging theory: responses should switch clusters when the marginal value of finding another item within the current cluster is less than the overall rate of return across memory. Hills et al. (2012) found that human memory search was consistent with this prediction of optimal foraging theory. Based on these results and a comparison of the performance of several different computational models, they proposed that human memory search involves two distinct processes, a “clustering” and a “switching” process, with the strategy for switching being consistent with the marginal value theorem.¹

In this chapter, we find that behavioral phenomena consistent with a two-stage search process can also be produced by a random walk on a semantic network derived from human word association data. This potentially provides an alternative account of human performance on semantic fluency tasks, and is consistent with previous work linking random walks on semantic networks with memory search (Griffiths, Steyvers, & Firl, 2007; Rhodes & Turvey, 2007; Thompson, Kello, & Montez, 2013). We show that predictions consistent with the results of Hills et al. (2012) are produced by random walks on semantic networks in which items that belong to the same cluster are close together in the network.

By providing an alternative account of the behavioral data that does not explicitly encode aspects of optimal foraging, our analyses suggest that further experiments will be required to determine whether the processes underlying human memory search involve optimal foraging. Furthermore, these results provide a concrete illustration of a theoretical problem for cognitive psychology that was identified by Anderson (1978) in the context of the mental imagery debate: different algorithms operating over different representations can produce the same predictions. In this case, a one-stage search process (a random walk) operating on one representation (a semantic network) can resemble a two-stage search process (optimal foraging) operating on another representation (a semantic space). The mimicry may not be complete – it might be possible to construct experiments that differentiate these two accounts – but both models produce key behavioral phenomena from the semantic memory literature.

The remainder of the chapter is organized as follows. First, we provide relevant background information on the retrieval phenomena predicted by an optimal foraging account of semantic fluency. We then discuss random walks as an alternative framework for modeling memory search, beginning with a model considered by Hills et al. (2012). This random walk provided a poor fit to human data and does not produce behavior consistent with optimal foraging. We then show that a random walk operating on a different representation – a semantic network based on free association data – does produce behavior consistent with optimal foraging. An analysis of the two representations on which these random walks are based suggests that the critical difference is

¹We note that Hills et al. (2012) are not the first to suggest that a switching process is involved in memory search – similar ideas appear in previous work (Dougherty, Harbison, & Davelaar, 2014; Raaijmakers & Shiffrin, 1981).

that the semantic network better captures the clustering of animals, and a minimal model confirms that a random walk based purely on such a cluster structure produces the key phenomena. We conclude by discussing the implications of our work for understanding the role of representations and algorithms in human foraging behavior and outlining possible directions for future research.

2.2 Optimal foraging as an account of semantic fluency

Optimal foraging theory covers a wide range of situations that a hungry animal might encounter (Stephens & Krebs, 1986), but the most basic scenario involves deciding how to navigate a “patchy” environment for resources. In this environment, food is contained in a set of discrete patches, which are depleted as the animal consumes the food. Staying in a patch thus provides diminishing returns, and the animal has to decide when to leave the patch and seek food elsewhere. The solution is provided by the marginal value theorem (Charnov, 1976), which indicates that the animal should leave the patch when the rate of return for staying drops below the average rate of return in the environment. Hills et al. (2012) suggested that retrieval from semantic memory is analogous to animals foraging for food, where a patch corresponds to a cluster of semantically-related items and acquiring food corresponds to retrieving an item from this cluster.

To investigate whether optimal foraging theory might account for human search through semantic memory, Hills et al. (2012) had people perform a semantic fluency task, where people were asked to “Name as many animals as you can in 3 minutes”. They then analyzed the search paths taken through memory, as indicated by the time between the animal names people produced, called the inter-item response time (IRT). These names were assigned to the predetermined animal categories identified by Troyer et al. (1997), which were used to analyze patterns in people’s responses: if an item shares a category with the item immediately before it, it is considered part of the same cluster, otherwise, that item defines a transition between clusters. For example, given the sequence “dog-cat-giraffe”, “dog” and “cat” are considered elements of the same cluster, while “giraffe” is considered the point of transition to a new cluster.

As a first measure of correspondence with optimal foraging theory, the ratio between IRTs and the long-term average IRTs for each participant were examined at different retrieval positions relative to a cluster switch. Figure 2.1a displays the results of this analysis. The first word in a cluster (indicated by an order of entry of “1”) takes longer to produce than the overall long-term average IRT (indicated by the dotted line), and the second word in a cluster (indicated by “2”) takes much less time to produce (reported results of a within-participant paired t -test, $t(140) = 13.14, p < 0.001$ and $t(140) = 11.92, p < 0.001$, for first and second words respectively). Furthermore, the IRTs for words preceding a cluster switch (indicated by “-1”) did not differ significantly from most participants’ own long-term average IRTs (reported results using a one-sample t -test, 132 of 141 participants were not significantly different, and the nine that were significantly different all had pre-switch IRT averages less than their long-term averages). These results are in line with the marginal value theorem, which predicts that IRTs should increase monotonically towards the long-term average IRT prior to a cluster switch, going above this average only when switching to a new cluster.

As a further test of the marginal value theorem’s predictions, the absolute difference between the pre-switch IRT and long-term average IRT was plotted against the number of words a participant produced (see Figure 2.1b). Participants with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words, as predicted by the marginal value theorem (reported results using a linear regression model found a significant negative relationship between participants’ deviation from the marginal value theorem policy for patch departures and the total number of words the participants produced, with a slope of -5.35 , $t(139) = -5.77$, $p < 0.001$).

2.3 Initial comparison of optimal foraging and random walks

Inspired by the marginal value theorem, Hills et al. (2012) suggested a two-part process model to account for the results of their experiment: when the IRT following a word exceeds the long-term average IRT, search switches from local to global cues. They compared this model to several simpler alternatives, including one in which memory search is simply construed as a random walk over a set of items (called the “one cue – static” model in their paper). Random walks have a long history as models of memory (Anderson, 1972), and recent work has shown that random walks on semantic networks can produce a distribution of IRTs in fluency tasks (Rhodes & Turvey, 2007;

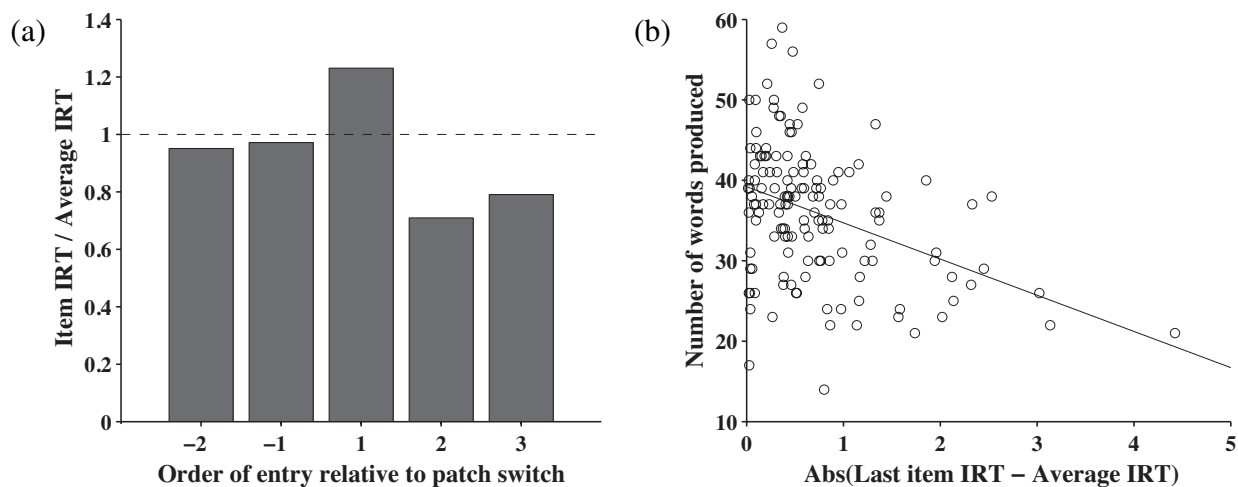


Figure 2.1: Experimental results of a semantic fluency experiment (free recall from the category of animals) reproduced from Hills et al. (2012). (a) The mean ratio between the inter-item response time (IRT) for an item and the participant’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the participant’s average IRT for the entire task. (b) The relationship between a participant’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a participant produced.

Thompson et al., 2013) and a pattern of responses in free association tasks (Griffiths, Steyvers, & Firl, 2007) similar to those produced by people.

The random walk considered by Hills et al. (2012) operated over a set of 771 animals, being possible responses in the semantic fluency task. The model assumed that each response people produced was sampled from a distribution based on the previous response, with the probability of each animal given by

$$P(X_i|X_j) \propto S(X_i, X_j)^\beta \quad (2.1)$$

where $S(X_i, X_j)$ is the similarity between the previous animal response X_j and the current animal response X_i , given by the BEAGLE model of semantic representation (Jones & Mewhort, 2007). In this model, each word is represented by a vector in a semantic space, and the similarity between words is based on the cosine similarity of their vectors. β is a free parameter of the model controlling the saliency (attention weight) assigned to a given cue.

Hills et al. (2012) compared this model with a two-part model that switched between exploring a cluster using a similar random walk and making a larger leap between clusters (called the “combined cue – dynamic” model in their paper). In this two-part model, the global switching process was carried out using a generic model of memory retrieval based on the ACT-R and SAM architectures (Anderson, 1990; Raaijmakers & Shiffrin, 1981). This makes it possible to calculate the probability of each participant’s sequence of responses under both models, and Hills et al. (2012) found that the two-part model gave a better fit to the human data than the random walk model.

Another way to evaluate the performance of the random walk model is to examine whether it can produce the key phenomena of human behavior that are suggestive of optimal foraging: the correspondence between the average IRT and the time at which people switch clusters, and the relationship between deviation from the marginal value theorem and overall performance (as shown in Figure 2.1). To examine this, we simulated random walks generating responses via Equation 2.1 and subjected these responses to the same analyses that Hills et al. (2012) used on their data. We used their reported mean $\beta = 4.34$ in the simulations below.

To connect the output of a simulation (the sequence of items visited by the random walk model) to the experimental results (i.e., IRTs), we need to define a method for mapping the sequence of items to IRTs. In our analyses, we consider only the time between first visits to animals, which we denote as $\tau(k)$ for the k^{th} unique animal item seen (out of the K unique animal items visited on the random walk). For example the output of a simulated random walk might be:

$$X_0 = \text{“dog”}, X_1 = \text{“cat”}, X_2 = \text{“dog”}, X_3 = \text{“mouse”}.$$

Here, $K = 3$ with $k = 1$ referring to “dog”, $k = 2$ referring to “cat”, and $k = 3$ referring to “mouse”. Our $\tau(k)$ function would return $\tau(1) = 1$, $\tau(2) = 2$, and $\tau(3) = 4$ for this example since we only consider the first time “dog” is visited (at timestep $n = 1$). Thus, we define the IRT between animals k and $k - 1$ in a sequence of nodes visited along a random walk as

$$IRT(k) = \tau(k) - \tau(k - 1). \quad (2.2)$$

where $\tau(k)$ is the first hitting time of animal $X_{\tau(k)}$. For the above example, the IRT between “mouse” ($k = 3$) and “cat” ($k = 2$) is

$$IRT(3) = \tau(3) - \tau(2) = 4 - 2 = 2.$$

With this mapping defined, we can perform the same set of analyses in Hills et al. (2012) on IRTs between animal words for our random walker simulations. Although Hills et al. (2012) consider only animals in their search space, we present this method to operate over multi-domain spaces more generally.

A total of 141 simulated random walks were run for 45 iterations, which was selected so that the average number of animals produced by a simulated random walk was approximately equal to the average number of animals typed by a participant in Hills et al. (2012). Figure 2.2 shows the results. Consistent with its poor fit to people’s responses, the random walk model did not produce behavior that resembles optimal foraging. While there was a negative linear relationship between the deviation from the marginal value theorem and overall performance (slope of -31.21, $t(138) = -4.00, p < 0.001$)², there are few differences between the IRTs and long-term average IRTs, regardless of retrieval position. This latter difference is particularly important when analyzing whether transitions between clusters occur at the point predicted by optimal foraging.

2.4 Exploring a different semantic representation

In arguing that people engage in a two-stage process based on optimal foraging theory, Hills et al. (2012) are making a commitment to a particular *algorithm* for memory search. In particular, they show that this algorithm accounts for human behavior better than a random walk. However, in making this comparison they also need to commit to a *representation* of semantic memory – in this case the spatial representation provided by BEAGLE (Jones & Mewhort, 2007). But, as Anderson (1978) pointed out, claims with respect to representation cannot be evaluated on behavioral evidence alone without assuming a particular algorithm or process on that representation (due to mimicry): with a different representation, a random walk might produce a closer match to human behavior.

While Hills et al. (2012) focused on spatial representations, other researchers in the memory literature have used random walks to capture aspects of human memory search (Griffiths, Steyvers, & Firl, 2007; Rhodes & Turvey, 2007; Thompson et al., 2013) by assuming a different kind of representation: a semantic network (Collins & Loftus, 1975). In a semantic network, nodes and edges in a graph encode words and pairwise associations, respectively. Technically, any random walk on a discrete set of objects can be interpreted as a random walk on a graph. From this perspective, the random walk considered by Hills et al. (2012) could be viewed as a random walk on a semantic network. However, the probabilities of moving between nodes on this graph were derived from the spatial representation used in BEAGLE, which places constraints on what kinds of conditional

²We removed outliers from all such analyses, which were defined as foragers whose deviation from the marginal value theorem or the number of words produced were more than three standard deviations from their respective means.

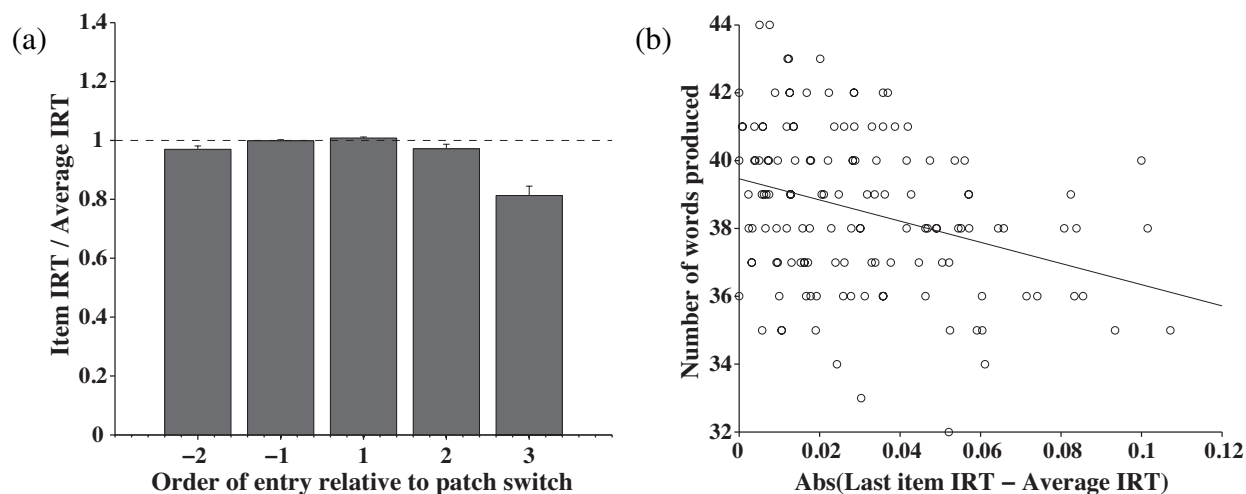


Figure 2.2: Results from 141 simulations of the random walk model from Hills et al. (2012), submitted to the same analyses as their human data. (a) The mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the participant’s average IRT for the entire task. (b) The relationship between a participant’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a participant produced.

probabilities (and graph structures) are possible. For a discussion of these constraints, please see (Griffiths, Steyvers, & Tenenbaum, 2007; Tversky, 1977). In particular, low-dimensional spatial representations constrain the number of items to which an item can be the nearest neighbor (Tversky & Hutchinson, 1986) – a property that might be relevant to the behavior of a random walk. Previous work has explored how semantic networks can be used to explore questions about the structure of semantic memory (Griffiths, Steyvers, & Firl, 2007; Griffiths, Steyvers, & Tenenbaum, 2007; Romney et al., 1993; Steyvers et al., 2004; Steyvers & Tenenbaum, 2005). Following this work, we approximate the structure of semantic memory with a semantic network constructed from people’s behavior in a word association task, where people are given a cue and list words associated with the cue (Nelson, McEvoy, & Schreiber, 2004). For example, if a participant were told the cue “bed,” they might respond with “pillow,” “blanket,” and “sheet.” The result is a semantic network with 5018 nodes, representing the associations between words from “a” to “zucchini.” There are 178 animals in the semantic network that were also one of the 373 animals produced by at least one of the 141 participants in Hills et al.’s (2012) experiment. Of these 178 animals, 13 were not the associate of any other word and so, we removed them from this and subsequent analyses (leaving 165 animals for analysis). Our random walk models operate over all 5018 nodes in the semantic network, however our $\tau(k)$ function operates over just these 165 animals.

A random walk on a semantic network searches memory in the following manner. Initially,

it starts at the node whose label corresponds to the cue. It moves to a new node by following an edge, selected at random, from the current node to the new node. A random walk on a semantic network could retrieve items in a patchy manner, appearing to make deliberate switches between clusters, if the clusters correspond to densely linked sets of nodes with few links between them. Thus, if the clusters that appear in people’s responses are reflected in the structure of the semantic network, this non-strategic search process might be sufficient to capture the phenomena reported by Hills et al. (2012).

Other than the network, there are two other steps to defining a random walk: (1) defining what node the random walk starts at (or a probability distribution over nodes), and (2) defining the probability distribution for transitioning from one node to the next node (a transition probability matrix). We assume that the random walk starts at the node that represents the cue C given to the participant. So, to capture the results of Hills et al. (2012), we assume that C is “animal”, and $X_0 = l^{-1}(C)$, where $l(\cdot)$ is a function whose input is a node and output is its corresponding label, and $l^{-1}(\cdot)$ is the inverse function, whose input is the label and its output is the node with that label. We explore four possible transition probability matrices defined by the orthogonal combination of two factors: whether the probability of transitioning out of a node is uniform or weighted over its edges and whether there is a non-zero probability of jumping back to the node corresponding to the cue.

The first factor is whether the probability of transitioning out of a node is *uniform* over its outgoing edges from the current node (as discussed in the previous example) or *weighted*, allowing the model to represent the degree of association between two items. In the case of the *weighted* model, the probability of transitioning from the current node to a new node is proportional to the frequency that the label of the new node was said by a participant given the current node as a cue in the word-association database (Nelson et al., 2004). Formally, the associations between a set of n items can be represented as a $n \times n$ matrix \mathbf{L} , where $L_{ij} = 1$ when item j is associated with item i and is zero otherwise. A random walk over these n items is defined by a matrix \mathbf{M} of transition probabilities, where M_{ij} denotes the conditional probability of jumping to item i given the random walk is currently at item j . In the *uniform* model, this is

$$M_{ij} = \frac{L_{ij}}{\sum_{k=1}^n L_{kj}}. \quad (2.3)$$

The denominator is called the *out-degree* of node j or the number of items that are associates of item j and so it is the number of possible items that the random walk could move to from node j . For the *weighted* model, L_{ij} is proportional to the number of times that i was an associate of j . The *weighted* model can capture that some associations (e.g., “dog” and “cat”) are stronger than others (e.g., “dog” and “house”).

The second factor is either *non-jumping*, where there is no effect of the cue besides for initializing the random walk or *jumping*, where at each time step, the model “jumps” back to the cue with probability ρ , but otherwise (with probability $1 - \rho$) the model transitions in the same manner as described above. We note this is a qualitatively different operation than the proposal of “jumping” between different search cues made by Hills et al. (2012). Rather than reflecting a strategic decision to switch between clusters, the jumps are executed at random and simply “prime” the search

process by returning to the initial state. In simulations not presented in the paper, we also examined the consequences of jumping to random nodes – a process which is more similar to the move from local to global cues in the model proposed by Hills et al. (2012), and has precedent in other work on random walks and semantic memory (Goñi et al., 2011, 2010). However, this jumping process produces qualitatively similar results to those described in the main text.

In sum, we will explore four different random walk models, which are formed by combining two factors: whether the edges are *uniform* or *weighted* and whether the random walk randomly jumps back to the cue (*jumping*) or not (*non-jumping*). Formally, they are all defined by the following equation

$$P(X_{n+1}|C = \text{“animal”}, X_n = x_n) = \rho P(X_{n+1}|X_n = l^{-1}(\text{“animal”})) + (1 - \rho)P(X_{n+1}|X_n = x_n) \quad (2.4)$$

where $P(X_{n+1}|X_n)$ is defined by Equation 2.3 with \mathbf{L} defined according to whether the model is *uniform* or *weighted*, and $\rho = 0$ when the model is *non-jumping* or $0 < \rho \leq 1$ when the model is *jumping*.

A direct quantitative comparison between these models and the models considered by Hills et al. (2012) is difficult, as there are different numbers of animals in the free association data and in the BEAGLE representation. This makes comparison hard because the probabilities assigned to participant responses by each model is determined in part by the number of possible responses (roughly speaking, the more animals in the model, the less probability assigned to each animal by the model). Instead, we perform qualitative comparisons of these models in the same manner as we did for the random walk earlier in the article.

A total of 141 simulated random walks were run for each of the four models. Each simulation was run for 2000 iterations, which was selected so that the average number of animals produced by a simulated random walk was approximately equal to the average number of animals typed by a participant in Hills et al. (2012). On average, each participant responded with 36.8 animals, and an average of 33.5, 42.5, 23.8, and 30.9 animals were produced by the uniform-non jumping, uniform jumping, weighted non-jumping, and weighted jumping random walk models, respectively. We expected the jumping models to produce more animals than the non-jumping models because “jumps” got the model away from nodes already visited by the random walk. Additionally, we expected slightly fewer animals to be produced by the random walk models than participants because of the small number of animals included in the semantic network (most people have probably encountered more than 165 animals). For the jumping models, we selected the probability of jumping on a given trial, ρ , to be 0.05. Other values for ρ produced similar results (assuming the value was small).

Figure 2.3 shows the results of analyzing the simulations of the four random walk models for optimal foraging-like behavior in the same manner as Hills et al. (2012) performed for participants in their experiment. In the left column of Figure 2.3 is the average ratio of the IRT of an item relative to its distance to the closest cluster switch (“order of entry”), and, the overall average IRT for each random walk model. Like people, the first item in a cluster (indicated by “1”) has a significantly longer IRT than the overall average IRT ($t(140) > 17, p < 0.001$ for all four models), and the second item in a cluster (indicated by “2”) has a significantly shorter IRT than the overall

average IRT ($t(140) < -15, p < 0.001$ for all four models). The introduction of jumps primarily reduces the difference for IRTs before (at “-2”) and after a cluster switch (at “2” and “3”), while increasing the amount of time it takes to find the first item in a cluster. This can be explained by the model randomly jumping back to the cue anywhere along the search path, making it difficult to find a new animal, yet once one is found, there are more unseen animals left to find nearby. In addition, the IRTs for words preceding a cluster switch (indicated by “-1”) were not significantly different from most walkers’ long-term average IRTs. The IRT for words preceding cluster switches (indicated by “-1”) of 140, 138, 139, and 138 out of 141 walkers were not significantly different for the uniform non-jumping, uniform jumping, weighted non-jumping, and weighted jumping models respectively, and all of the walkers that were significantly different had pre-switch IRT averages less than their long-term averages for each of the four models. This pattern of results is consistent with the results of participants in Hills et al. (2012)’s experiment, the marginal value theorem, and optimal foraging. Each time the IRT increases dramatically (at “1”) and then decreases dramatically (at “2”), one might be tempted to suggest that the model “found” another “patch” of relevant items in the semantic network. However, there are no search strategies being used by the model. It is simply walking randomly over the semantic network and emitting the labels of nodes that it visits. Thus, a simple process over a structured representation is sufficient to capture optimal foraging-like behavior.

The right column of Figure 2.3 examines the marginal value theorem’s cluster-switching policy, where the absolute difference between the pre-switch IRT and long-term average IRT was plotted against the number of words a random walker produced along with a regression line through this data (as in Figure 1b). Across all four models, walkers with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words (a linear regression model found a significant negative relationship between axes for each of the four models: slope = -0.19, $t(137) = 2.51, p < 0.05$, slope = -0.21, $t(137) = 1.98, p < 0.05$, slope = -0.10, $t(135) = 3.25, p < 0.05$, slope = -0.09, $t(137) = 2.15, p < 0.05$ for the uniform non-jumping, uniform jumping, weighted non-jumping, and weighted jumping models respectively). Intriguingly, each of the models produces the basic phenomena taken as evidence for the use of the marginal value theorem in memory search. These results show that behavior consistent with following the marginal value theorem can be produced by surprisingly simple search algorithms, at least when measured along these metrics. In the following sections, we turn to examining how the structure of semantic memory affects the behavior of these random walks.

2.5 The importance of clustering

Our results so far show that a random walk on a semantic network derived from free associations produces phenomena suggestive of optimal foraging, while a random walk on a spatial representation generated by BEAGLE (Jones & Mewhort, 2007) does not. This raises a natural question: Why? What is the critical difference between these two representations?

To address this question, we examined whether the similarity between items in these two representations reflects the clusters used by Troyer et al. (1997). According to the semantic network,

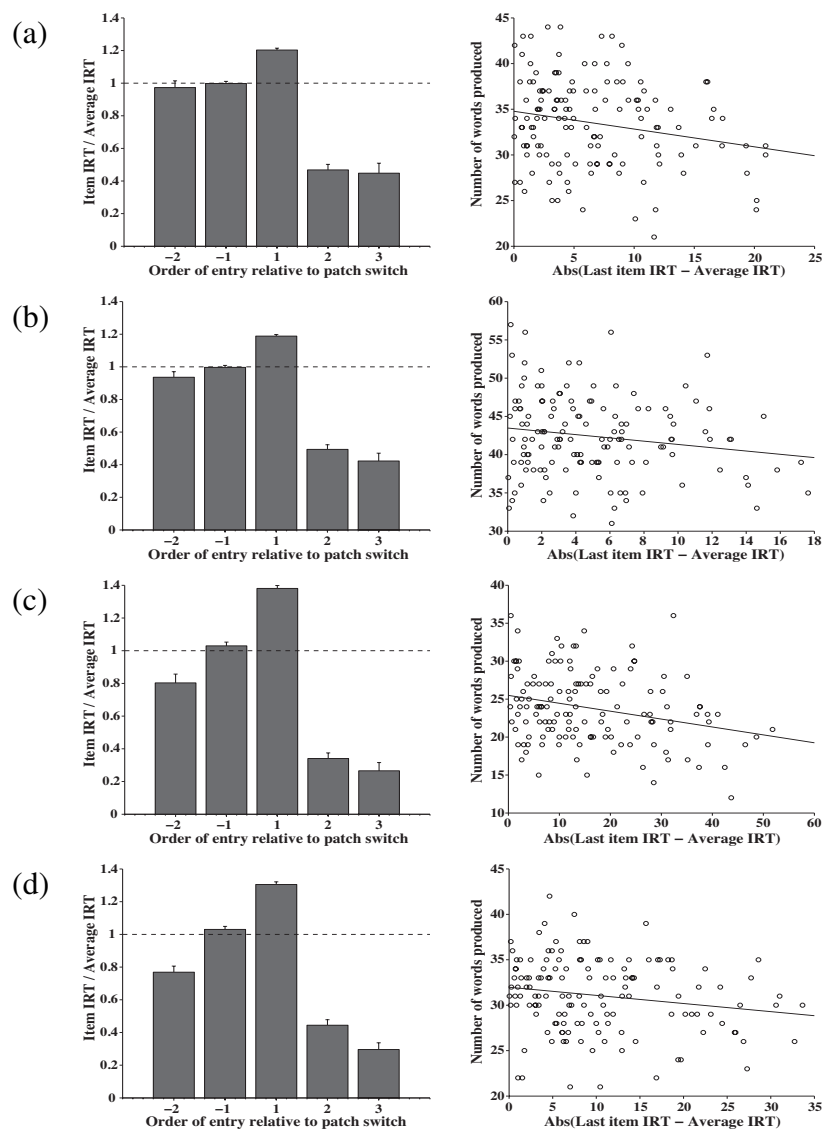


Figure 2.3: Results after 141 simulations for the four random walk models: (a) the uniform transition model with no jumps, (b) the weighted transition model with no jumps, (c) the uniform transition model with a jump probability of 0.05, and (d) the weighted transition model with a jump probability of 0.05. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced.

the similarity between the animals corresponding to nodes i and j was encoded as $s_{ij} = \exp\{-d_{ij}\}$, where d_{ij} is the shortest path distance between the nodes i and j in the semantic network. To derive similarities from the clusters, we used an additive clustering model (Shepard & Arabie, 1979), where the (nonexclusive) clusters from Troyer et al. (1997) were interpreted as features. To do so, we formed a 165×165 similarity matrix \mathbf{S} . According to additive clustering, the similarity matrix is defined as

$$\mathbf{S} = \mathbf{F}\mathbf{W}\mathbf{F}' \quad (2.5)$$

where \mathbf{F} is the matrix of clusters interpreted as features ($f_{ac} = 1$ when animal a is in cluster c), and \mathbf{W} is a diagonal weight matrix, whose elements are non-negative and represent the psychological weights of the clusters. We used the 22 animal clusters defined by Troyer et al. (1997) to define \mathbf{F} . We inferred \mathbf{W} by maximizing the posterior distribution of reconstructing \mathbf{S} based on graph distances using additive clustering, assuming a Gaussian prior on \mathbf{W} and Gaussian reconstruction error as outlined in Navarro and Griffiths (2008).

The top row of Figure 2.4 shows the graph-based similarity matrix and the similarity matrix reconstructed using additive clustering. Visual inspection of the block structure in both similarity matrices confirms that they are very similar and provides evidence that the semantic network implicitly encodes the clusters. The distance between the nodes corresponding to animals in the same cluster is smaller than the distance between animals in different clusters and the retrieval process depends (implicitly or explicitly) on this distance. This may be why a random walk on a semantic network can produce behavior that resembles optimal foraging.³

By comparison, the representation used in the random walk evaluated by Hills et al. (2012) does not show the same pattern of clustering. We used the same additive-clustering technique on the similarity data from BEAGLE, examining how well the similarity data could be predicted from the cluster membership of different animals. The bottom row of Figure 2.4 shows the results: there is only a weak signature of the animal clusters in these data. Consequently, the poor performance of this model could be a result of the underlying representation not encoding a clear cluster structure.

These results suggest that the critical difference between these two representations may be the extent to which they capture the cluster structure of animals. Because items that are in the same cluster are close in the semantic network, a random walk will tend to stay within clusters and occasionally switch between clusters, creating the illusion of a two-stage search process. To evaluate this idea, and to demonstrate that the performance of our model does not depend on any of the specifics of the free association data from which our semantic network was formed, we conducted a further simulation using a minimal random walk model. In this model, we assume that the probability of a transition from item j to item i is given by

$$L_{ij} = \begin{cases} 0 & i = j \\ (1-p)/C_j & i \text{ and } j \text{ are in the same cluster} \\ p/(n-C_j-1) & i \text{ and } j \text{ are not in the same cluster} \end{cases} \quad (2.6)$$

³We also examined various structural modifications of the network we operate upon in Appendix A, exploring how degree distributions, edge direction, and connectivity structure in the semantic network effect the observed optimal foraging phenomena.

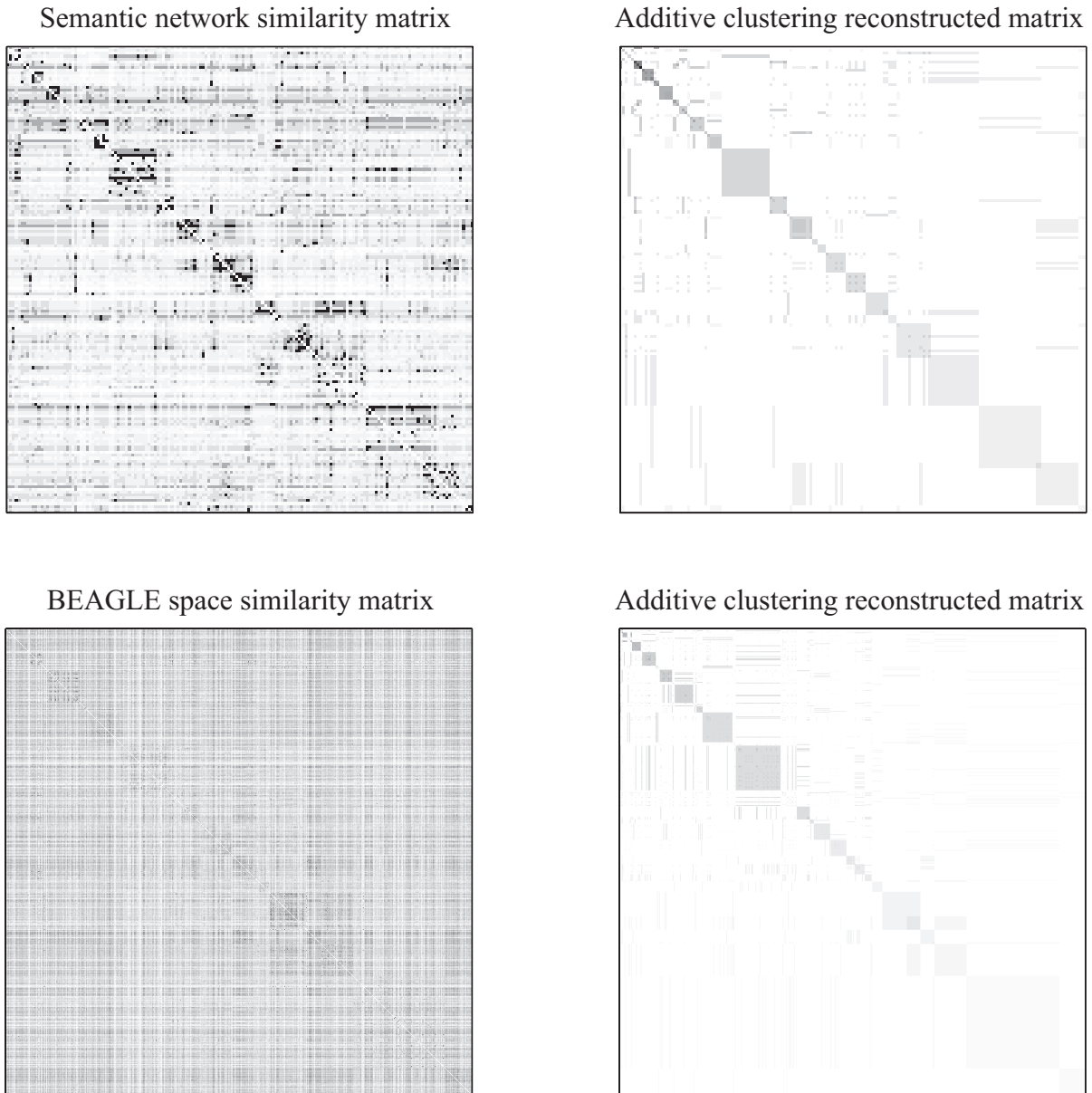


Figure 2.4: (Top row) A visualization of the similarity between pairs of animals in the semantic network (left panel) and an additive clustering model (right panel), where darker colors represent stronger similarities. (Bottom row) A visualization of the BEAGLE animal similarity space (left panel) and an additive clustering model (right panel). The rows and columns of the each matrix were reordered to display animals in the clusters with largest weight first.

where C_j is the number of items that belong to the same cluster as item j (excluding item j) and n is the total number of items. This model only makes use of the cluster structure, assigning a high probability to transitions within a cluster when p is small, but uses no other information about the items to determine the transition probabilities.

The random walk was run over the subset of 165 animals from the semantic network used in our previous simulation, and p was determined by calculating the average probability of making a transition outside a cluster in the *uniform non-jumping* random walk based on the word association network. We ran 141 simulations for a total of 45 steps each, and submitted them to the same analyses as Hills et al. (2012). The results are shown in Figure 2.5. The left column shows the key phenomena associated with optimal foraging, with the first word in a patch taking significantly longer to produce on average ($t(140) = 9.49, p < 0.001$), and the second word taking much less time to produce than the long-term mean ($t(140) = -11.11, p < 0.001$). The right column of Figure 2.5 examines consistency with the cluster-leaving policy indicated by the marginal value theorem, where again we find that walkers with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words (a linear regression model found a significant negative relationship between axes: slope = $-4.88, t(132) = 4.09, p < 0.001$).

The fact that this minimal model produces behavior similar to optimal foraging suggests that random walks can mimic a two-stage search process, provided they are on a representation that captures the underlying cluster structure. This suggests the success of the random walk model using the semantic network based on free associations in producing behavior that resembles optimal foraging, and the failure of the random walk using the BEAGLE representation considered by Hills et al. (2012) may be considered a consequence in the extent to which they capture this cluster structure.

2.6 Discussion

In this chapter, we examined two potential explanations for why people show optimal foraging-like behavior when they retrieve items from semantic memory. Both explanations produced behavior consistent with the predictions of optimal foraging, but they propose that very different representations and processes are responsible for this behavior. Hills et al. (2012) suggested that semantic memory is based on spatial representations and search is a dynamic process, retrieving items from one cluster at a time and switching between clusters when the retrieval rate falls below a threshold. We proposed an alternative explanation, where semantic memory is represented by a network and search is simply a random walk on the network. In support of this proposal, we showed that predictions consistent with the results of Hills et al. (2012) are produced by a random walk model on a network where semantically-related items are close together in the network.

Representations and algorithms

Taken together, our simulations show that the behavior in semantic fluency tasks that Hills et al. (2012) viewed as evidence for optimal foraging is also predicted by a random walk on a semantic

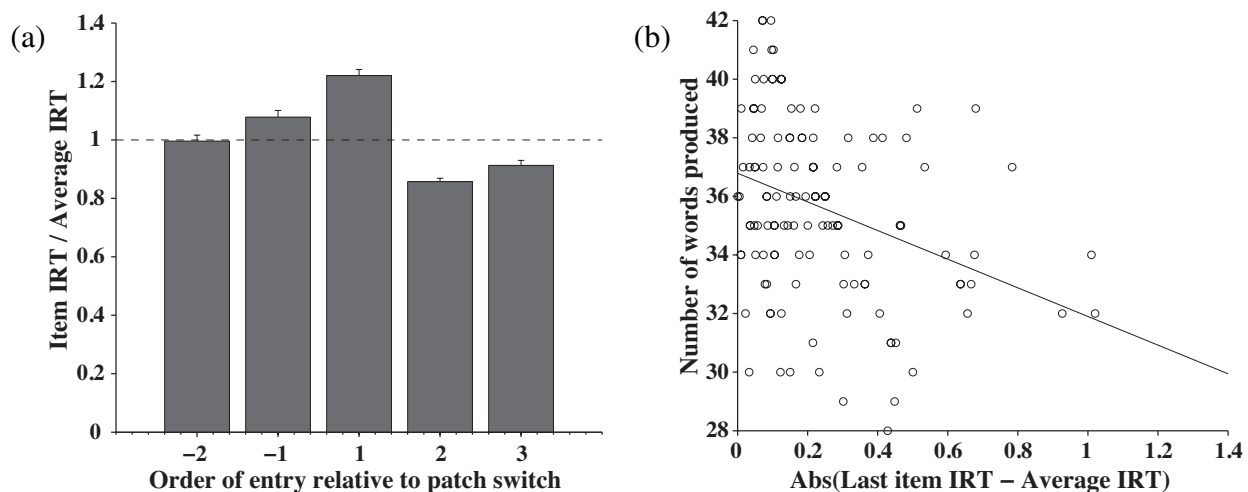


Figure 2.5: Results for the minimal model on our semantic network with p estimated from the uniform transition word association matrix. (a) The mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a patch and the last word in the preceding patch). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. (b) The relationship between a walker’s deviation from the marginal value theorem policy for patch departures (horizontal-axis) and the total number of words a walker produced.

network. Crucially, this behavior depends on the representation used by the random walk model: a random walk on a semantic network produces optimal foraging behavior, but a random walk on corresponding spatial representations does not. Consequently, it seems that there is something special about the semantic network representation that allows the simple random walk to appear similar to optimal foraging.

Finding that different algorithms operating on different representations can produce the same behavior might seem surprising, but has a precedent in cognitive psychology. The mental imagery debate (Kosslyn, 1994; Pylyshyn, 1973) depended crucially on this issue – whether there are effective ways of identifying the algorithms and representations that human minds employ. Anderson (1978) convincingly argued that we should not be surprised to find cases where algorithms and representations that seem quite different nonetheless end up producing similar behavior – such cases are the rule, rather than the exception. In fact, for a sufficiently rich set of algorithms and representations, we can always find algorithm-representation pairs that cannot be discriminated based purely on behavior.

The situation illustrated by our analyses is not necessarily as extreme as the cases that Anderson considered, but it does illustrate one of the fundamental challenges of cognitive psychology: possible psychological representations and mechanisms are always underdetermined by the available behavioral data, and even behavior that seems like the signature of one mechanism can sometimes

be produced by others. In this case, further experiments may be able to discriminate between optimal foraging and random walks, but these experiments must be specifically designed to distinguish between these two accounts rather than motivated by the predictions of one account alone.

Conclusion

Identifying and retrieving information relevant to a cue is one of the basic capabilities of the human memory system. Understanding how people solve the task of searching this vast store of information is likely to give us insight not just into the human mind, but into how to build better artificial information retrieval systems. Optimal foraging and random walks on semantic networks offer two quite different accounts of this process – one based on an intelligent search strategy, the other on a rich representational framework. Both algorithms and representations also have links to other disciplines, offering links to literatures in biology and computer science respectively. That both accounts can produce similar behavior is surprising, but also exciting, in that it creates new opportunities to explore these connections more deeply and develop a more complete picture of this remarkable human capacity.

Chapter 3

Universals and variation in color categories

3.1 Introduction

Focal colors, or best examples of color terms, lie at the center of the debate over language and color cognition. An influential view (Kay & McDaniell, 1978) is that color naming across languages is constrained by the Hering primaries (Hering, 1964) in the opponent pairs *red* vs. *green*, *yellow* vs. *blue*, and *black* vs. *white*. The best examples of these six color terms are often understood to be universal privileged points, or foci, in color space, such that languages differ in their color naming systems primarily by grouping these universal foci into categories in different ways. There is some empirical support for this view: the best examples of color terms across languages tend to cluster near these six points (Berlin & Kay, 1969; Regier, Kay, & Cook, 2005), and an early study (Heider, 1972b)—but not a recent followup (Roberson, Davies, & Davidoff, 2000)—also found these colors to be cognitively privileged.

However, Roberson and colleagues (Roberson et al., 2000) claimed that this influential view has matters exactly backwards. They argued that color categories are not constrained by universal foci, but are instead defined at their boundaries by local linguistic convention, which varies across languages. They proposed: “Once a category has been delineated at the boundaries, exposure to exemplars may lead to the abstraction of a central tendency so that observers behave as if their categories have prototypes” (p. 395). On this view, best examples do not reflect a universal cognitive or perceptual substrate, but are merely an after-effect of category construction by language.

A proposal by Jameson and D’Andrade (Jameson & D’Andrade, 1997) has the potential to reconcile these two opposed stances. They suggested that there are genuine universals of color naming, but that these do not stem from a small set of focal colors. Instead, on their view, universals of color naming stem from irregularities in the overall shape of perceptual color space, which is partitioned into categories by language in a near-optimally informative way. This proposal (see also (Jameson, 2005a, 2005b, 2010; Komarova, Jameson, & Narens, 2007)) has been shown to explain universal tendencies and cross-language variation in the *extensions* of color categories (Regier et al., 2007, 2015) (see also (Baronchelli, Gong, Puglisi, & Loreto, 2010, 2015; Dowman, 2007; Griffin, 2006; Lindsey & A. Brown, 2009; Puglisi, Baronchelli, & Loreto, 2008; Steels &

Belpaeme, 2005; Yendrikhovskij, 2001) for other approaches to the same question). However it is the *best examples* of color categories, not their extensions, which lie at the heart of the debate.

Here, we address this open issue, completing the reconciliation of the two standardly opposed views. Following Roberson et al. (Roberson et al., 2000), we argue that best examples of color categories across languages are not reflections of underlying universal focal colors. However we argue that best examples do not vary arbitrarily either. Instead, we note that color categories across languages reflect the functional need for informative communication about color (Jameson & D’Andrade, 1997; Regier et al., 2007, 2015), and argue that best examples are derived from the resulting informative categories. On this view, all languages are driven by the same functional forces, and thus unrelated languages will often independently settle on similar informative color naming systems—and when they do, the best examples of those color categories should also be similar. But color categories may also vary across languages, representing different informative partitions of color space—and when categories do vary, the best examples of those categories should vary with them. Here we test this account by asking whether best examples of categories across languages can be predicted from category extensions, and whether such predictions account both for universal tendencies and for cross-language variation in focal colors.

Pursuing these ideas requires an account of how the best example of a category is determined. To this end, we use a rational model which formally characterizes the best example of a category in terms of the support that it provides to a Bayesian inference. This model was originally proposed (Tenenbaum & Griffiths, 2001b) to account for patterns of human inference that have been taken to suggest a cognitive heuristic of “representativeness” (Kahneman & Tversky, 1972), as described below. To preview our results, we find that this model accounts both for universal tendencies and for variation in focal color choices across languages.

The remainder of the chapter proceeds as follows. We first describe the data we consider, and a set of competing models, including the representativeness model, that predict best examples of color categories from the extensions of those categories. We test these models against universal tendencies in the data, and find that the representativeness model outperforms competing models, consistent with preliminary results using a different measure of model performance (Abbott, Regier, & Griffiths, 2012). In a separate test, we then consider cross-language variation in the same data, and again find that the representativeness model outperforms its competitors. We close by discussing the implications of our findings.

3.2 Predicting best examples of color categories

Evaluating formal models of color foci requires a good source of color naming data. The primary data we considered were those of the World Color Survey (WCS)¹, which collected color naming data from native speakers of 110 unwritten languages worldwide (Cook, Kay, & Regier, 2005).

¹The WCS color naming data we analyze are available at <http://www.icsi.berkeley.edu/wcs/data.html>. Because our analyses concern the relation between category extension (naming data) and best examples (focus data) on a per-speaker basis, we considered only those categories for which both naming and focus data were available for the same speaker.

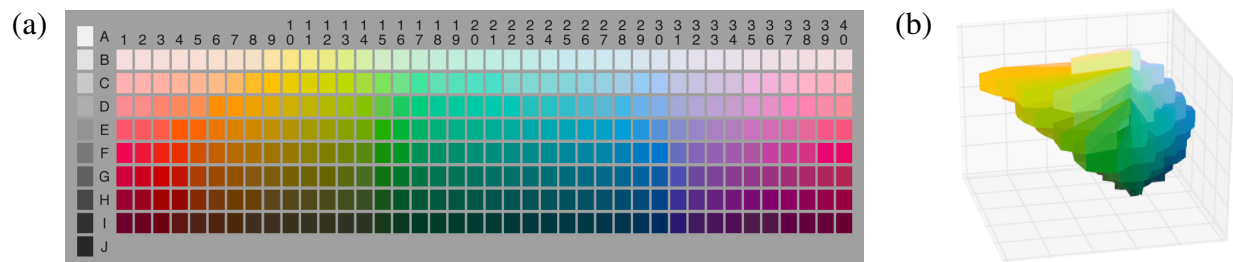


Figure 3.1: (a) Color naming stimulus array. The rows correspond to 10 levels of Munsell value (lightness), and the columns correspond to 40 equally spaced Munsell hues. The color in each cell corresponds approximately to the maximum available Munsell chroma for that hue-value combination. (b) The chips of the stimulus array as plotted in CIELAB color space. The irregularity of the outer surface of the color solid can be seen, most notably in the yellow region.

Participants in the WCS were shown each of the 330 color chips of the stimulus array in Figure 3.1(a), and were asked to name each chip with a color term in their native language; we refer to the resulting data as “naming data”. Afterwards, participants were asked to pick out those cells in the stimulus array that were the best examples (foci) of each color term they used; we refer to these as “focus data”.

We compared several models that predict best examples of color categories from the extensions of those categories. We represented each color in the stimulus array as a point in 3-dimensional CIELAB color space (Kay & Regier, 2003) (see Figure 3.1(b)). For short distances at least, Euclidean distance between two colors in CIELAB is roughly proportional to the perceptual dissimilarity of those colors (Brainard, 2003) (but see also (Komarova & Jameson, 2013)). For each named color category used by each speaker in each language of the WCS, we used each model to predict that speaker’s focus data from that speaker’s naming data. We provide overviews of our models and analyses below.

Representativeness model

Why do people believe that the sequence of coin flips HHTHT (where H=heads, T=tails) is more likely than the sequence HHHHH to be produced by a fair coin? Using simple probability theory, it is easy to show that the two sequences are in fact equally likely. Cognitive psychologists have proposed that people use a heuristic of “representativeness” instead of performing probabilistic computations in such scenarios (Kahneman & Tversky, 1972). We might then explain why people believe HHTHT is more likely than HHHHH to be produced by a fair coin by arguing that the former is more representative of the output produced by a fair coin than the latter. But how do we define the notion of representativeness that the heuristic appeals to? Numerous proposals have been made, connecting representativeness to existing quantities such as similarity (Kahneman & Tversky, 1972) and likelihood (Gigerenzer & Hoffrage, 1995). Tenenbaum and Griffiths (2001b) provided a *rational analysis* (Anderson, 1990) of representativeness by trying to identify the problem that such a quantity solves. They noted that one sense of representativeness is being a good

example of a concept, and they showed how this could be quantified in the context of Bayesian inference.

Formally, given some observed data d and a set of hypothetical sources, \mathcal{H} , we assume that a learner uses Bayesian inference to infer which $h \in \mathcal{H}$ generated d . In that context, Tenenbaum and Griffiths (2001b) defined the representativeness of data d for hypothesis h to be the evidence that d provides in favor of a specific h relative to its alternatives:

$$R(d, h) = \log \frac{p(d|h)}{\sum_{h' \neq h} p(d|h')p(h')} \quad (3.1)$$

where $p(h')$ in the denominator is the prior distribution on hypotheses, re-normalized over $h' \neq h$. This measure was shown to outperform similarity and likelihood in predicting human representativeness judgments for a number of simple stimuli. We propose that this measure can also be used to predict focal colors, or best examples of named color categories, from the extensions of those categories.

We first need a way to represent a speaker's naming data as a probability distribution. We do so as follows. For each named color category used by each speaker in each language of the WCS, we modeled that category as a 3-dimensional Gaussian distribution in CIELAB space, and estimated the parameters of that distribution using a normal-inverse-Wishart prior, a standard estimation method for multivariate Gaussian distributions of unknown mean and unknown variance (Gelman, Carlin, Stern, & Rubin, 2004). Specifically, given a set of M chips \mathbf{x}_i in color category t , where \mathbf{x}_i holds the coordinates of that chip in CIELAB space, we obtain the estimates:

$$\mu_t = \frac{1}{M} \sum_i^M \mathbf{x}_i, \quad \Sigma_t = \frac{SS_t + \lambda_0}{M + \nu_0} \quad (3.2)$$

where SS_t is the sum of squares for category t : $\sum_i^M (\mathbf{x}_i - \mu_t)(\mathbf{x}_i - \mu_t)^\top$, and λ_0 and ν_0 are the parameters of the prior. λ_0 was set by taking an empirical estimate of the variance in CIELAB coordinates over all chips in the stimulus array, and ν_0 was set to 1.

With a Gaussian distribution that characterizes the category named by color term t , we can adopt the representativeness measure given in Equation 3.1 to determine how good an example each color chip x is of color term t . Substituting x in for the observed data d and t for hypothesis h we obtain the expression:

$$R(x, t) = \log \frac{p(x|t)}{\sum_{t' \neq t} p(x|t')p(t')} \quad (3.3)$$

where $p(x|t)$ is given by the density function of the Gaussian described above and the priors $p(t')$ are proportional to the number of chips in named color category t' . This model can be seen as formalizing the claim of Rosch and Mervis (1975) that category prototypes, or best examples, reflect not just high similarity to other members of the category (captured here in the numerator), but also low similarity to members of other categories (captured in the denominator).

We test this measure against the alternative proposals mentioned above (Gigerenzer & Hoffrage, 1995; Kahneman & Tversky, 1972): a likelihood model and two similarity models: a prototype model and an exemplar model. In addition, we explore a model that selects as the focus

for category t that chip in the extension of t that has the highest chroma. Chroma, or saturation, corresponds to how colorful or “un-gray” a given color is, and in exploring this model we follow the suggestion (Jameson & D’Andrade, 1997; Regier et al., 2007) that focal colors tend to be those with high chroma (but see also (Witzel & Franklin, 2014)). These models represent different ways in which the best example of a category may be predicted from the extension of that category. As with the representativeness model, for a given color x and color term t , each model assigns a score indicating how good x is as an example of t .

Likelihood model

In this model, the goodness score of color x as an example of color category t is given by the density function of the Gaussian distribution that was fit to the naming data for t . Thus:

$$L(x, t) = \log p(x|t) \quad (3.4)$$

This model is similar to the representativeness model, but lacks the denominator which captures competition among categories in that model.

Prototype model

In this model we define the focus, or prototype, of color category t to be the mean μ_t of the distribution characterizing t (Reed, 1972). The score for this measure then becomes the similarity of x to that prototype:

$$P(x, t) = \text{Sim}(x, \mu_t) \quad (3.5)$$

where $\text{Sim}(\cdot, \cdot)$ characterizes the similarity between two colors as a function of CIELAB distance:

$$\text{Sim}(x, y) = \exp\{-c \text{dist}(x, y)^2\} \quad (3.6)$$

and $c = 0.001$, following previous work (Regier et al., 2007).

Exemplar model

The exemplar model uses a scoring metric similar to that of the prototype model, except rather than computing the similarity of color x to a single prototype, we compute its similarity to each color chip that falls in the extension of category t (Nosofsky, 1988), and sum the results:

$$E(x, t) = \sum_{x_j \in \mathbb{X}_t} \text{Sim}(x, x_j) \quad (3.7)$$

where \mathbb{X}_t is the set of color chips that fall in the extension of category t .

Chroma model

The score for this model is given by the similarity of color x to that color chip c_t which has the highest chroma (saturation) value within the extension of category t . Thus we compute:

$$C(x,t) = \text{Sim}(x, c_t) \quad (3.8)$$

where c_t is the chip within the extension of t that has the highest chroma value. In the case of ties for c_t —that is, several chips with the same maximum value for chroma—we randomly select a chip from the set of ties.

3.3 Results

We assessed these models as follows. For each speaker of each language in the WCS, we first considered that speaker's naming data, and modeled the categories in those data as either a set of Gaussian probability distributions (for the Representativeness and Likelihood models), or as a set of 3-dimensional points in CIELAB (for the Prototype, Exemplar, and Chroma models). Then, for each such category, we determined how good an example of that category each of the 330 chips in the stimulus array is, according to each model. This yielded, for each model, a ranking of chips in the array by predicted goodness, and we compared this model prediction with empirical WCS focus data, that specify which chip(s) were in fact selected by that speaker as the best example(s) of that category. Thus, we compare model predictions to empirical data on a per-speaker basis. Below we present both qualitative and quantitative evaluations of the models.

Distribution of foci

A simple means of assessing the models is to generate predicted focal choices from each model's ranking of chips, and to then compare the distribution of those predicted focal choices with the distribution of actual focus data from the WCS. Some speakers in the WCS provided more than one focus (best example) for some categories; if a speaker provided n foci for a given category, we selected the n top-ranked chips as a given model's predicted focal choices for that category and speaker. In this manner we obtained, for each model, one predicted focal choice for each empirical focal choice in the data. We then counted the number of times each of the 330 color chips in the stimulus array was selected as a focal choice, yielding a distribution of focal choices over the stimulus array. Interestingly, every chip in the stimulus array was selected at least once as a focus for some color term by some speaker of some language. We compared the empirical distribution of foci across the array with the distribution predicted by each of the models. Following an earlier analysis of WCS focus data (Regier et al., 2005), we plotted these distributions over the chromatic portion of the array, where the 2-dimensional layout makes contours easily interpretable. The resulting contour plots, of the empirical WCS focus distribution and the five models' predicted focus distributions, are shown in Figure 3.2.

The empirical distribution is shown in panel (a), and replicates earlier findings (Regier et al., 2005). The distribution predicted by the representativeness model (panel b) matches this empirical

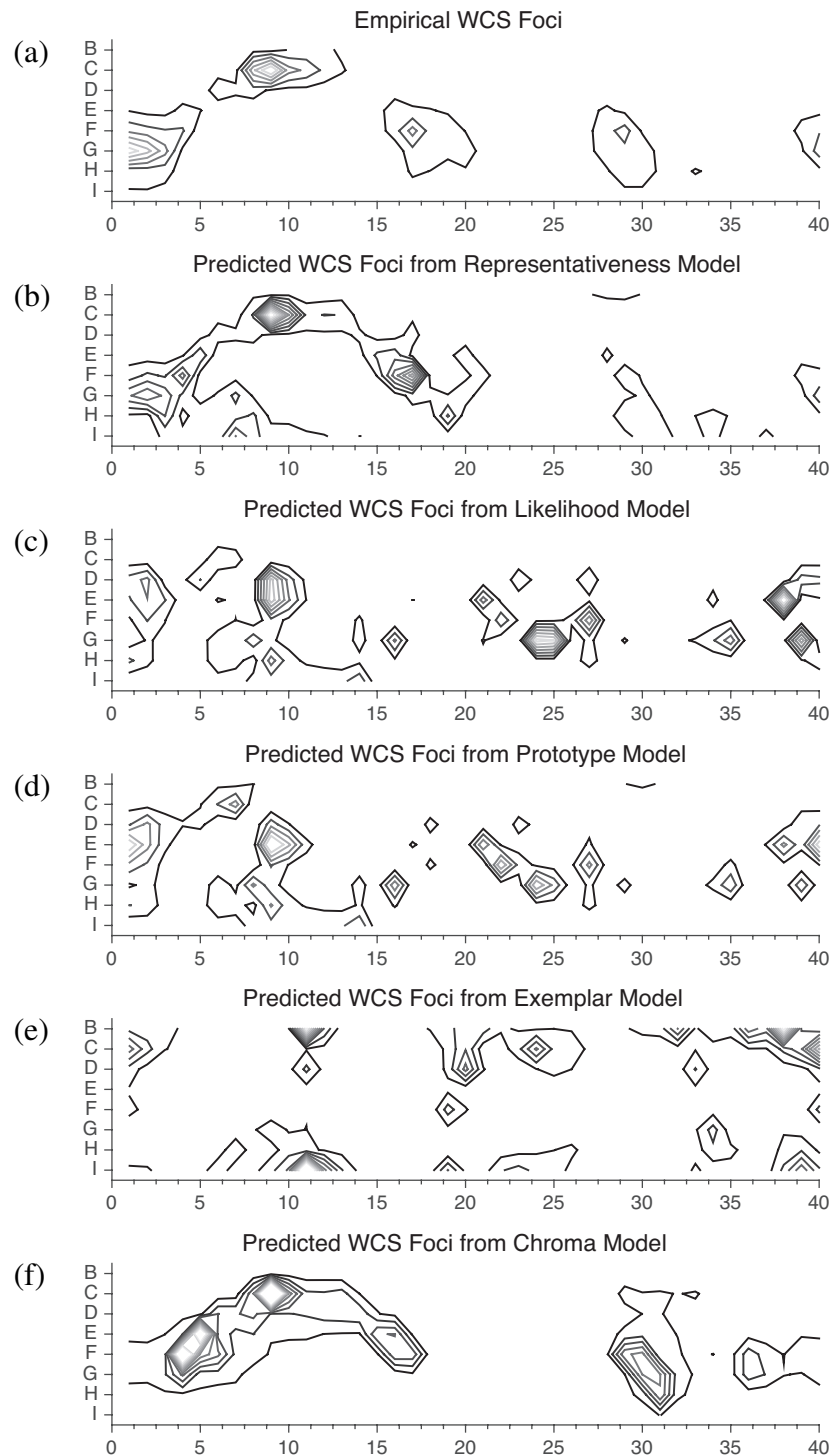


Figure 3.2: Contour plots of the focus distributions in (a) the WCS, and as predicted by (b) the representativeness model, (c) the likelihood model, (d) the prototype model, (e) the exemplar model, and (f) the chroma model. Each contour line corresponds to 100 focal choices.

Table 3.1: Quantitative assessment of each model against WCS focus distribution (parentheses give number of languages for which this is the best performing model).

Model	QF distance	Rank Position
Representativeness	1.17 (76)	27.67 (104)
Likelihood	1.74 (6)	42.65 (1)
Prototype	1.96 (3)	48.30 (0)
Exemplar	1.64 (24)	38.98 (5)
Chroma	2.13 (1)	78.51 (1)

distribution qualitatively fairly well. Moreover, at least on informal inspection, the representativeness model appears to approximate the empirical distribution more closely than do the competing models. The chroma model (panel f) at first appears to also approximate the empirical distribution fairly well, but closer inspection reveals that several of the peaks of the model distribution do not align correctly with those of the empirical distribution (see also (Witzel & Franklin, 2014)).

This qualitative assessment is reinforced by a quantitative one that considered all chips of the array, not just the chromatic portion. The quadratic form (QF) distance is a measure of the difference between two histograms, H_1 and H_2 , over the same set of points in space, and it takes into account the similarities between those points (Hafner, Sawhney, Equitz, Flickner, & Niblack, 1995). QF distance is defined as:

$$QF_M(H_1, H_2) = \sqrt{(H_1 - H_2)^\top \mathbf{M} (H_1 - H_2)} \quad (3.9)$$

where \mathbf{M} is an inter-point similarity matrix. In our analyses we defined \mathbf{M} over the color chips of the stimulus array with $m_{i,j} = \text{Sim}(i, j)$, where $\text{Sim}(\cdot, \cdot)$ characterizes the similarity between two colors as a function of CIELAB distance:

$$\text{Sim}(x, y) = \exp\{-c \text{dist}(x, y)^2\} \quad (3.10)$$

and $c = 0.001$, following previous work (Regier et al., 2007).

We computed the QF distance between the WCS empirical focus distribution shown in Figure 3.2(a) and each of the model distributions shown in Figure 3.2(b) through 3.2(f), with similarity determined by the function $\text{Sim}(\cdot, \cdot)$ defined above. The results are shown in Table 3.1, with the best model score shown in **bold**. The representativeness model outperforms the other models, diverging less from the empirical distribution than its competitors do.

Each model produces as output a ranking of the stimulus chips, where rank is assigned in descending order. Thus, another natural way to assess the models is to note the position of the true empirical focal choice in this ranked list. The average rank position for each model is presented in Table 3.1. As before, we find that the representativeness model outperforms the other models, ranking the true foci higher on average.

3.4 Language level analysis

The analyses above considered all focus choices in the WCS as a single distribution, pooling together choices made by different speakers of different languages. Color naming varies across languages, so a natural question is whether the representativeness model also outperforms its competitors when each language is considered separately. Such a language-level analysis would be appropriate if we are to take seriously the hypothesis that category boundaries are determined in part by local linguistic convention (Roberson et al., 2000).

We considered separately each of the 110 languages of the WCS, and conducted analyses like those described above, but at the language level, pooling together focus choices that were made by speakers of a single language. For each language, we noted which model best predicted focus choices by speakers of that language, by each of our two metrics. Table 3.1 (in parentheses) shows that by both metrics, the representativeness model again outperforms its competitors: it exhibits the best performance for a majority of the WCS languages. Four paired t-tests compared average QF distance per language predicted by the representativeness model with that predicted by each other model, in each case averaging across speakers and color terms for each language. The representativeness model outperformed each other model, $p \ll 0.001$, Bonferroni-corrected for multiple comparisons. Analogous results were also obtained when measuring rank position rather than QF distance. Full details of this cross-language analysis, including results for individual languages, are presented in Appendix B.

We also conducted similar analyses for two languages outside the WCS: Berinmo (Roberson et al., 2000) and Dani (Heider, 1972a), which we present below. These analyses highlight both cross-language and within-language (i.e. inter-individual) variation in focus choices.

Dani

We considered Dani color naming data as reported by Heider (1972a). Dani has been reported to use primarily a two-term color system, *mili* and *mola*, corresponding roughly to “cool” and “warm” colors, respectively, although Heider also found that roughly half the Dani participants also provided other terms for regions corresponding roughly to English *red*, *yellow*, and *blue*. Dani, with only two major color terms, has fewer major color terms than any of the languages of the WCS. Dani thus provides an opportunity to test our models against a system that is qualitatively different in an important respect from those of the WCS.

Our models require two sorts of data relative to the same set of stimuli: naming data and best example data. The experimental stimuli and procedure used by Heider (1972a) differed slightly from those used in the WCS and presented in the main text. A reduced set of 160 maximally saturated Munsell color chips were used, corresponding to every other column in the WCS chromatic grid, and data were not provided at the level of individual speakers. Instead, naming data for Dani were reported in the form of language-level responses, aggregated over speakers, distributed over the 160 chip chromatic grid. Such data were provided for all terms except for *mola*, which was described as the complement of *mili*. Focus data were also provided at the language level,

i.e. again aggregated over speakers, in the form of a histogram of reported best examples for each term. These focus histograms were provided for *mili* and *mola*, but not for the less dominant terms mentioned above, consistent with the view of Dani as primarily a two-term language.

Preparation of the data. Since we do not have individual speaker data, we constructed a single naming map from the reported naming distributions to provide as input to our models. Naming data were not provided for *mola*, so we inferred data for that term by assuming that all 40 Dani participants provided a naming response for each chip in the stimulus array and that *mola* was the only missing term after summing the counts for other color terms at a given stimulus chip. This allowed us to create a mode map for Dani over the 160 chip chromatic grid, where each chip is assigned the color term used by a plurality of speakers (the modal term for that chip). Figure 3.3 below displays the resulting Dani mode map. Here, the extension of each color term is shown as a colored region, and the color assigned to that region is determined by taking the average RGB coordinates of the chips in the region. The number of focus hits for *mili* and *mola* per chip, aggregated over speakers in the language, is overlaid on top. Although Dani is generally regarded as a two-term color system, two color chips in the mode map were given names other than *mili* and *mola* by a plurality of speakers.

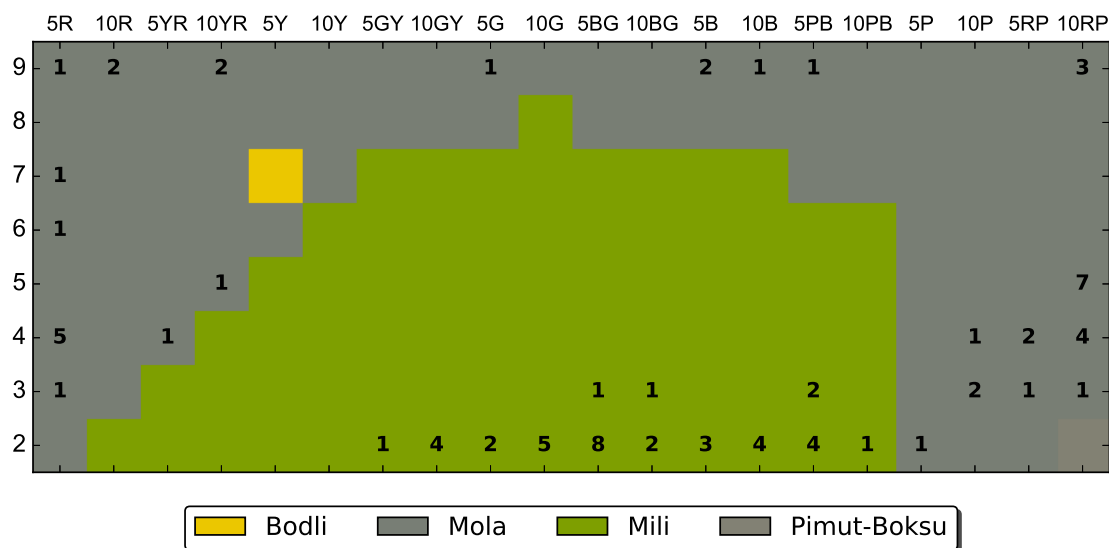


Figure 3.3: Naming data for the Dani language, overlaid with the empirical focus distributions for *mili* and *mola*.

Analyses and results. Our analyses were conducted at the language level, because the data for this language were reported at the language level. We analyzed model predictions for only the terms *mili* and *mola* because these were the only terms for which focus data were reported. Each model returned a ranked list of all chips in the array. For each term, we recorded the number n of chips that received 1 or more focus choices. We then computed rank position by averaging

together the rank positions of these n chips in the ranked list produced by the model. To compute QF distance, we compared the empirical histogram of focus choices for a given term to a model-predicted histogram in which each of the n top-ranked chips received a count of 1 and each other chip received a count of 0. The results of these analyses are provided in Table 3.2 below. We find that the Representativeness model outperforms its competitors by both metrics.

Table 3.2: Quantitative assessment of each model against Dani focus distribution.

Model	QF	RP
Representativeness	2.09	35.29
Likelihood	2.54	39.99
Prototype	2.39	42.70
Exemplar	2.88	36.44
Chroma	2.64	45.59

Berinmo

The Berinmo data we consider (Roberson et al., 2000) were originally collected in an attempt to replicate and extend earlier work based on Dani (Heider, 1972b). For this reason, the stimuli were the same as those of the earlier Dani work. Roberson et al. (2000) reported results on Berinmo color memory that differed in important respects from those obtained from Dani, but the similarity in stimuli and procedure make the two studies directly comparable with respect to naming and focus data. As in the case of Dani, Berinmo naming data and focus data were both reported at the language level. The Berinmo naming data provided by Roberson et al. (2000) were presented in the form of a mode map, which we illustrate in Figure 3.4 below. The focus data for each term were reported in the form of a histogram over the naming grid, here shown overlaid on top of the naming data.

Analyses and results. Because the data format for Berinmo was the same as that for Dani, we followed the same procedure as with Dani. The results of the Berinmo analyses are provided in Table 3.3 below. The Representativeness model outperforms its competitors by both metrics.

3.5 Color categories with unusual extensions

So far, we have shown that a model of focal colors as representative members of categories accounts well for the distribution of WCS best example choices across the stimulus array, as well as for the distribution of best example choices within many languages. These results are consistent with our proposal that color foci are representative members of categories, and that their location in color space reflects category extensions, which are in turn shaped by the functional need for color

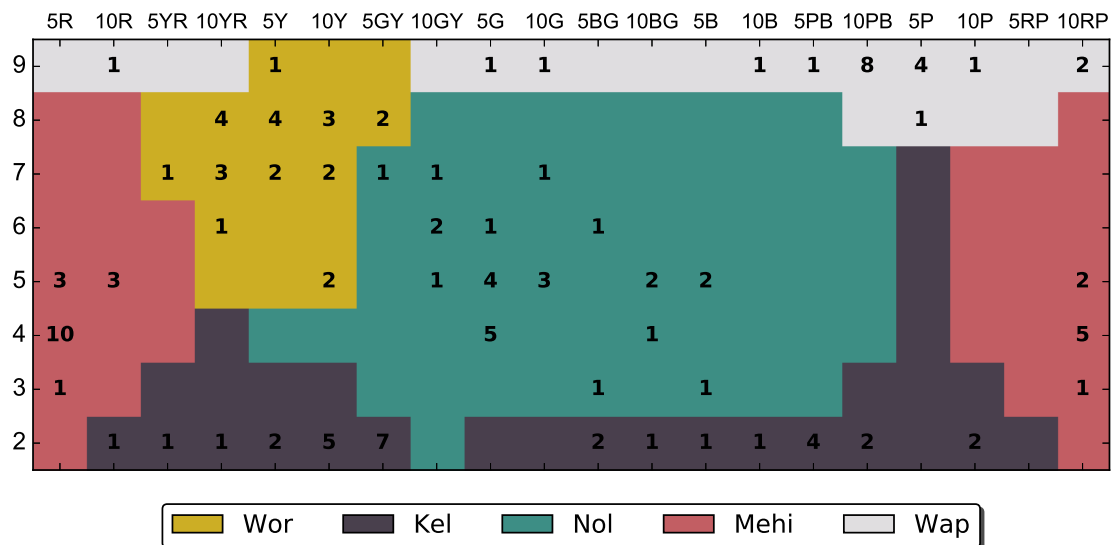


Figure 3.4: Naming data for the Berinmo language, overlaid with the empirical focus distribution.

Table 3.3: Quantitative assessment of each model against Berinmo focus distribution.

Model	QF	RP
Representativeness	1.37	12.27
Likelihood	1.69	14.31
Prototype	1.68	16.58
Exemplar	1.66	15.52
Chroma	2.64	24.75

naming systems to be informative (Regier et al., 2007, 2015). However, the analyses we have seen so far do not discriminate between this hypothesis and a natural alternative: the traditional view of color foci as reflections of unalterably universal privileged points in color space. For languages with common color-naming systems, the two hypotheses make the same prediction: foci should tend to fall in the canonical positions shown in Figure 3.2(a). This is predicted on the traditional universal-foci account, because these are the proposed locations of the universal foci. Roughly the same outcome is predicted by our account, as seen in Figure 3.2(b).

In a final investigation, then, we attempt to discriminate between these two hypotheses. The hypotheses diverge in their predictions for color categories that have unusual extensions. If foci are a universal groundwork for color naming, then in such unusual cases, foci will fall in the universal (canonical) positions, despite the non-canonicity of the category extensions. In contrast, our account predicts that in such cases, foci should follow the category extensions, and fall in non-canonical positions. What is not yet known is: (a) whether the representativeness model

accounts for non-canonical empirical distributions better than universal foci do, and (b) whether the representativeness model also outperforms the other competing models considered above on non-canonical or unusual categories generally. To test these open questions, we began by defining a formal model of the universalist account and a measure of category unusualness.

Universalist model

Like the other models we consider, the universalist model assigns a score indicating how good a given color chip x is as an example of color term t . The score for the universalist model is determined by the empirical WCS focus distribution shown in Figure 3.2(a), gated by the extension of category t :

$$U(x,t) = W(x) \times I(x,t) \quad (3.11)$$

where $W(x)$ is the number of times color chip x was chosen as a best example of any term by any speaker in any language in the WCS, and $I(x,t)$ is 1 if $x \in t$ and 0 otherwise.

Category unusualness

We took the extension of a major color term² to be that set of chips in the stimulus array that were named by that color term by a plurality of speakers, and represented that set of chips as a set of points in CIELAB space. We took the dissimilarity between any two categories $X = \{x_1, \dots, x_p\}$ and $Y = \{y_1, \dots, y_q\}$ to be the Hausdorff distance $H(X, Y)$ between the two corresponding sets of points. The Hausdorff distance (Huttenlocher, Klanderman, & Rucklidge, 1993) is determined by finding, for each point in each set, the nearest point in the other set, and selecting the largest of the resulting distances:

$$H(X, Y) = \max(h(X, Y), h(Y, X)) \quad (3.12)$$

where

$$h(X, Y) = \max_{x \in X} \min_{y \in Y} \|x - y\| \quad (3.13)$$

and $\|x - y\|$ is the Euclidean distance between points x and y in CIELAB space. The unusualness of category c , $u(c)$, is the average dissimilarity of c to all major color categories in the WCS:

$$u(c) = \frac{1}{N} \sum_{i=1}^N H(c, c_i) \quad (3.14)$$

where i indexes over all N major color categories in the entire WCS. Example categories at varying levels of unusualness are presented in Appendix B. We also pooled together the focus choices for this term across speakers of the language in question, as well as the analogous focus predictions by each of the models. Finally, for both evaluation measures, we noted which model performed

best for this category (had the lowest rank position or lowest QF distance of the empirical focus distribution).

Figure 3.5 shows the results of this analysis. The scatterplots (left panels) show each category as a dot. The dot's position represents the category's unusualness (horizontal axis), and the score (Rank position in the top panel and QF distance in the bottom panel) of the best-performing model for that category (vertical axis: lower is better for both measures). The dot's color represents the best-performing model for that category. In both scatterplots it can be seen that the universalist model (red) performs well for the least unusual (most usual or common) categories; this is particularly apparent when using QF distance. This is unsurprising because the universalist model is based on universal tendencies in focus choices. However for higher values of unusualness, the rep-

²We considered a color term to be a major color term in a language if it was used by a plurality of speakers of the language for at least 10 of the 330 chips of the stimulus array (Regier et al., 2015); otherwise we considered it a minor term and excluded it from this analysis.

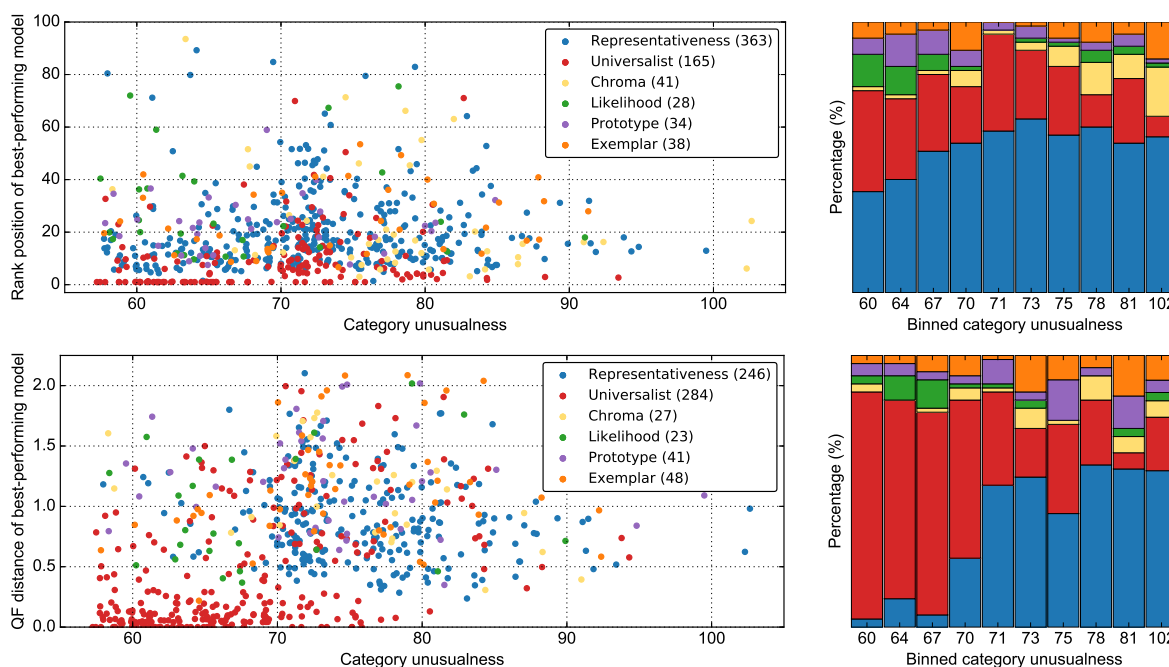


Figure 3.5: Effect of category unusualness. Left panels (scatterplots): Each dot represents a color category in the WCS, and the dot's color represents the best-performing model for that category. The horizontal axis represents category unusualness, and the vertical axis represents the model performance: rank position (top panel) and QF distance (bottom panel) of the best-performing model for that category. Right panels (bar charts): The horizontal axis again represents category unusualness, this time partitioned into 10 bins with the same number of categories per bin. The stacked bars show, for each level of unusualness, the proportion of categories at that level of unusualness that were best predicted by each model.

representativeness model (blue) begins to outperform the universalist model, and others, as predicted. This progression of increasing dominance for the representativeness model with increasing category unusualness is shown more schematically in the stacked barplots of the right panels. These findings suggest that when boundaries fall in non-canonical positions, foci do as well. Moreover, foci for unusual categories are better predicted by representativeness than they are by expectations based on strictly universal foci, or by the other models.

3.6 Discussion

Focal colors, or best examples of color terms, lie at the center of the debate over color naming. Focal colors have traditionally been viewed either as the underlying source of color naming universals, or as derived from category boundaries that vary widely with local linguistic convention. In contrast, we have argued for a novel account of this disputed construct, that synthesizes aspects of the traditionally opposed views and accounts for data that challenge those views. We have proposed that focal colors are representative members of color categories. This simple idea accounts for universal tendencies in focal colors, yet also correctly predicts some deviation from those universal tendencies, particularly for color categories with unusual extensions. Our proposal coheres naturally with a recent explanation of color naming in terms of the functional need for informative communication over irregularly shaped perceptual color space (Jameson & D'Andrade, 1997; Regier et al., 2007). That view explains cross-language universals and variation in color naming without reference to a small set of focal colors, and it leaves the nature of focal colors unexplained. Our proposal fills that gap. Taken together, the two proposals suggest a single overall account of color naming: color categories across languages assume the forms they do because of functional pressure for informative communication given the structure of color space, and foci are representative members of those categories.

Chapter 4

Large-scale word learning

4.1 Introduction

Many problems solved by the mind conform to the same abstract computational formulation: How should a property be generalized to novel stimuli from a set of stimuli observed to have the property? As there are many ways to extend the property that are consistent with some observed evidence, these are problems of *induction*, where the evidence constrains, but does not determine, the solution to a problem. The Bayesian generalization framework (Shepard, 1987; Tenenbaum & Griffiths, 2001a) has been remarkably successful at explaining human generalization behavior in a wide range of domains. However, its success is largely dependent on the choice of a hypothesis space and a prior probability distribution on hypotheses, which are usually hand constructed by the researcher for each specific problem. This is unsatisfying practically, because the models do not scale beyond the originally modeled problem, and theoretically, as it is unclear whether their success is due to the cleverness of the modeler and not because of a deep mathematical property of the computational problem that people solve.

One possible solution is to use existing sources of information about the organization of a domain as the basis for specifying a hypothesis space and prior. This helps address both the practical and the theoretical concerns raised by the Bayesian generalization model. In this chapter, we use this approach to show how a hypothesis space and prior can be constructed automatically from a large online database, making it possible to apply the Bayesian generalization framework to a wide range of naturalistic stimuli. We focus on one specific generalization problem, word learning, where people learn new words from observing a few objects that can be labeled with that word. Given that the number of possible extensions of a word is essentially infinite, learning the objects referred to by a word is a very difficult inductive problem (Quine, 1975). F. Xu and Tenenbaum (2007b) showed how the Bayesian generalization framework could be used to explain how people learn new words. However, to construct the hypothesis space of their Bayesian model, F. Xu and Tenenbaum (2007b) elicited approximately 400 similarity judgments from their participants. Clearly this is not practical to extend into every domain where people learn words. Thus, word learning is an appropriate setting for exploring novel methods of constructing hypothesis spaces

and prior distributions.

We propose a method for automatically constructing the hypothesis space and prior distribution of a Bayesian word learning model using freely available online resources. In particular, we use WordNet (Fellbaum, 2010; Miller, 1995) as an initial source for automatically creating the hypothesis space, and ImageNet (Deng et al., 2009) as a source of naturalistic images that can be used as stimuli to test the resulting model in behavioral experiments. WordNet is a popular lexical database of English comprised of over 100,000 relational sets of synonyms. ImageNet is a large ontology of images conforming to the hierarchical structure of WordNet, with the aim of providing over 500 high-quality images per noun in WordNet. These resources allow us to construct hypothesis spaces and prior distributions for word learning without eliciting a single judgment from participants and test the resulting model on a much larger scale than was previously possible. We demonstrate that the Bayesian model formulated from WordNet captures participant judgments in three behavioral experiments, addressing the practical and theoretical issues with Bayesian models discussed earlier.

The plan of the rest of the chapter is as follows. In the next section we review the Bayesian generalization model and examine how F. Xu and Tenenbaum (2007b) constructed the hypothesis space for their Bayesian word learning model. We then show how to build a hypothesis space from WordNet that can be used to evaluate word learning models on arbitrary conceptual domains. Afterwards, we present two experiments utilizing this hypothesis space to validate the method: one that replicates a previous study of adult word learning, and one that investigates word learning for a set of complex concepts in novel domains. We then demonstrate how to adopt an existing large-scale computer vision challenge to automatically construct word learning experiments for 4,000 concepts. Finally, the chapter concludes with a discussion of the implications of this work.

4.2 The Bayesian generalization framework

The Bayesian word learning model is a special case of the Bayesian generalization framework. This framework has been used to model generalization in a number of domains including dimensional concepts (Austerweil & Griffiths, 2010; Shepard, 1987; Tenenbaum, 1999), numerical concepts (Tenenbaum, 2000), sequential rules (Austerweil & Griffiths, 2011), rule-based categorical concepts (Goodman et al., 2008), and word learning (F. Xu & Tenenbaum, 2007b). Typically, problems are formulated in this framework as follows: Assume we observe n positive examples $\mathbf{x} = \{x_1, \dots, x_n\}$ of concept C and want to compute $P(y \in C | \mathbf{x})$, the probability that some new object y belongs to C given the observations \mathbf{x} . We compute this probability by using a hypothesis space \mathcal{H} , which is a set of hypothetical concepts, where each hypothesis is defined by the objects that would be members of the concept if the hypothesis were true, $P(\mathbf{x}|h)$.

Defining a Bayesian generalization model amounts to defining a hypothesis space \mathcal{H} , a prior probability distribution over hypotheses, $P(h)$, and for each hypothesis, a likelihood function, $P(\mathbf{x}|h)$, indicating the probability of observing a set of objects \mathbf{x} given that the hypothesis is true. A typical definition of the likelihood follows from assuming *strong sampling*, where objects are

generated uniformly at random from the true hypothesis (Tenenbaum & Griffiths, 2001a)

$$P(\mathbf{x}|h) = \begin{cases} 1/|h|^n & \text{if } \mathbf{x} \subset h \\ 0 & \text{otherwise} \end{cases}. \quad (4.1)$$

This likelihood function instantiates the *size principle* for scoring hypotheses: hypotheses containing a smaller number of objects assign greater likelihood than hypotheses with more objects to the same set of objects (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001a).¹ The prior distribution over hypotheses, $P(h)$ depends on the domain and in previous literature has ranged from a simple uniform distribution over the hypothesis space (Shepard, 1987) to a stochastic process over tree structures (Kemp & Tenenbaum, 2009). Given the prior and likelihood, the posterior probability that a hypothesis is true given a set of objects belonging to a novel concept, $P(h|\mathbf{x})$, follows from Bayes' rule: $P(h|\mathbf{x}) \propto P(\mathbf{x}|h)P(h)$. From this, we can compute the probability that a new object y is also a member of the concept C by averaging the predictions of all hypotheses weighted by their posterior probabilities:

$$P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \quad (4.2)$$

where $P(y \in C|h) = 1$ if the new object y is in hypothesis h , and 0 otherwise.

Word Learning as Bayesian Inference

F. Xu and Tenenbaum (2007b) used the Bayesian generalization model to explore how people learn the appropriate generalizations for new words. It is commonly held that a child's word learning development (especially for nouns) follows a *taxonomic* assumption, where new words refer to classes in a tree-structured hierarchy (Markman, 1991; Waxman, 1990). Furthermore, there is also a *basic-level* bias (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) in which both children and adults tend to generalize new words more often to a category at an intermediate level in a taxonomy. Thus, F. Xu and Tenenbaum (2007b) derived a taxonomic-based hypothesis space for their Bayesian word learning model by applying hierarchical clustering (Duda & Hart, 1973) to the perceived similarity of every pair of objects. The hypothesis space, prior and likelihood are defined by the tree resulting from this hierarchical clustering.

Nodes in the tree represent potential words (hypotheses) which extend to all the leaves they cover, where the leaves of the tree correspond to the domain of possible objects. The height of a node h (minimal distance from the node to a leaf) is a measure of the average pairwise dissimilarity of objects covered by node h and approximates the heterogeneity of the objects that can be called that word. The intuition that more distinctive clusters are more likely to have distinguishing names,

¹Using a likelihood with strong sampling (analogous to a knowledgeable teacher showing positive examples of the concept) as opposed to *weak sampling*, where objects are generated uniformly at random from any hypothesis (analogous to learners searching for examples themselves), has support in developmental research showing both adults and children are sensitive to these assumptions and generalize more conservatively under the former (F. Xu & Tenenbaum, 2007a). See also Navarro, Dry, and Lee (2012) for a further empirical investigation of different types of likelihood functions.

was incorporated by defining the prior $P(h)$ to be proportional to the branch length separating node h from its parent:

$$P(h) \propto \text{height}(\text{parent}(h)) - \text{height}(h), \quad (4.3)$$

where $\text{parent}(h)$ returns the parent of node h . To incorporate a *basic-level* bias (Rosch et al., 1976) in the prior, the probability of hypotheses at the basic level were 10 times the value given by Equation 4.3 (see below for example basic-level concepts). As the height of node h also approximates the number of objects in the extension of the possible word h , the likelihood of observing n objects called word h is defined as

$$P(\mathbf{x}|h) \propto \left[\frac{1}{\text{height}(h) + \epsilon} \right]^n, \quad (4.4)$$

where ϵ is a small constant so that the leaf hypotheses (those that refer to only a single object) do not have infinite likelihood (as their height is zero).

Using this framework, F. Xu and Tenenbaum (2007b) accurately predicted how people extend words to new objects depending on the diversity and number of objects labeled with that word. In a set of experiments on both adults and children, they showed participants one or more positive examples of a novel word while manipulating the taxonomic relationship of the objects the word referred to. For example, participants might observe one Dalmatian, three Dalmatians (exemplars at the subordinate-level), a Dalmatian, terrier, and mutt (exemplars at the basic-level), or a Dalmatian, pig, and toucan (exemplars at the superordinate-level) being labeled with a novel word (e.g. “fep”). After observing a word refer to one or three example objects at the subordinate, basic, or superordinate-level, they were asked whether the word referred to novel subordinate, basic, superordinate, and out-of-domain objects.

When participants were given one example of an object that refers to a word (e.g. one Dalmatian), they tended to select the subordinate-level matches (e.g. the two other Dalmatians) and the basic-level matches (e.g. the two non-Dalmatian dogs). However, when they were shown three subordinate-level examples of a concept (e.g. three Dalmatians), the participants tended to choose only the subordinate-level matches (e.g. they only believed the word referred to the two other Dalmatians). The Bayesian word learning model captured this phenomenon because the prior favors words at the basic-level, but the likelihood favors words at the subordinate-level, and the likelihood’s weight increases exponentially in the number of objects.

Unfortunately, the manner in which the hypothesis space was constructed (through hierarchical clustering on pairs of similarity judgments) poses a serious constraint to assessing the model’s validity. To construct the hypothesis space in the three domains tested by F. Xu and Tenenbaum (2007b), where there are 15 images per concept, each participant had to provide roughly 400 similarity judgments. To test how well this framework extends to new concepts and domains using their method for constructing the hypothesis space, an impractically large quantity of human judgments would need to be elicited. In the following section, we introduce an alternative method of constructing a hypothesis space for the Bayesian word learning model, which allows for testing the framework without eliciting any judgments from participants.

4.3 Constructing a hypothesis space for Bayesian word learning

Using an online word ontology, we can automatically construct the hypothesis space of a Bayesian word learning model. WordNet is a large lexical database of English represented as a network of words linked by directed edges denoting semantic relatedness (Fellbaum, 2010; Miller, 1995). Its structure was manually designed to group lexical concepts in an “is-a” hierarchy based on the many-to-one mapping of synonyms. For example, a Poodle “is-a” type of dog, thus WordNet has a directed edge from the node for *dog* to the node for *Poodle*. As WordNet is hierarchically structured like the hypothesis space used by F. Xu and Tenenbaum (2007b), it is an ideal candidate for constructing our hypothesis space.

Using a hypothesis space derived from an existing online ontology, we can better test the predictions of different generalization theories for word learning by examining their predictions for a large range of concepts. In the rest of this section, we present the method used to construct a hypothesis space from WordNet and outline the implementations of three generalization models using this hypothesis space for large-scale word learning.

In the context of the Bayesian generalization framework, the hypotheses correspond to subsets of the universe of objects that are psychologically plausible candidates as extensions of concepts (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001a). Using WordNet as the basis of our hypothesis space, the set of objects is the set of leaf nodes from the noun-space of the directed graph and the hypotheses correspond to both the inner nodes of the directed graph and the leaf nodes, which distinguish between objects at the subordinate-level. To construct a hypothesis space from WordNet, we first extracted a tree from the 82,115 noun nodes of WordNet.² The nodes are hypotheses, which represent possible words, and form the hypothesis space for the model.

From this graph we create a hypothesis space that is a binary matrix, \mathcal{H} , whose rows are the objects (64,958 leaf nodes from the graph) and columns are the hypotheses (82,115 nodes, 17,157 of which are inner nodes and 64,958 are leaf nodes). Each entry (i, j) of the matrix \mathcal{H} denotes whether or not hypothesis node j is an ancestor of leaf node i in the WordNet graph (with a 1 indicating it is). The leaf nodes are included as hypotheses so that the model distinguishes between subordinate objects.

Generalization models

With a hypothesis space derived from WordNet, we now have the ability to test the Bayesian model of word learning on a much larger scale. In addition, we can use the hypothesis matrix as a feature space for testing alternative models. We compare the Bayesian model against two similarity models: a prototype model and an exemplar model. Given a set of examples $\mathbf{x} = \{x_1, \dots, x_n\}$ representing some concept C (where the elements of \mathbf{x} correspond to rows in the hypothesis matrix \mathcal{H}), we can compute a score for each row $y \in \mathcal{H}$ denoting the probability that y is also a member

²Technically WordNet is a directed acyclic graph because some nodes have multiple parents (the method still works in these cases).

of C . We present the different ways to compute this score below.

Bayesian model. This is the Bayesian generalization framework that we discussed earlier. We used strong sampling for the likelihood, $P(\mathbf{x}|h)$, which is computed via Equation 4.1, where the size of h is the number of nodes that can be reached by a directed path from h . This simply corresponds to the sum of the elements in the column corresponding to h .

The prior $P(h)$ was defined to be Erlang distributed in the size of the hypothesis (a standard prior over sizes in Bayesian models with preference for intermediate sized hypotheses; Shepard (1987), Tenenbaum (2000))

$$P(h) \propto (|h|/\sigma^2)\exp\{-|h|/\sigma\}, \quad (4.5)$$

where the σ parameter was set to 200 by hand fitting the model predictions to all human responses (the same value was used in both experiments). This value favors medium sized hypotheses, which is roughly equivalent to a basic-level bias. The probability that word C extends to object y after observing a set of objects called C is

$$\text{Bscore}(y) = P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \quad (4.6)$$

where $P(y \in C) = 1$ if $y \in h$ and 0 otherwise, and $P(h|\mathbf{x})$, is the posterior distribution over hypotheses.

Prototype model. In this model, we define the prototype of a set of objects, x_{proto} , to have those features owned by a majority of the objects in the set. The generalization measure for an object y is

$$\text{Pscore}(y) = \exp\{-\lambda_p \text{dist}(y, x_{\text{proto}})\}, \quad (4.7)$$

where $\text{dist}(\cdot, \cdot)$ is the Hamming distance between the two vectors and λ_p is a free parameter (for all of the results presented here, $\lambda_p = 0.15$, optimized by hand using half-interval search). Pscore was then normalized over all objects y in the hypothesis space (all leaf nodes).

Exemplar model. We define the exemplar model using a similar scoring metric as the prototype model, except rather than computing the distance of object y to a single prototype vector, we compute a distance for each item x_j in the set of observations \mathbf{x} . The exemplar generalization measure is thus computed as

$$\text{Escore}(y) = \sum_{x_j \in \mathbf{x}} \exp\{-\lambda_e \text{dist}(y, x_j)\}, \quad (4.8)$$

where $\text{dist}(\cdot, \cdot)$ is the Hamming distance between two vectors and λ_e is a free parameter (for all of the results presented here, $\lambda_e = 0.20$, optimized using half-interval search). Escore was then normalized over all objects y in the hypothesis space (all leaf nodes).

4.4 Behavioral experiments to validate our approach

To evaluate the performance of our models using the WordNet-based hypothesis space, we conducted two experiments using the paradigm of Xu and Tenenbaum (2007). The first experiment replicates F. Xu and Tenenbaum (2007b) on their three object taxonomies (animals, vehicles, and vegetables), which validates our approach for constructing a hypothesis space from WordNet and using images from ImageNet as stimuli. The second experiment extends the paradigm into three previously unexplored domains (clothing, containers, and seats), which have hierarchical structure, but it is not as clear how well they conform to a natural basic-level taxonomy (Rosch et al., 1976).

Experiment 1: Replicating previous results

Participants. Thirty four participants were recruited via Amazon Mechanical Turk and compensated \$0.05 for each trial (training set) completed out of twelve possible. Each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

Stimuli and Procedure. Within each taxonomy, the stimuli consisted of the images of objects distributed across the superordinate, basic and subordinate-levels, and subsequently split into training and test sets. The training sets were the labeled objects given to participants of which there were four conditions: a single subordinate-level example (e.g. a Dalmatian); three examples of the same subordinate-level object (e.g. three Dalmatians); the subordinate-level object and two basic-level objects (e.g. a Dalmatian, a Shih Tzu, and a Beagle); and the subordinate object and two superordinate-level objects (e.g. a Dalmatian, a hippopotamus, and a toucan). This corresponds to twelve trials total (four conditions for each of the three object taxonomies).

The test sets were the same regardless of the training set and consisted of eight objects matching the currently tested taxonomy: two subordinate examples (e.g. two other Dalmatians); two basic-level examples (e.g. a Cocker Spaniel and a Corgi); and four superordinate examples (e.g. a cat, a bear, a sea lion, and a horse). There were also sixteen non-matching objects in the test set corresponding to the objects that match the two other taxonomies.

For each trial, participants were instructed that they needed to help a cartoon frog who speaks a different language from us, pick out objects that he wants. The frog shows one or more examples of a novel word (e.g. “FEP”) and the participant is instructed to select other items that are a “FEP” from the objects comprising the test set. A unique novel word was associated with each of the twelve trials. See Figure 4.1 for an example.

Results. Figure 4.2 shows the results of this experiment, along with the predictions of the different generalization models. For each training set condition, the data for each test item has been averaged over participants and domains. The generalization judgments of participants (left-most panel of Figure 4.2) follows the same qualitative trend as those reported in F. Xu and Tenenbaum (2007b). There is a sharp drop in generalization to basic-level objects when seeing only a single subordinate example compared to the condition when seeing three subordinate examples.

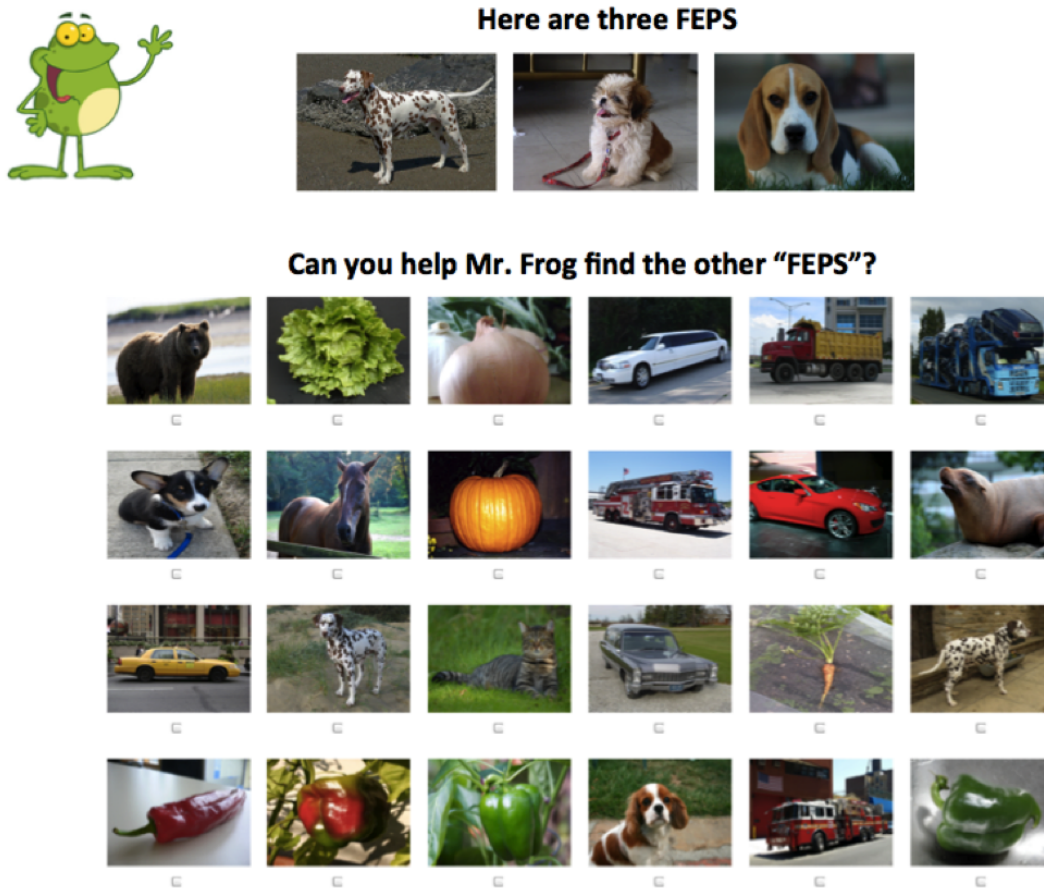


Figure 4.1: Example word learning experiment trial using ImageNet as a source of stimuli.

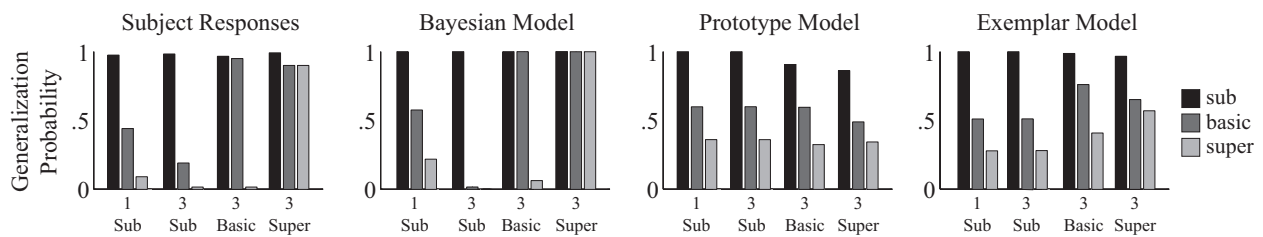


Figure 4.2: Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the three domains in Experiment 1. The generalizations for non-matching items are omitted for brevity (neither the participants chose nor the Bayesian model predicted non-matching objects, while the prototype and exemplar models predicted non-matches less than 4% of the time for each condition).

The Bayesian model predictions (second panel from the left) exhibits this same generalization pattern ($r^2 = 0.98$), while the prototype and exemplar models do not ($r^2 = 0.66$ and $r^2 = 0.84$, respectively). This validates our method of automatically creating hypothesis spaces with WordNet.

Experiment 2: Novel Domains

Constructing hypothesis spaces automatically from online resources allows us to easily extend the Bayesian model to new domains. In this experiment, we demonstrate the power of our approach by applying the model to three taxonomic domains (clothing, containers, and seats), and empirically validate its predictions.

Participants. Thirty six participants were recruited via Amazon Mechanical Turk and compensated \$0.05 for each trial completed out of twelve possible. As in Experiment 1, each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

Stimuli and Procedure. Table 4.1 contains examples of the objects we used for training in the three hierarchical domains (clothing, containers, and seats). As in Experiment 1, the same test objects were used for every training set, and the “non-match” test objects were the objects in the test set which match the two other taxonomies that are not contained in the training set. As before, this corresponds to twelve trials total. The procedure was identical to Experiment 1.

Results. Figure 4.3 presents the averaged results of how participants and the Bayesian model generalized the learned words to the test objects based on the observed training set across the different domains in Experiment 2. Across the three domains, the generalization probabilities of the participants and Bayesian model with the same parameters are extremely similar. This is



















Object level	Clothing		Containers		Seats	
	1	2	1	2	1	2
Subordinate						
Basic						
Superordinate						

Table 4.1: Training domains and example stimuli at various taxonomic levels for Experiment 2.

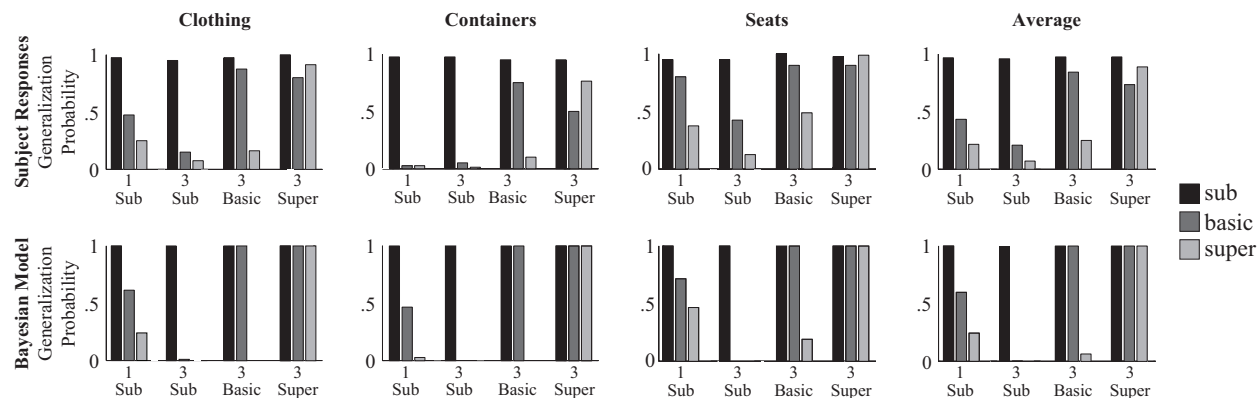


Figure 4.3: Participant generalization judgments and the predictions of the Bayesian model for Experiment 2. From left to right, the columns present the results for the three taxonomies (clothing, containers, and seats) and average results. Non-matching items are omitted for brevity (participants only chose non-matches twice, both in the containers domain).

exemplified in the very good quantitative model fit on the averaged data ($r^2 = 0.95$). Due to poor performance in the previous experiment, the prototype and exemplar models are omitted from Figure 4.3 for brevity ($r^2 = 0.80$ and $r^2 = 0.90$ averaged over domains, respectively).

Using the hypothesis space constructed automatically from WordNet explains the idiosyncrasies of participant generalization behavior in each domain ($r^2 = 0.97, 0.88$, and 0.91 , for clothing, containers, and seats respectively). For example, the model accurately predicts that participants would generalize most broadly in the seats domain for the single exemplar and three basic-level exemplar training sets. Additionally, the model captures that people generalized the least in the containers domain for the three subordinate-level exemplar training sets. This would not have been possible if the hypothesis space for each domain had the same structure.

Note that there is a larger amount of variance between model predictions and human performance in Experiment 2 than Experiment 1. We believe that this is due to the domains not conforming to a natural taxonomy. For example, it is unclear if box should be the basic-level category for a mail box and a cigar box; however, this is the basic level of these objects provided by WordNet. Regardless, the good quantitative fit of the Bayesian model’s predictions provides evidence that using WordNet as a hypothesis space for word learning can capture people’s generalizations even for hierarchies without clearly defined basic-level concepts. In the next section we explore how to connect our extended word learning model to problems in Computer Vision and Machine Learning, further leveraging the use of online resources to perform an even larger-scale evaluation.

4.5 Large-scale word learning

Now that we have validated the model, we consider a much larger and conceptually-diverse experiment to better understand how people generalize at different levels of categorization and familiarity

with objects. In constructing the stimuli for Experiments 1 and 2, we needed to manually select the nodes used to represent different taxonomic concepts. However, large-scale experimentation requires an efficient scheme to generate test data across varying levels of a concept hierarchy. To this end, we developed a fully-automated procedure for constructing a large-scale dataset suitable for a challenge problem focused on visual concept learning. In what follows we first show how to construct this dataset and then use it to run a large-scale experiment over 4,000 different concepts.

Constructing a large-scale dataset

We used the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) data as the basis for automatically constructing a hierarchically-organized set of concepts at four different levels of abstraction. We had two goals in constructing the dataset: to cover concepts at various levels of abstraction (from subordinate concepts to super-ordinate concepts, such as from Dalmatian to living things), and to find query images that comprehensively test human generalization behavior. We address these two goals in turn.

To generate concepts at various levels of abstraction, we use all 1,000 words in the ILSVRC hierarchy as concept candidates, associating each word with its leaf node in WordNet as the most specific level concept. We then generate three more levels of increasingly broad concepts along the path from the leaf to the root for each leaf node in the hierarchy. Examples from these concepts are used as the training set. Specifically, we use the leaf node class itself as the most specific trial type L_0 , and select three levels of nested concepts L_1, L_2, L_3 which correspond to three intermediate nodes along the path from the leaf node to the root. We choose the three nodes that maximize the combined information gain across these levels:

$$C(L_{1\dots 3}) = \sum_{i=0}^3 \log(|L_{i+1}| - |L_i|) - \log |L_{i+1}|, \quad (4.9)$$

where $|L_i|$ is the number of leaf nodes under the subtree rooted at L_i , and L_4 is the whole taxonomy tree. As a result, we obtain levels that are “evenly” distributed over the taxonomy tree. Such levels coarsely correspond to the sub-category, basic, super-basic, and super-category levels in the taxonomy. For each concept, the training images shown to participants as examples of that concept were randomly sampled from five different leaf node categories from the corresponding subtree in the ILSVRC test images. For example, the four levels used in Figure 4.4 are *blueberry*, *berry*, *edible fruit*, and *natural object* for the leaf node *blueberry*.

To construct the test set, we randomly sample twenty query images as follows: three each from the L_0, L_1, L_2 and L_3 subtrees, and eight distractor images from L_4 , nodes found outside the broadest subtree rooted at L_3 . This ensures a complete coverage over in-concept and out-of-concept queries. We explicitly made sure that the leaf node classes of the query images were different from those of the examples if possible, and no duplicates exist among the 20 queries. Note that we always sampled the example and query images from the ILSVRC test images, allowing us to subsequently train our models with the training and validation images from the ILSVRC dataset while keeping those in the visual concept learning dataset as novel test images.

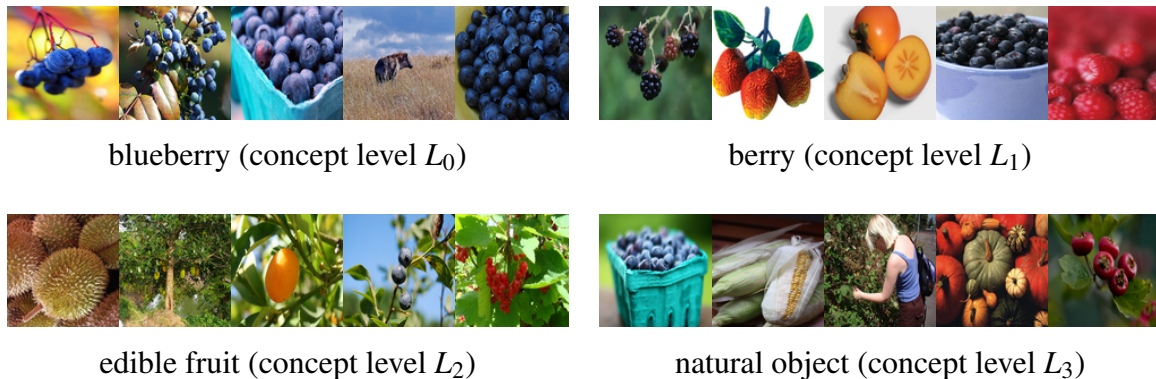


Figure 4.4: Concepts constructed from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Training set images sampled from the four levels for the leaf node *blueberry*, with the levels L_0, \dots, L_3 corresponding to the concepts *blueberry*, *berry*, *edible fruit*, and *natural object*, respectively.

Experiment 3: The ImageNet Large Scale Visual Recognition Challenge

Participants. Participants were recruited via Amazon Mechanical Turk and compensated \$0.05 for each trial completed out of 4,000 possible. As in the previous experiments, each participant completed as many trials as he or she wished, and ten unique participants completed each trial. All participant responses were used.

Stimuli and Procedure. We created 4,000 identical concepts (four for each leaf node) using the protocol above. For each concept, the training set consisted of five example images and the test set consisted twenty query images. Following the previous experiments, participants were asked to help a cartoon frog pick out the objects he wants from the test set. A total of 40,000 trials were collected, and a total of 100,000 images were shown to the participants.

Results. Figure 4.5 presents the averaged results of how participants and each of the models generalized to the test objects based on the observed training set across the different domains in this large-scale experiment. Here we see the exemplar and prototype models approach human performance much more closely than the Bayesian model. This is supported quantitatively by better model fits averaged across all data ($r^2 = 0.99, 0.92$, and 0.74 , for the exemplar, prototype, and Bayesian models, respectively). However, we note the Bayesian model assumes perfect identification of all images, and we question whether five examples is too few for grasping the conceptual coverage of categories in level L_3 . For example, in Figure 4.4, “natural object” might be difficult to infer from just these five example images. This provides an opportunity for future work: finding the appropriate number of examples given the level of generalization one is trying to convey or learn.

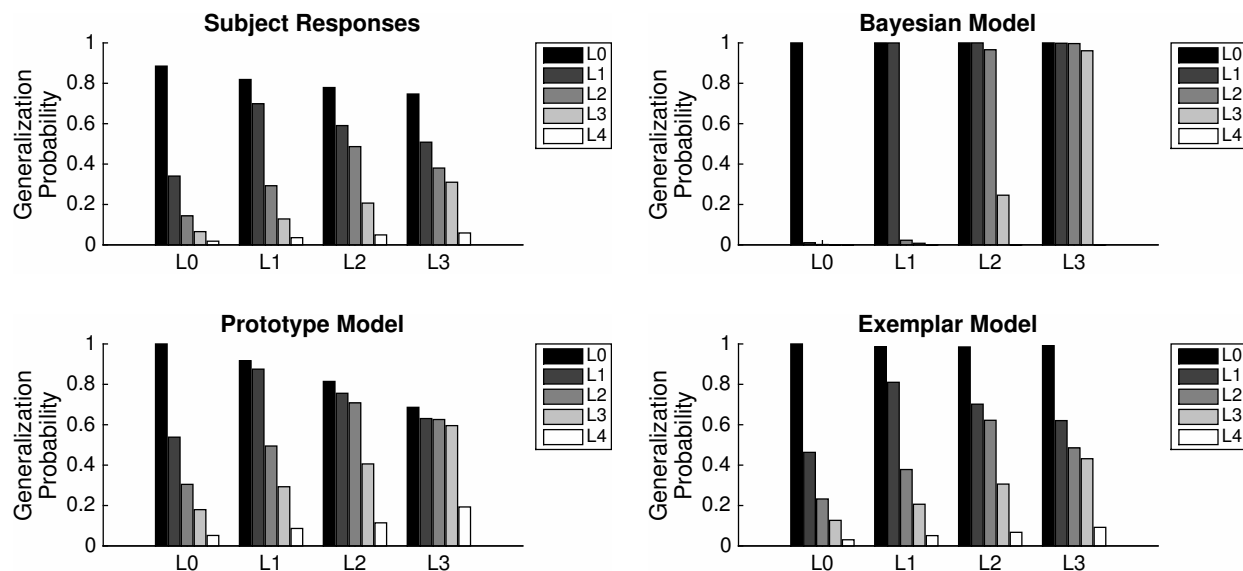


Figure 4.5: Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the 4,000 concepts in Experiment 3. The horizontal axis for each presents four levels at which training set examples were provided (L_0 to L_3). At each level, five bars show the proportion of test set images from levels L_0 to L_4 that were selected as instances of the concept (where L_4 denotes non-matching items), with the results averaged over all domains.

4.6 Discussion

Although the Bayesian generalization framework has had success in explaining human generalization behavior across a number of domains, the hypothesis spaces are typically small-scale and hand-constructed, which is unsatisfying both practically and theoretically. In this chapter, we explored a proposal for automatically constructing the hypothesis space using an online resource as a potential solution to the methodological challenges posed by this problem. We validated a Bayesian model using this hypothesis space by showing it can replicate previously observed word learning phenomena, and demonstrating it can also explain how participants learned words in three novel domains of varying taxonomic assumptions. Using the automatically constructed hypothesis space, the model predicted the subtle changes in participants' word learning behavior across multiple domains, demonstrating the practical and theoretical benefits of our approach. We then conducted a large-scale evaluation of this model by adopting an existing computer vision challenge which utilizes the same resources we draw from, WordNet and ImageNet, with different results. In Experiment 3, we found that the exemplar model had the best predictions of people's generalization behaviors, most likely due to difficulty inferring high-level concepts that have wide coverage over leaf nodes from very few examples; a difference in the conceptual structure automatically sampled from WordNet for higher-level concepts. This opens more questions about the taxonomic nature of concepts more broadly, and provides many opportunities for future research.

Chapter 5

Conclusions

The work in the present dissertation explores how people learn and reason with abstract knowledge, focusing on the kinds of processes and representations used in semantic memory. In particular, I presented three case studies, each of which investigated different assumptions for the semantic representations and algorithms used for modeling a cognitive task. In this final chapter I review some remaining questions from each of the case studies and outline potential directions for future research. I then present theoretical connections across the chapters and to larger bodies of work, concluding with a summary of the dissertation contributions.

5.1 Remaining questions and future directions

Chapter 2 demonstrated that simple random walks over rich semantic representations can produce behavior consistent with optimal foraging in semantic fluency tasks, providing some interesting directions for future research. Having two competing accounts of the same phenomena suggests that the next step in exploring semantic fluency is designing an experiment that distinguishes between these accounts. One way to do this might be to explore the extent to which human memory search really is strategic – offering people the opportunity to get a “hint” (say, an example category member) might provide the way to do this, as it would be possible to examine whether people seek hints at the moments predicted by the marginal value theorem.

An alternative approach to distinguish these models is considering whether the optimal foraging account can also predict results that the random walk model has previously been used to explain. One such result is the correspondence of word fluency with PageRank (Griffiths, Steyvers, & Firl, 2007) – something that follows directly from the random walk account, but might be more challenging to account for in terms of optimal foraging. Likewise, additional support for optimal foraging in memory has been found which directly measures variables associated with working memory capacity and relates them to features associated with the search process (Hills, Mata, Wilke, & Samanez-Larkin, 2013; Hills & Pachur, 2012). Accordingly, these findings provide future tests for the random walk account.

Exploring some of the nuances of optimal foraging as an account for human memory search

is likely to be a productive direction of future research as well. Human foraging behavior has been examined in a few other domains, including information foraging (Pirolli & Card, 1999) and searching for resources in a simulated spatial environment (Cain, Vul, Clark, & Mitroff, 2012; Hutchinson, Wilke, & Todd, 2008; Kalff, Hills, & Wiener, 2010; Wolfe, 2013). In particular, studies in simulated environments investigate the strategies people use in multiple-target search and examine whether searchers adapt their strategies based on the target distribution statistics. The common finding is that people are in fact sensitive to the resource distributions of their environment, spending more time in resource-dense patches as predicted by optimal foraging theory. However, their actual departure times from these patches tend to be at non-optimal rates (e.g., dependent on patch quality and not the long-term average rate of return as predicted by the marginal value theorem). It would be interesting to see whether modifying the optimal foraging model considered by Hills et al. (2012) to produce behavior more consistent with human search in these other domains would increase or decrease its fit to the data from semantic fluency tasks.

Chapter 3 advances a proposal that reconciles traditionally opposed accounts of color naming across cultures. In this view, people across cultures share a universal representation of color space which constrains what a good color naming system is. We proposed that focal colors, or best examples of color terms, can also be derived from this representation, as representative members of color categories. Given our proposed reconciliation, we should be able to create experimentally-manipulated color categories and force participants to select best examples from them, to further evaluate the predictions of the representativeness model. Furthermore, we can develop experiments to investigate the social influences in color naming, under the iterated learning paradigm (e.g., following J. Xu, Dowman, and Griffiths (2013)). This would provide the opportunity to experimentally test how evolutionary pressures for informative communication interact with general principles of categorization.

The color data examined in Chapter 3 had minimal assumptions for agreement and handling of noise. Future work will explore cleaner versions of the World Color Survey (WCS) data in both individual-speaker and language-level analyses. In addition, a similar dataset to the WCS is currently being digitally transcribed: The Meso-American Color Survey (MCS), currently only partially presented in MacLaury (1997). Investigating how our account generalizes across a different set of languages and color naming data would provide further support that color cognition is constrained by a universally shared perceptual color space, following broader cognitive principles of categorization.

Chapter 4 presented only a limited analysis of the results from Experiment 3, obtained using the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) data. Given the range of conceptual domains for which we have collected human generalization judgments, we can perform a much more detailed investigation of the prior knowledge over different types of conceptual structures that people use when they learn words (e.g. do people prefer shallow or deep taxonomies? what is the variance in basic-level generalization across conceptual domains?). Additionally, we can incorporate how participant behavior is affected by the visual similarity of the images in the training and tests sets (and its interaction with conceptual structure), which would not be possible to explore with the previous Bayesian word learning model.

Developing a system for determining the appropriate referents of a word from labeled images

also has the potential to extend state of the art performance in the field of computer vision. Over the last decade, computer vision researchers have developed algorithms that can classify images and their contents into a large number of categories (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; Krizhevsky, Sutskever, & Hinton, 2012; Sermanet et al., 2014). Despite such success, existing image classification algorithms work in a “yes-or-no” fashion. That is to say, given an image and a category (e.g. “dog”), the classifier predicts if the image belongs to the category or not. The categories could be mutually exclusive (as in early problems such as digit classification), or nested (as in the context of e.g., ImageNet (Deng et al., 2009)), in which case the classifier would predict multiple categories that the image belongs to. However, given a set of categories that are all true for an image or a set of images, existing vision algorithms are not able to further infer what level of the hierarchy is the true underlying concept. Although recent work has proposed using hierarchical structures in object categories (Torralba, Murphy, & Freeman, 2007) or shared attributes (Parikh & Grauman, 2011), learning which objects in an object hierarchy can be referred to by a word (e.g., just Dalmatians or all dog species?) has only recently been explored in computer vision. Jia, Abbott, Austerweil, Griffiths, and Darrell (2013) extended the word learning model in Chapter 4 to include perceptual uncertainty in the recognition of the stimuli, finding the generalization performance of this model dropped substantially from the performance of previous models assuming perfect recognition. However, the perceptual classifier used in Jia et al. (2013) has since been improved dramatically in both accuracy and speed (Jia et al., 2014), offering further opportunities to investigate the role of perceptual features and recognition in word learning.

As word learning is a special case of the more general problem of generalization, our approach potentially could be applied to automatically construct hypothesis spaces for generalization problems in other domains. For example, a Bayesian model of commonsense reasoning could be formulated by automatically deriving hypothesis spaces from ConceptNet (Liu & Singh, 2004) or OpenCyc (Matuszek, Cabral, Witbrock, & DeOliveira, 2006). Each of these resources can be explored as potential representations to support different inferences for different inductive tasks. This follows a development in modern machine learning, which has leveraged online resources to make more successful learning algorithms (Medelyan, Legg, Milne, & Witten, 2009; Ponzetto & Strube, 2006). We hope that this draws a closer connection between computer science and cognitive science, which can lead to more psychologically valid, yet still scalable, artificial intelligence systems.

5.2 Broader theoretical and practical implications

The research in Chapter 2 demonstrates the importance of careful distinction between process and representation: a simple random walk search process over a rich structured representation can produce results consistent with an optimal foraging search over a simpler vector-space representation. The observation that different algorithms operating on different representations can yield similar behavioral predictions echoes previous arguments about the challenges of identifying cognitive representations and processes (Anderson, 1978; Kosslyn & Pomerantz, 1977; Pylyshyn, 1973). One may reasonably question how this effects the remaining dissertation work in Chapters 3

and 4, where I propose specific processes and representations to account for particular behavioral phenomena. Anderson (1978) advises that model parsimony and efficiency may help constrain representations, and using the probabilistic framework provided us the leverage to reveal similar results in Chapter 2. In Chapters 3 and 4 we also use probabilistic models as a framework to evaluate proposals for process-representation pairs, and as in Chapter 2, we argue that future accounts will need to be tested on experiments that distinguish their algorithmic and representational commitments from ours.

The structural analysis of our semantic network in Chapter 2 showed that the categories of animals identified by Troyer et al. (1997) were implicitly reflected in the distances between animal nodes in the network. This relationship provides a potential explanation for why a random walk will exhibit behavior that resembles strategically switching between clusters. However, the semantic network also has a number of other structural attributes that might contribute to this behavior. The recent development of “network science” offers a variety of graph-theoretic properties of networks that can be investigated (Baronchelli et al., 2013). In particular, network science has focused on graphs that form “scale-free networks” – where most of the nodes have few connections but some nodes have many connections – and “small-world networks” – where all nodes are within a few links apart (Barabási & Albert, 1999; Milgram, 1967; Strogatz, 2001; Watts & Strogatz, 1998). These properties have been found to exist in numerous semantic networks built from English language resources, including the network from free associations that we consider (Steyvers & Tenenbaum, 2005). In Appendix A we examine various structural modifications of our semantic network, exploring how degree distributions, edge direction, and connectivity structure in the semantic network effect the observed optimal foraging phenomena. These graph-theoretic structural properties might also uncover interesting properties in semantic networks built from cross-cultural data, which have recently been used to explain universals and variation in conceptual meaning across languages (Borin et al., 2013; Youn et al., 2016).

Informative communication (Corter & Gluck, 1992; Regier et al., 2015) from Chapter 3, used to explain universals and variation in color naming boundaries across cultures, can also provide an account for the *basic-level bias* (Rosch et al., 1976) from Chapter 4, in which people have a bias to generalize new words towards and intermediate taxonomic level. Here, the basic-level is argued to be a natural solution to the problem of efficient communication. However, the analyses which support this proposal are particularly small-scale and artificial. This provides another opportunity to utilize existing online databases, like WordNet, to formally evaluate principles of cognition on a large-scale (Griffiths, 2015). We propose that the structure of WordNet could be used to discover which partitions of a taxonomic subset give rise to basic-level categories. A parallel line of work explores how having multiple meanings for the same word effects word learning. Recently, Dautriche, Chemla, and Christophe (2016) looked at the role of homophony (e.g., “bat”, an animal or baseball item) in word learning and found children used the distribution of exemplars to inform their generalizations and inferences about homophonous words. Srinivasan and Snedeker (2014) on the otherhand explored how the *taxonomic-assumption* (Markman, 1991) constrains learning polysemous words (e.g., “chicken”, the animal or its meat), and found that children are guided by taxonomies, but utilize the lexical structures more for inference. An analogous set of studies could be conducted with WordNet to find the extent in which these results generalize to adults using our

method of large-scale evaluation from Chapter 4.

The Bayesian model of word learning has also been applied to other problems of inferring linguistic meaning, utilizing different hypothesis spaces for learning verb frames (Niyogi, 2002), and syntax acquisition (Chater & Manning, 2006). Besides WordNet, there are other online databases of linguistic representation which could be explored as sources of hypothesis spaces for these different generalization tasks. Whereas WordNet is a popular resource for the hierarchical hyponymy structure of nouns, the Proposition Bank (Palmer, Gildea, & Kingsbury, 2005) is a large resource of verbs and their predicate-arguments, or semantic roles, built on top of syntactic parses from the Penn Treebank (Marcus, Marcinkiewicz, & Santorini, 1993). FrameNet is another online database created by linguists, providing the context of an event, or *semantic frame*, which gives rise to the appropriate conceptual meaning (Baker, Fillmore, & Lowe, 1998). Furthermore, each of these representations has been translated for multiple languages, allowing us to explore potential cross-linguistic effects of semantic organization on generalization (Waxman, Senghas, & Benveniste, 1997).

Probabilistic models of cognition have helped reconcile traditionally opposing accounts of behavioral phenomena (Tenenbaum, 2000), and have uncovered inductive biases which previous accounts could not find (F. Xu & Tenenbaum, 2007a). They have also been used to re-interpret existing computational accounts of cognition such as in formal models of categorization, by viewing categories as probability distributions over collections of objects (Anderson, 1991; Fried & Holyoak, 1984). The traditional models of categorization explored in Chapter 2 and Chapter 3 can be accounted for through this probabilistic lens. For example, prototype models (Reed, 1972) can be interpreted as parametric density estimation (Ashby & Alfonso-Reese, 1995), and exemplar models (Medin & Schaffer, 1978; Nosofsky, 1986) can be explained as non-parametric (kernel) density estimation (Griffiths, Sanborn, Canini, & Navarro, 2008). Furthermore, these popular models of categorization, which have been traditionally approached from an algorithmic level of analysis, have recently been re-formalized as rational approximations to Bayesian inference (Sanborn, Griffiths, & Navarro, 2010; Shi, Griffiths, Feldman, & Sanborn, 2010).

The machine learning community has also benefited from re-interpreting cognitive models to address large-scale inference problems, providing the opportunity for further theoretical and practical advances. For example, a popular machine learning algorithm for *clustering-on-demand* with sparse data, Bayesian Sets (Ghahramani & Heller, 2005; Heller & Ghahramani, 2006), was inspired by the Bayesian model of generalization. Recent work has shown a formal connection between Bayesian Sets and representativeness (from Chapter 3) over sets as well (Abbott, Heller, Ghahramani, & Griffiths, 2011). This opens an interesting set of domain-general principle to explore for future directions: relating the problems of learning categories from a few examples, and finding good examples from categories. This formal connection additionally highlights the practical applications of probabilistic models of cognition (Griffiths, Abbott, & Hsu, 2016). Recent trends in machine learning and computer vision utilizing deep convolutional neural networks provide new opportunities for collaboration with cognitive scientists to solve challenging problems of induction. A particular problem of recent interest involves generating natural language image captions which describe the objects and events in a given scene (Hendricks et al., 2016; Karpathy & Fei-Fei, 2015; Mao et al., 2016). There is a detailed history of similar studies in cognitive

psychology which have explored the kinds of descriptions that people generate for simple and complex image scenes, and have revealed particular biases people use given the semantics of the scene, the relationship between objects, and the different levels of concepts presented (R. Brown, 1958; Kosslyn, 1975; Lupyan, Thompson-Schill, & Swingley, 2010; Mervis & Rosch, 1981; Murphy & Gregory, 2004; Osherson, Smith, Wilkie, Lopez, & Shafir, 1990; Rosch & Mervis, 1975; Rosch et al., 1976; Tanaka & M. Taylor, 1991; K. Taylor, Devereux, Acres, Randall, & Tyler, 2012). Formalizing these inductive biases has the potential to greatly improve the performance and “naturalness” of automated image captioning models.

5.3 Concluding remarks

Probabilistic models of cognition combine statistical evidence from the environment with structured knowledge representations to support the kinds of challenging inductive tasks people seem to solve effortlessly. The case studies considered in this dissertation show how this framework can be used to guide formal psychological investigations. In particular, this work presents alternative accounts, synthesizes traditionally opposed views, and scales up existing models into more powerful explanatory tools. These studies contribute to a richer understanding of human cognition, and to the development of machine learning algorithms that perform more like people.

Bibliography

- Abbott, J., Heller, K., Ghahramani, Z., & Griffiths, T. (2011). Testing a Bayesian measure of representativeness using a large image database. In *Advances in Neural Information Processing Systems* (Vol. 24, pp. 2321–2329).
- Abbott, J., Regier, T., & Griffiths, T. (2012). Predicting focal colors with a rational model of representativeness. In N. Miyake, D. Peebles, & R. Cooper (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society* (pp. 60–65).
- Anderson, J. (1972). Fran: a simulation model of free recall. *Psychology of learning and motivation*, 5, 315–378.
- Anderson, J. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, 85, 249–277.
- Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.
- Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.
- Ashby, F. & Alfonso-Reese, L. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, 39(2), 216–233.
- Austerweil, J. & Griffiths, T. (2010). Learning hypothesis spaces and dimensions through concept learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.
- Austerweil, J. & Griffiths, T. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, 35, 499–526.
- Baker, C., Fillmore, C., & Lowe, J. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics* (Vol. 1, pp. 86–90). Association for Computational Linguistics.
- Barabási, A.-L. & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439), 509–512.
- Baronchelli, A., Ferrer-i-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, 17(7), 348–360.
- Baronchelli, A., Gong, T., Puglisi, A., & Loreto, V. (2010). Modeling the emergence of universality in color naming patterns. *Proceedings of the National Academy of Sciences*, 107, 2403–2407.
- Baronchelli, A., Loreto, V., & Puglisi, A. (2015). Individual biases, cultural evolution, and the statistical nature of language universals. *PLoS ONE*, 10, e0125019.

- Berlin, B. & Kay, P. (1969). *Basic color terms: Their universality and evolution*. University of California Press.
- Borin, L., Comrie, B., & Saxena, A. (2013). The Intercontinental Dictionary Series - A rich and principled database for language comparison. In L. Borin & A. Saxena (Eds.), *Approaches to measuring linguistic differences* (pp. 285–302). Mouton de Gruyter.
- Bousfield, W. A. & Sedgewick, C. H. W. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology*, 30, 149–165.
- Brainard, D. (2003). Color appearance and color difference specification. In S. K. Shevell (Ed.), *The science of color: Second edition* (pp. 191–216). Elsevier.
- Brown, R. (1958). How shall a thing be called? *Psychological Review*, 65(1), 14–21.
- Cain, M. S., Vul, E., Clark, K., & Mitroff, S. R. (2012). A Bayesian optimal foraging model of human visual search. *Psychological Science*, 23(9), 1047–1054.
- Carey, S. (1978). The child as word learner. *Linguistic theory and psychological reality*.
- Carey, S. (2000). The origin of concepts. *Journal of Cognition and Development*, 1(1), 37–41.
- Charnov, E. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, 9(2), 129–136.
- Chater, N. & Manning, C. (2006). Probabilistic models of language processing and acquisition. *Trends in Cognitive Sciences*, 10(7), 335–344.
- Chomsky, N. (1957). *Syntactic structures*. Mouton.
- Collins, A. & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, 82(6), 407–428.
- Cook, R., Kay, P., & Regier, T. (2005). The World Color Survey database: History and use. In H. Cohen & C. Lefebvre (Eds.), *Handbook of categorization in cognitive science* (pp. 223–242). Elsevier.
- Corter, J. & Gluck, M. (1992). Explaining basic categories: feature predictability and information. *Psychological Bulletin*, 111(2), 291–303.
- Dautriche, I., Chemla, E., & Christophe, A. (2016). Word learning: homophony and the distribution of learning exemplars. *Language Learning and Development*, 12(3), 231–251.
- Davidoff, J., Davies, I., & Roberson, D. (1999). Colour categories in a stone-age tribe. *Nature*, 398, 203–204.
- Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: a large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.
- Dougherty, M., Harbison, J., & Davelaar, E. (2014). Optional stopping and the termination of memory retrieval. *Current Directions in Psychological Science*, 23(5), 332–337.
- Dowman, M. (2007). Explaining color term typology with an evolutionary model. *Cognitive Science*, 31, 99–132.
- Duda, R. O. & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.
- Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2010). The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2), 303–338.
- Fellbaum, C. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*, 231–243.

- Fried, L. & Holyoak, K. (1984). Induction of category distributions: a framework for classification learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 10(2), 234–257.
- Gelman, A., Carlin, J., Stern, H., & Rubin, D. (2004). *Bayesian data analysis: Second edition*. Chapman & Hall / CRC Press.
- Ghahramani, Z. & Heller, K. A. (2005). Bayesian sets. *Advances in Neural Information Processing Systems*, 18, 435–442.
- Gigerenzer, G. & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, 102, 684–704.
- Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N. V., Corominas-Murtra, B., ... Villoslada, P. (2011). The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive Processing*, 12(2), 183–196.
- Goñi, J., Martincorena, I., Corominas-Murtra, B., Arrondo, G., Ardanza-Trevijano, S., & Villoslada, P. (2010). Switcher-random-walks: a cognitive-inspired mechanism for network exploration. *International Journal of Bifurcation and Chaos*, 20(3), 913–922.
- Goodman, N., Tenenbaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, 32(1), 108–154.
- Grice, H., Cole, P., & Morgan, J. (1975). Syntax and semantics. *Logic and Conversation*, 3, 41–58.
- Griffin, L. (2006). Optimality of the basic colour categories for classification. *Journal of the Royal Society: Interface*, 3, 71–85.
- Griffiths, T. (2015). Manifesto for a new (computational) cognitive revolution. *Cognition*, 135, 21–23.
- Griffiths, T., Abbott, J., & Hsu, A. (2016). Exploring human cognition using large image databases. *Topics in Cognitive Science*, 8(3), 569–588.
- Griffiths, T., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8), 357–364.
- Griffiths, T., Kemp, C., & Tenenbaum, J. (2008). Bayesian models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational psychology* (pp. 59–100). Chapter 3. Cambridge University Press.
- Griffiths, T., Sanborn, A., Canini, K., & Navarro, D. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind*. Oxford: Oxford University Press.
- Griffiths, T., Steyvers, M., & Firl, A. (2007). Google and the mind. *Psychological Science*, 18(12), 1069–1076.
- Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, 114, 211–244.
- Griffiths, T. & Tenenbaum, J. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.
- Hafner, J., Sawhney, H. S., Equitz, W., Flickner, M., & Niblack, W. (1995). Efficient color histogram indexing for quadratic form distance functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(7), 729–736.

- Heider, E. (1972a). Probabilities, sampling, and ethnographic method: The case of Dani colour names. *Man, New Series*, 7, 448–466.
- Heider, E. (1972b). Universals in color naming and memory. *Journal of Experimental Psychology*, 93, 10–20.
- Heller, K. A. & Ghahramani, Z. (2006). A simple Bayesian framework for content-based image retrieval. *IEEE Conference on Computer Vision and Pattern Recognition*, 2, 2110–2117.
- Hendricks, L. A., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., & Darrell, T. (2016). Deep compositional captioning: describing novel object categories without paired training data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hering, E. (1964). *Outlines of a theory of the light sense*. Translated by L.M. Hurvich and D. Jameson. Harvard University Press.
- Hills, T., Jones, M., & Todd, P. (2012). Optimal foraging in semantic memory. *Psychological Review*, 119(2), 431–440.
- Hills, T., Mata, R., Wilke, A., & Samanez-Larkin, G. R. (2013). Mechanisms of age-related decline in memory search across the adult life span. *Developmental psychology*, 49(12), 2396.
- Hills, T. & Pachur, T. (2012). Dynamic search and working memory in social recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(1), 218–228.
- Hutchinson, J., Wilke, A., & Todd, P. (2008). Patch leaving in humans: can a generalist adapt its rules to dispersal of items across patches? *Animal Behaviour*, 75(4), 1331–1349.
- Huttenlocher, D., Klanderman, G., & Rucklidge, W. (1993). Comparing images using the Hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15, 850–863.
- Jameson, K. (2005a). Culture and cognition: what is universal about the representation of color experience? *Journal of Cognition and Culture*, 5, 293–348.
- Jameson, K. (2005b). Why GRUE? An interpoint-distance model analysis of composite color categories. *Cross-Cultural Research*, 39, 159–204.
- Jameson, K. (2010). Where in the World Color Survey is the support for the Hering primaries as the basis for color categorization? In J. Cohen & M. Matthen (Eds.), *Color ontology and color science* (pp. 179–202). MIT Press.
- Jameson, K. & D’Andrade, R. (1997). It’s not really red, green, yellow, blue: An inquiry into perceptual color space. In C. L. Hardin & L. Maffi (Eds.), *Color categories in thought and language* (pp. 295–319). Cambridge University Press.
- Jia, Y., Abbott, J., Austerweil, J., Griffiths, T., & Darrell, T. (2013). Visual concept learning: combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems 26* (pp. 1842–1850).
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 675–678). ACM.
- Jones, M. N. & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, 114(1), 1–37.
- Kahneman, D. & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.

- Kalff, C., Hills, T., & Wiener, J. (2010). Human foraging behavior: a virtual reality investigation on area restricted search in humans. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 168–173).
- Karpathy, A. & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3128–3137).
- Kay, P. & McDaniel, C. (1978). The linguistic significance of the meanings of basic color terms. *Language, 54*, 610–646.
- Kay, P. & Regier, T. (2003). Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences, 100*, 9085–9089.
- Kay, P. & Regier, T. (2007). Color naming universals: the case of Berinmo. *Cognition, 102*, 289–298.
- Kemp, C. & Tenenbaum, J. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687–10692.
- Kemp, C. & Tenenbaum, J. (2009). Structured statistical models of inductive reasoning. *Psychological Review, 116*(1), 20–58.
- Komarova, N. & Jameson, K. (2013). A quantitative theory of human color choices. *PLoS ONE, 8*, e55986.
- Komarova, N., Jameson, K., & Narens, L. (2007). Evolutionary models of color categorization based on discrimination. *Journal of Mathematical Psychology, 51*, 359–382.
- Kosslyn, S. (1975). Information representation in visual images. *Cognitive Psychology, 7*(3), 341–370.
- Kosslyn, S. (1994). *Image and brain: the resolution of the imagery debate*. The MIT Press.
- Kosslyn, S. & Pomerantz, J. (1977). Imagery, propositions, and the form of internal representations. *Cognitive Psychology, 9*(1), 52–76.
- Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25* (pp. 1097–1105).
- Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder-Mead simplex method in low dimensions. *SIAM Journal of Optimization, 9*, 112–147.
- Lezak, M. (1995). *Neuropsychological assessment*. Oxford University Press, USA.
- Lindsey, D. & Brown, A. (2006). Universality of color names. *Proceedings of the National Academy of Sciences, 103*, 16608–16613.
- Lindsey, D. & Brown, A. (2009). World Color Survey color naming reveals universal motifs and their within-language diversity. *Proceedings of the National Academy of Sciences, 106*, 19785–19790.
- Liu, H. & Singh, P. (2004). ConceptNet - A practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211–226.
- Lupyan, G., Thompson-Schill, S., & Swingley, D. (2010). Conceptual penetration of visual processing. *Psychological Science*.

- MacLaury, R. (1997). *Color and cognition in mesoamerica: constructing categories as vantages*. University of Texas Press.
- Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Marcus, M., Marcinkiewicz, M., & Santorini, B. (1993). Building a large annotated corpus of english: the Penn Treebank. *Computational Linguistics*, 19(2), 313–330.
- Markman, E. M. (1991). *Categorization and naming in children: problems of induction*. MIT Press.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In C. Baral (Ed.), *Proceedings of the AAAI 2006 Spring Symposium* (pp. 44–49). Menlo Park, CA: AAAI Press.
- McClelland, J., Botvinick, M., Noelle, D., Plaut, D., Rogers, T., Seidenberg, M., & Smith, L. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, 14(8), 348–356.
- Medelyan, O., Legg, C., Milne, D., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9), 1–76.
- Medin, D. L. & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, 85, 207–238.
- Mervis, C. & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology*, 32, 89–115.
- Milgram, S. (1967). The small world problem. *Psychology Today*, 2(1), 60–67.
- Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review*, 63(2), 81–97.
- Miller, G. (1995). WordNet: a lexical database for english. *Communications of the ACM*, 38(11), 39–41.
- Murphy, G. & Gregory, L. (2004). *The big book of concepts*. MIT press.
- Navarro, D., Dry, M., & Lee, M. (2012). Sampling assumptions in inductive generalization. *Cognitive Science*, 32, 187–223.
- Navarro, D. & Griffiths, T. (2008). Latent features in similarity judgments: a nonparametric Bayesian approach. *Neural Computation*, 20, 2597–2628.
- Nelson, D., McEvoy, C., & Schreiber, T. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3), 402–407.
- Newell, A., Shaw, J., & Simon, H. (1958). Elements of a theory of human problem solving. *Psychological Review*, 65(3), 151–166.
- Niyogi, S. (2002). Bayesian learning at the syntax-semantics interface. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (Vol. 36, pp. 697–702). Fairfax.
- Nosofsky, R. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1), 39–57.
- Nosofsky, R. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4), 700–708.

- Oaksford, M. & Chater, N. (Eds.). (1998). *Rational models of cognition*. Oxford: Oxford University Press.
- Osherson, D., Smith, E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, 97, 185.
- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999, November). *The PageRank citation ranking: bringing order to the web*. (Technical Report No. 1999-66). Stanford InfoLab. Stanford InfoLab.
- Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: an annotated corpus of semantic roles. *Computational Linguistics*, 31(1), 71–106.
- Parikh, D. & Grauman, K. (2011). Relative attributes. In *Proceedings of the International Conference on Computer Vision (ICCV)*.
- Pirolli, P. & Card, S. (1999). Information foraging. *Psychological Review*, 106(4), 643–675.
- Ponzetto, S. P. & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the HLT Conference of the NAACL* (pp. 192–199).
- Puglisi, A., Baronchelli, A., & Loreto, V. (2008). Cultural route to the emergence of linguistic categories. *Proceedings of the National Academy of Sciences*, 105, 7936–7940.
- Pylyshyn, Z. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, 80(1), 1–24.
- Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, UK: Cambridge University Press.
- Quine, W. V. O. (1975). *Word and object*. MIT Press.
- Raaijmakers, J. & Shiffrin, R. (1981). Search of associative memory. *Psychological Review*, 88(2), 93–134.
- Reed, S. (1972). Pattern recognition and categorization. *Cognitive Psychology*, 3(3), 382–407.
- Regier, T., Kay, P., & Cook, R. (2005). Focal colors are universal after all. *Proceedings of the National Academy of Sciences*, 102, 8386–8391.
- Regier, T., Kay, P., & Khetarpal, N. (2007). Color naming reflects optimal partitions of color space. *Proceedings of the National Academy of Sciences*, 104, 1436–1441.
- Regier, T., Kay, P., & Khetarpal, N. (2009). Color naming and the shape of color space. *Language*, 85, 884–892.
- Regier, T., Kemp, C., & Kay, P. (2015). Word meanings across languages support efficient communication. In B. MacWhinney & W. O'Grady (Eds.), *The handbook of language emergence* (pp. 237–263). Wiley-Blackwell.
- Rhodes, T. & Turvey, M. T. (2007). Human memory retrieval as Lévy foraging. *Physica A*, 385, 255–260.
- Roberson, D., Davidoff, J., Davies, I. R., & Shapiro, L. R. (2005). Color categories: Evidence for the cultural relativity hypothesis. *Cognitive Psychology*, 50, 378–411.
- Roberson, D., Davies, I. R., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, 129, 369–398.
- Rogers, T. & McClelland, J. (2004). *Semantic cognition: a parallel distributed processing approach*. MIT press.

- Romney, A. K., Brewer, D. D., & Batchelder, W. H. (1993). Predicting clustering from semantic structure. *Psychological Science*, *4*(1), 28–34.
- Rosch, E. & Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*, 573–605.
- Rosch, E., Mervis, C., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., . . . Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, *115*(3), 211–252. doi:10.1007/s11263-015-0816-y
- Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.
- Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). Overfeat: integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR)*.
- Shepard, R. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.
- Shepard, R. & Arabie, P. (1979). Additive clustering: representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.
- Shi, L., Griffiths, T., Feldman, N., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin and Review*, *17*, 443–464.
- Spelke, E. & Newport, E. (1998). Nativism, empiricism, and the development of knowledge. In R. Lerner (Ed.), *Handbook of child psychology, 5th ed., vol. 1: Theoretical models of human development*. NY: Wiley.
- Srinivasan, M. & Snedeker, J. (2014). Polysemy and the taxonomic constraint: children’s representation of words that label multiple kinds. *Language Learning and Development*, *10*(2), 97–128.
- Steels, L. & Belpaeme, T. (2005). Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, *28*, 469–489.
- Stephens, D. & Krebs, J. (1986). *Foraging theory*. Princeton University Press.
- Steyvers, M., Shiffrin, R., & Nelson, D. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, 237–249.
- Steyvers, M. & Tenenbaum, J. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.
- Strogatz, S. H. (2001). Exploring complex networks. *Nature*, *410*(6825), 268–276.
- Tanaka, J. & Taylor, M. (1991). Object categories and expertise: is the basic level in the eye of the beholder? *Cognitive Psychology*, *23*(3), 457–482.
- Taylor, K., Devereux, B., Acres, K., Randall, B., & Tyler, L. (2012). Contrasting effects of feature-based statistics on the categorisation and basic-level identification of visual objects. *Cognition*, *122*(3), 363–374.

- Tenenbaum, J. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (Vol. 11, pp. 59–65). Cambridge, MA: MIT Press.
- Tenenbaum, J. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12* (Vol. 12, pp. 59–65).
- Tenenbaum, J. & Griffiths, T. (2001a). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, 24(4), 629–640.
- Tenenbaum, J. & Griffiths, T. (2001b). The rational basis of representativeness. In *Proceedings of the 23th Annual Conference of the Cognitive Science Society* (pp. 1036–1041).
- Tenenbaum, J., Griffiths, T., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, 10(7), 309–318.
- Tenenbaum, J., Kemp, C., Griffiths, T., & Goodman, N. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, 331(6022), 1279–1285.
- Thompson, G. W., Kello, C. T., & Montez, P. (2013). Searching semantic memory as a scale-free network: evidence from category recall and a wikipedia model of semantics. In *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.
- Thurstone, L. (1938). Primary mental abilities. *Psychometric Monographs*.
- Torralba, A., Murphy, K., & Freeman, W. (2007). Sharing visual features for multiclass and multi-view object detection. *IEEE TPAMI*, 29(5), 854–869.
- Tröster, A., Salmon, D., McCullough, D., & Butters, N. (1989). A comparison of the category fluency deficits associated with Alzheimer’s and Huntington’s disease. *Brain and Language*, 37(3), 500–513.
- Troyer, A. K., Moscovitch, M., & Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, 11(1), 138–146.
- Troyer, A. K., Moscovitch, M., Winocur, G., Leach, L., & Freedman, M. (1998). Clustering and switching on verbal fluency tests in Alzheimer’s and Parkinson’s disease. *Journal of the International Neuropsychological Society*, 4(2), 137–143.
- Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84, 327–352.
- Tversky, A. & Hutchinson, J. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, 93(1), 3–22.
- Watts, D. J. & Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684), 440–442.
- Waxman, S. (1990). Linguistic biases and the establishment of conceptual hierarchies: evidence from preschool children. *Cognitive Development*, 5(2), 123–150.
- Waxman, S. & Markow, D. (1995). Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cognitive Psychology*, 29(3), 257–302.
- Waxman, S., Senghas, A., & Benveniste, S. (1997). A cross-linguistic examination of the noun-category bias: its existence and specificity in french-and spanish-speaking preschool-aged children. *Cognitive Psychology*, 32(3), 183–218.

- Witzel, C. & Franklin, A. (2014). Do focal colors look particularly “colorful”? *Journal of the Optical Society of America, A*, 31, A365–A374.
- Wolfe, J. M. (2013). When is it time to move to the next raspberry bush? foraging rules in human visual search. *Journal of Vision*, 13(3), 1–17.
- Xu, F. & Tenenbaum, J. (2007a). Sensitivity to sampling in Bayesian word learning. *Developmental Science*, 10(3), 288–297.
- Xu, F. & Tenenbaum, J. (2007b). Word learning as Bayesian inference. *Psychological Review*, 114(2), 245–272.
- Xu, J., Dowman, M., & Griffiths, T. (2013). Cultural transmission results in convergence towards colour term universals. *Proceedings of the Royal Society of London B: Biological Sciences*, 280(1758), 20123073.
- Yendrikhovskij, S. (2001). Computing color categories from statistics of natural images. *Journal of Imaging Science and Technology*, 45, 409–417.
- Youn, H., Sutton, L., Smith, E., Moore, C., Wilkins, J., Maddieson, I., . . . Bhattacharya, T. (2016). On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7), 1766–1771.

Appendix A

Supplementary analyses of semantic memory search

This supplementary material presents the results for simulations not included in Chapter 2 of the main text. These additional simulations explore the effects of our random walk model on various structural modifications of the semantic network we operate upon. In particular, we explore how degree distributions, edge direction, and connectivity structure in our semantic network effect the observed optimal foraging phenomena reported in the chapter.

A.1 Effects of degree distributions and edge direction

Our analysis of the structure of our semantic network in Chapter 2 found that the categories of animals identified by Troyer et al. (1997) were implicitly reflected in the distances between animal nodes in the network. This relationship provides a potential explanation for why a random walk will exhibit behavior that resembles strategically switching between clusters. However, the semantic network also has a variety of other properties that might contribute to this behavior.

In this section, we consider the effects that the *degree distribution* of the semantic network has on whether a random walk produces predictions consistent with optimal foraging. The degree of a node refers to the number of connections (or neighbors) it has, which can be differentiated into *out-degree* and *in-degree* for directed graphs, corresponding to the number of outward and inward edges respectively. The degree distribution indicates the probability that a random node will have a particular number of connections. Recent findings indicate that many real-world networks follow a power-law degree distribution, with “heavy tails” that result in a small number of nodes having a very large degree (Barabási & Albert, 1999; Strogatz, 2001). Since power-law distributions have no characteristic scale of node degree, networks exhibiting this property are referred to as “scale-free”. Steyvers and Tenenbaum (2005) examined the degree distribution of the semantic network derived from word associations that we have used in our analyses. This examination found that an undirected version of the network (where the direction of the edges was removed) had a power-law degree distribution, whereas the the in-degree distribution for the directed network was near

power-law but the out-degree distribution was not (while there is some variation, all nodes have a relatively small out-degree).

We now explore whether these different degree distributions play a role in producing predictions consistent with optimal foraging. By modifying the properties of the graph on which we conduct the random walk, we can examine whether maintaining directional information is important, and whether having a power-law distribution in in-degree or out-degree is important to producing the appropriate behavioral results.

Methods

We test two structural variations of our semantic network: a version with undirected edges, and a version with the edge directions reversed. Recall, in the directed semantic network derived from word-association data, a directed edge from word node j connects to word node i if the word i was given as a response to cue j (corresponding to the *link matrix* \mathbf{L}_{ij} from above). In the undirected network, two word nodes are connected by an edge if they were related associatively, regardless of association direction (ie. if a link existed from i to j , or from j to i). In the reverse-directed network, a directed edge from word node i connects to word node j if the word i was given as a response to cue j (corresponding to the transpose of the link matrix).

We explored the effects of these structural changes using the uniform-transition model, as the frequency of transitions in the word-association data account for only one direction of association. We ran 1000 simulations of the uniform non-jumping and uniform jumping models for a duration of 1750 iterations on both the reverse-directed and undirected networks. As before, the jumping models had a probability of $\rho = 0.05$ of making a jump back to “animals”, selected primarily to illustrate the impact of adding this additional component to the search process. Other small values of ρ produced the same qualitative results.

Results and Discussion

Reverse-directed network. Each of the models was analyzed in the same manner as our random walk simulation in the main text. The results for the uniform non-jumping and uniform jumping models on the reverse-directed network are presented in Figure A.1, where the top row displays the results of the uniform non-jumping model and the bottom row displays the results of the uniform jumping model. The left column shows the mean ratio between the IRT for an item and the mean IRT over all 1750 iterations in the simulations, relative to the order of entry for the item. As before, we see that the first word starting a cluster has the highest overall retrieval time for both networks ($t(999) = 31.62, p < 0.001$ and $t(999) = 85.05, p < 0.001$ for the uniform non-jumping and uniform jumping models respectively), and the second word in a cluster takes much less time to produce than the long-term mean ($t(999) = -28.63, p < 0.001$ and $t(999) = -73.40, p < 0.001$, respectively). In addition, the IRTs for words preceding a cluster switch (indicated by “-1”) did not differ significantly from most walkers’ long-term average IRTs (918 and 981 out of 1000 walkers were not significantly different for the uniform non-jumping and uniform jumping models

respectively, and all of the walkers that were significantly different had pre-switch IRT averages less than their long-term averages for each of the networks).

The right column of Figure A.1 examines consistency with the cluster-leaving policy indicated by the marginal value theorem, where again we find that walkers with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words (a linear regression model found a significant negative relationship between axes for both networks: slope = -0.70, $t(998) = 9.03$, $p < 0.001$ and slope = -1.42, $t(998) = 4.49$, $p < 0.001$ for the uniform non-jumping and uniform jumping models respectively).

Undirected network. The results for the uniform non-jumping and uniform jumping models on the undirected network are presented in Figure A.2, where the top row displays the results of the uniform non-jumping model and the bottom row displays the results of the uniform jumping model. The left column shows the mean ratio between the IRT for an item and the mean IRT over all 1750 iterations in the simulations, relative to the order of entry for the item. As before, we see that the first word starting a cluster has the highest overall retrieval time for both networks ($t(999) = 44.56$, $p < 0.001$ and $t(999) = 60.21$, $p < 0.001$ for the uniform non-jumping and uniform jumping models respectively), and the second word in a cluster takes much less time to produce than the long-term mean ($t(999) = -34.94$, $p < 0.001$ and $t(999) = -52.24$, $p < 0.001$, respectively). In addition, the IRTs for words preceding a cluster switch (indicated by “-1”) did not differ significantly from most walkers’ long-term average IRTs (986 and 992 out of 1000 walkers were not significantly different for the reverse-directed and undirected networks respectively, and all of the walkers that were significantly different had pre-switch IRT averages less than their long-term averages for each of the networks).

The right column of Figure A.2 examines consistency with the cluster-leaving policy indicated by the marginal value theorem, where again we find that walkers with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words (a linear regression model found a significant negative relationship between axes for both networks: slope = -0.54, $t(998) = 9.05$, $p < 0.001$ and slope = -1.19, $t(998) = 4.31$, $p < 0.001$ for the uniform non-jumping and uniform jumping models respectively).

We thus observe similar phenomena on these semantic networks when the edges are made to be undirected or when they are reversed. Making the edges to be undirected or reversing the edges results in a power-law degree distribution for both in-degree and out-degree, or power-law degree distribution for the out-degree rather than in-degree distribution, respectively. Because random walks on these transformed semantic networks produce similar behavior to random walks on the original semantic network, degree distribution does not have a strong effect on the foraging behavior produced by our random walker models.

A.2 Effects of connectivity structure

The simulations above show that the degree distribution of the semantic network has little effect on whether a random walk produces behavior consistent with optimal foraging. In this section,

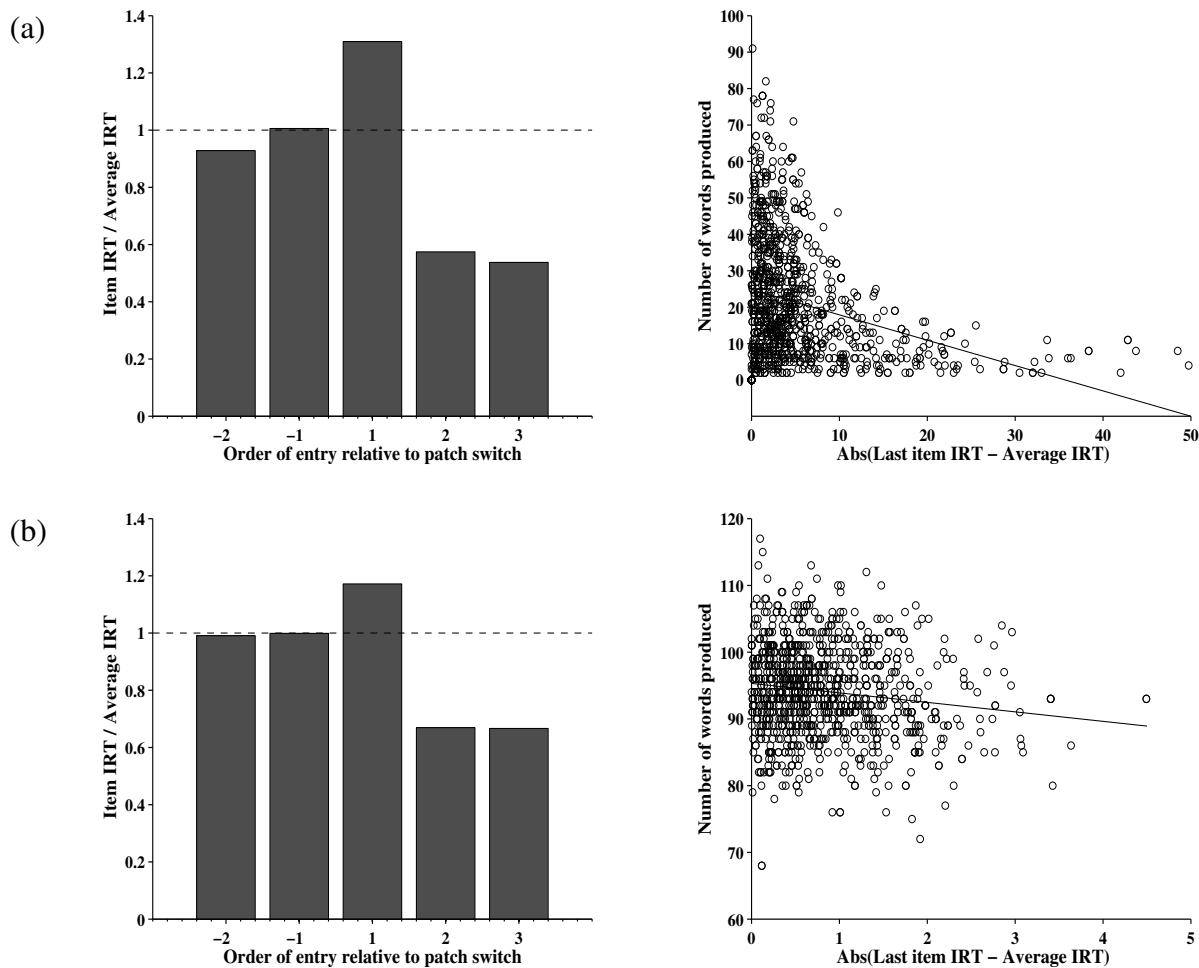


Figure A.1: Results for the uniform non-jumping model and uniform jumping model on the reverse-directed network. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced.

we explore how the connectivity structure of our network may affect these results. A common finding in real-world networks is that most nodes can be reached from any other node within a short number of traversed edges – similar to the idea that most people can be connected by a sequence of six friends or associates (Milgram, 1967; Watts & Strogatz, 1998). Networks with this property are called “small-world” networks. Steyvers and Tenenbaum (2005) examined the

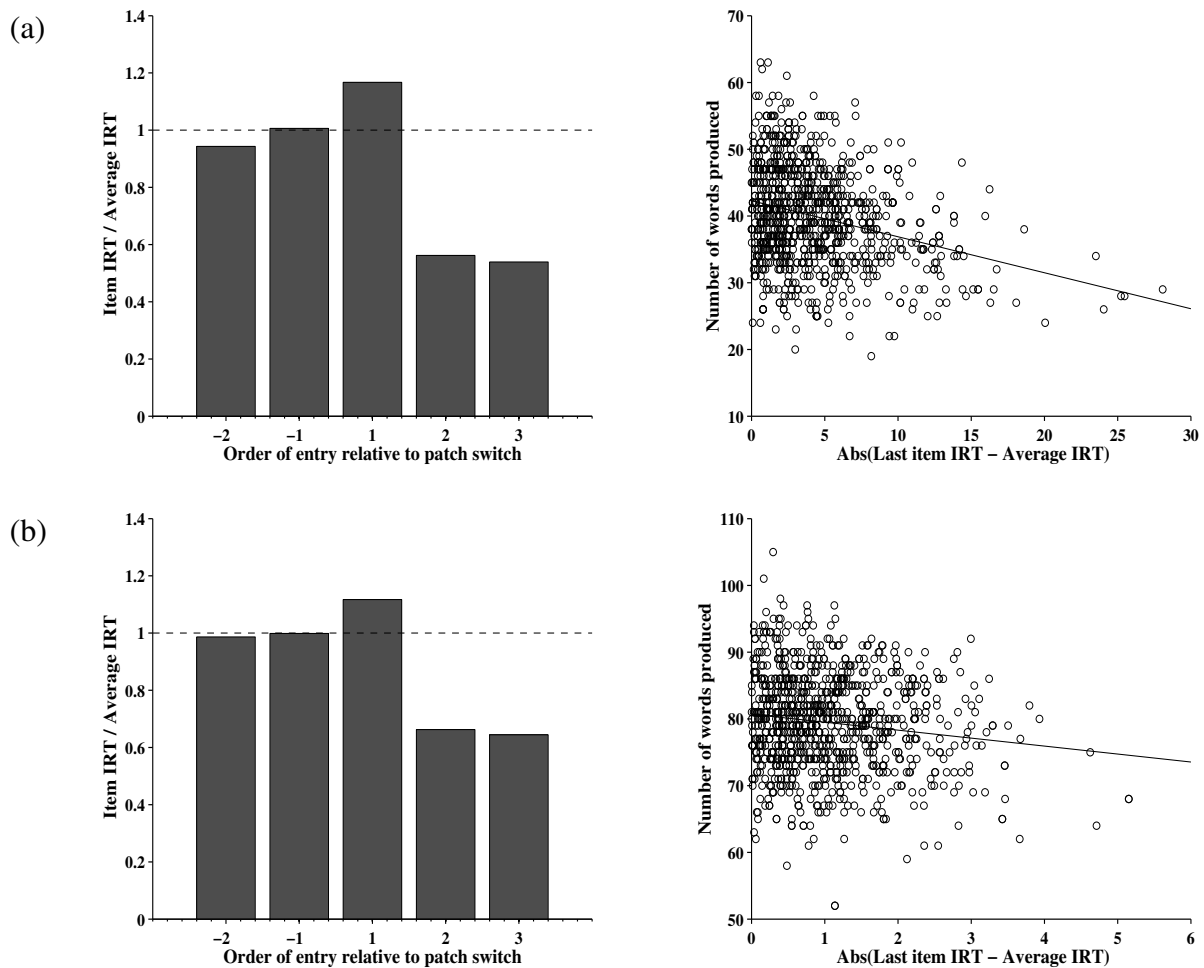


Figure A.2: Results for (a) the uniform non-jumping model and (b) uniform jumping model on the undirected network. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem policy for cluster departures (horizontal-axis) and the total number of words a walker produced.

average shortest-path lengths between nodes in the semantic network used in our analyses and found results that were consistent with other small-world networks of similar size.

In the following simulation we explored whether the small-world property is sufficient to produce predictions consistent with optimal foraging theory. By randomly relabeling the nodes in the semantic network, we can change which words are connected to one another without changing the

small-world structure. This random relabeling also allows us to examine whether the property of the semantic network that motivated our analyses – words from the same category tending to be close together – is necessary in order for a random walk to behave similarly to optimal foraging.

Methods

To determine the effects of the small-world structure in our semantic network, we created a new network with the same connections between nodes, but with a random relabeling of the words that correspond to those nodes. This maintains the small-world structure, while disrupting the relationship between edges and semantic relatedness. As before, we explored the effects of this change using the uniform-transition model. We ran 1000 simulations of the uniform non-jumping and uniform jumping models for a duration of 1750 iterations on the random-labeled network.

Results and Discussion

Each of the models was analyzed in the same manner as those in Simulation 1. The uniform non-jumping and uniform jumping model results are presented in Figure A.3 as panels (a) and (b), respectively. The left column shows the mean ratio between the IRT for an item and the mean IRT over all 1750 iterations in the simulations, relative to the order of entry for the item. Here we see that all words take relatively the same amount of time to produce, regardless of their order in a cluster (none of the bars is significantly different from the dotted line). These results violate the predictions of the marginal value theorem. However, the right column of Figure A.3 shows that walkers with a larger absolute difference (indicating they either left clusters too soon or too late) produced fewer words (a linear regression model found a significant negative relationship between axes: slope = -0.55, $t(998) = 6.26$, $p < 0.001$ and slope = -0.23, $t(998) = 5.42$, $p < 0.001$ for the uniform non-jumping and uniform jumping models, respectively). These results are consistent with the predictions of the marginal value theorem.

The results taken together indicate that having a small-world structure is not sufficient to produce results consistent with the marginal value theorem. Edges need to reflect semantic relatedness – and words in the same semantic categories need to be close to one another – in order for the IRT differences between cluster switches predicted by optimal foraging theory to be produced. The results also suggest that the relationship between absolute difference in cluster leaving times and number of words produced by a participant is a weaker indicator of whether the agent is guided by an optimal foraging policy, because a random walk on the semantic network where the node labels are shuffled produced an appropriate pattern with respect to this metric while failing to do so for the IRT.

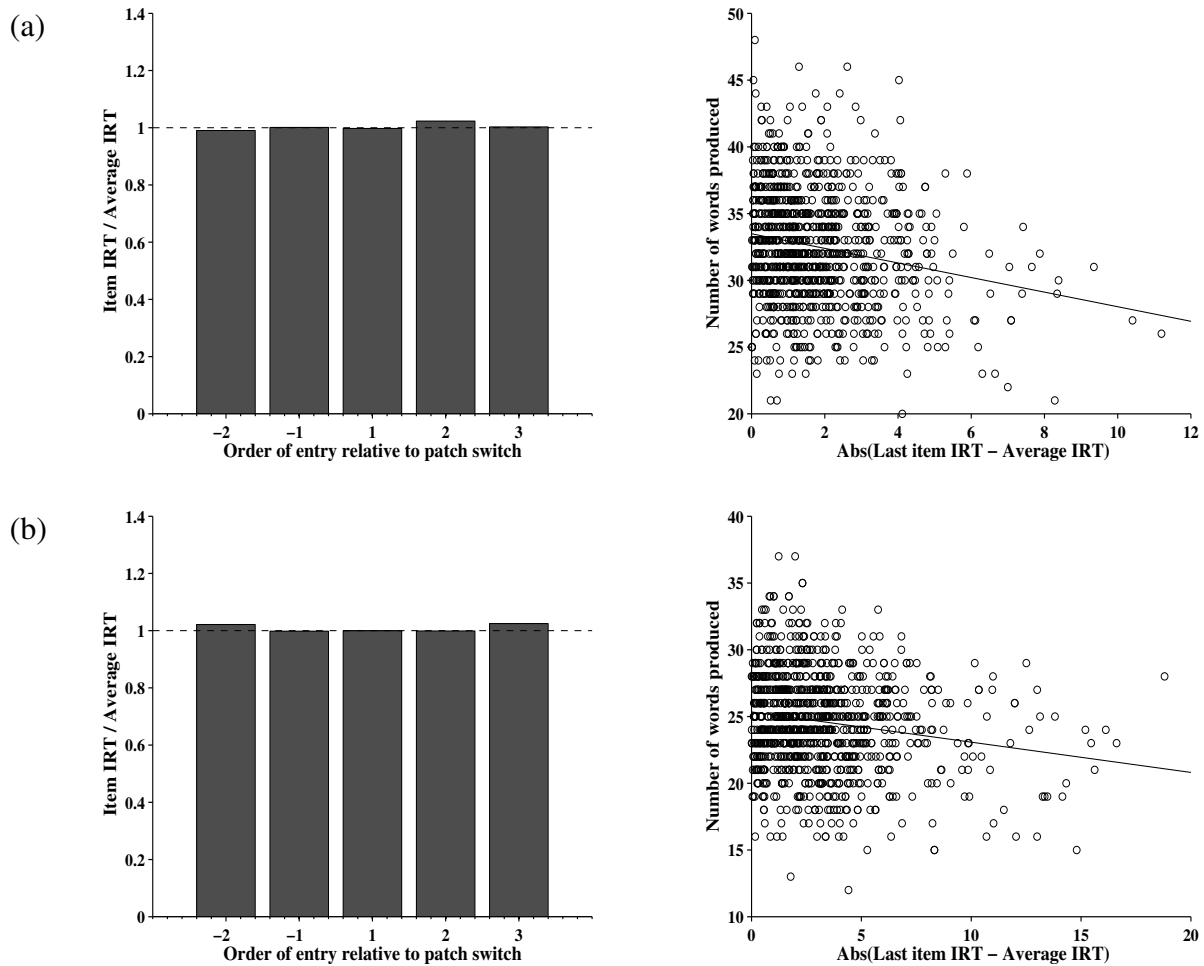


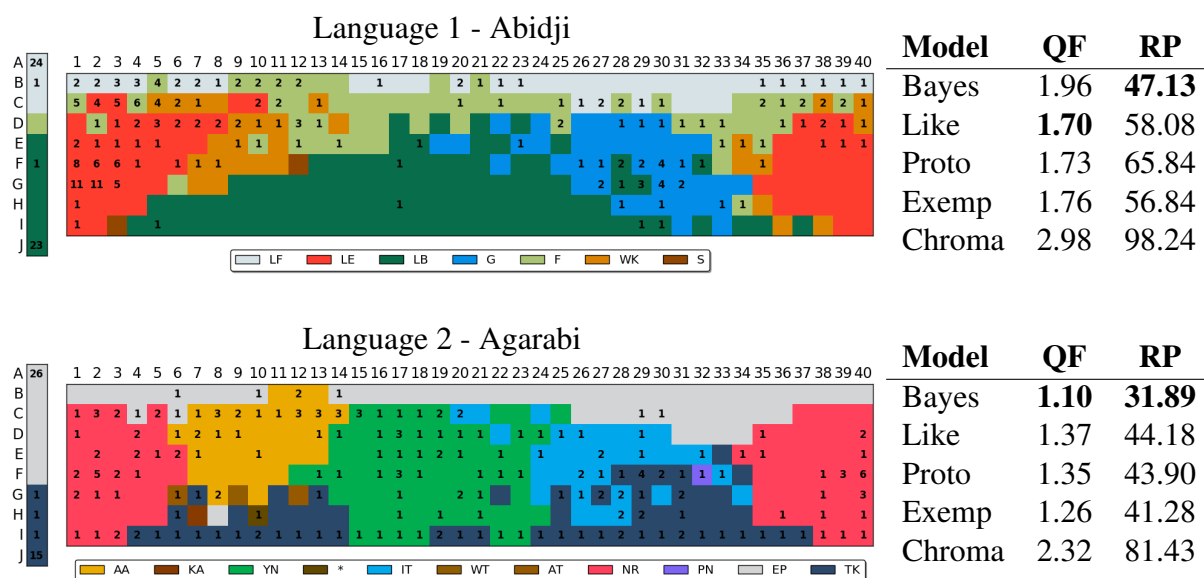
Figure A.3: Results for (a) the uniform non-jumping model and (b) uniform jumping model on our semantic network with random node relabelling. The left column displays the mean ratio between the inter-item response time (IRT) for an item and the walker’s long-term average IRT over the entire task, relative to the order of entry for the item (where “1” refers to the relative IRT between the first word in a cluster and the last word in the preceding cluster). The dotted line indicates where item IRTs would be the same as the walker’s average IRT for the entire task. The right column displays the relationship between a walker’s deviation from the marginal value theorem for cluster departures (horizontal-axis) and the total number of words a walker produced.

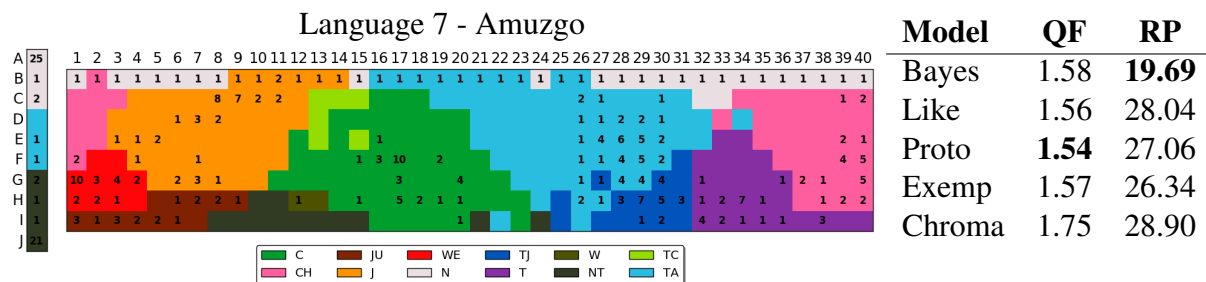
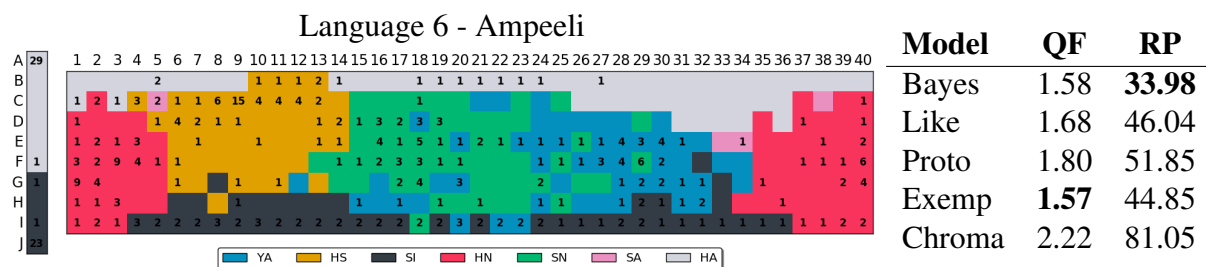
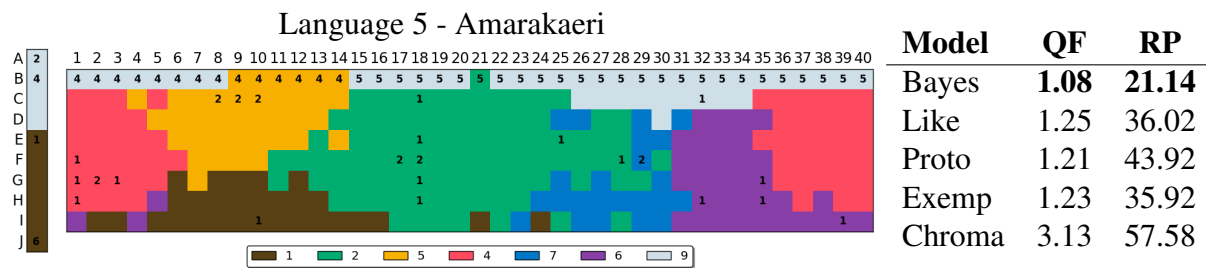
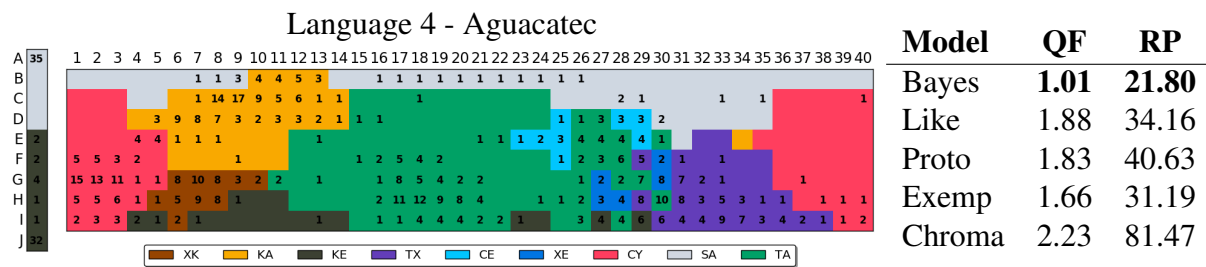
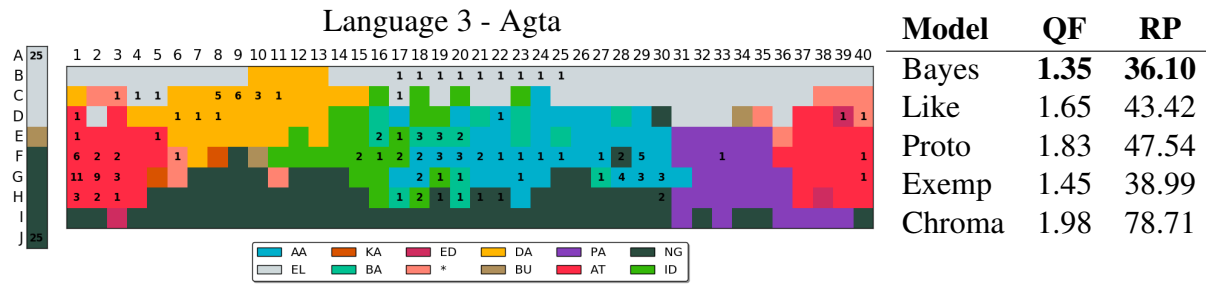
Appendix B

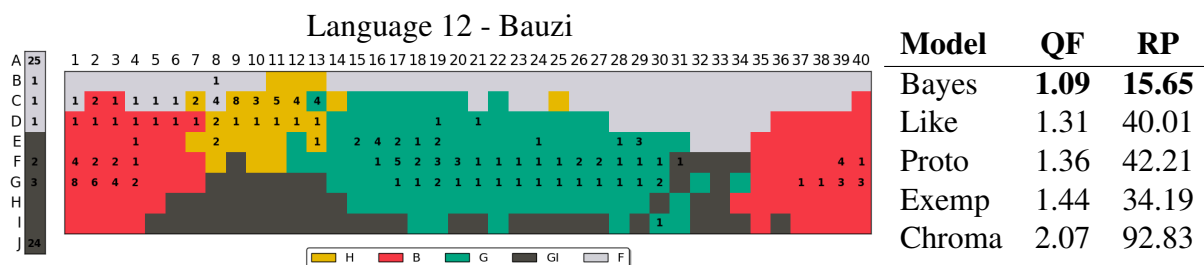
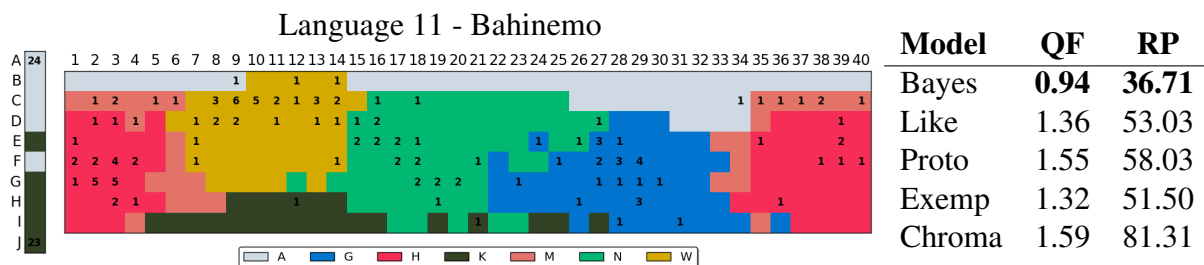
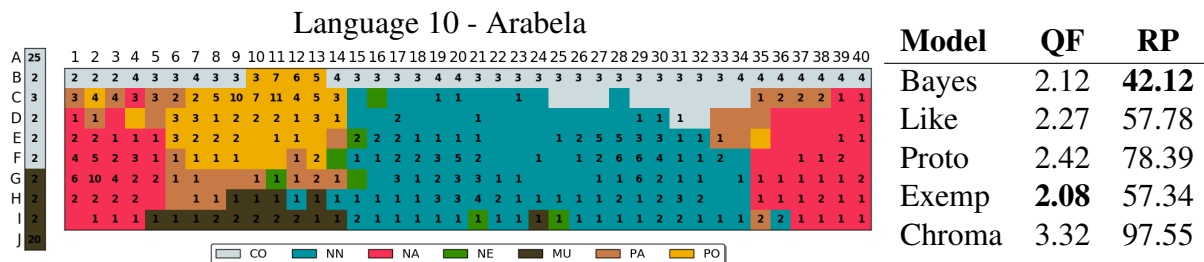
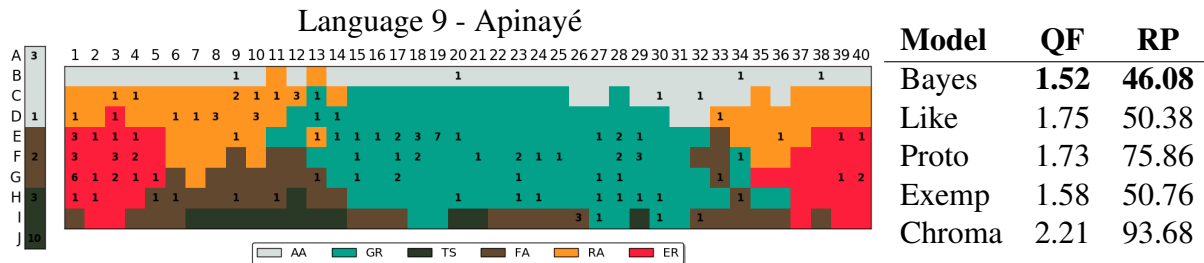
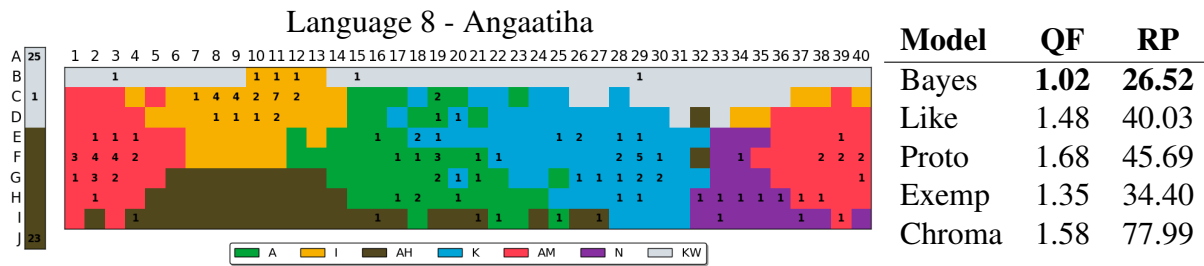
Supplementary analyses of the World Color Survey

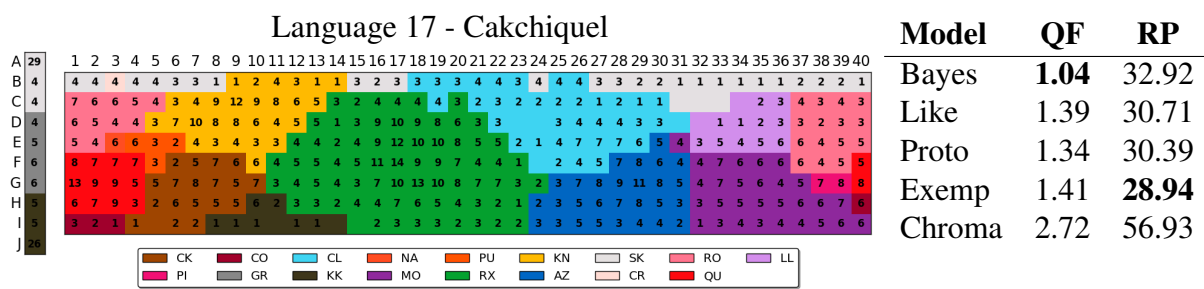
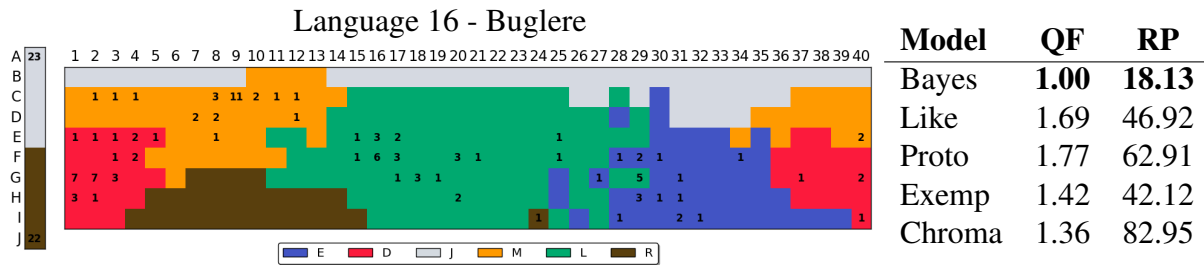
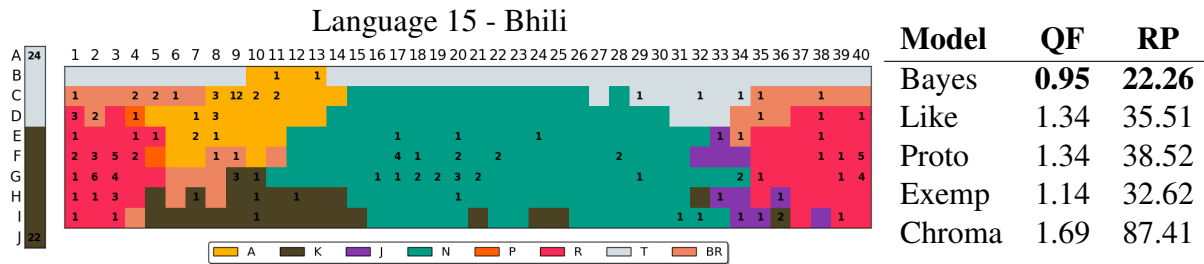
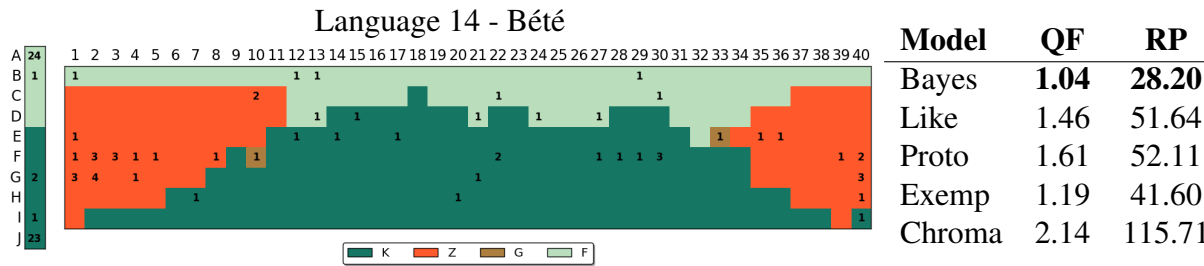
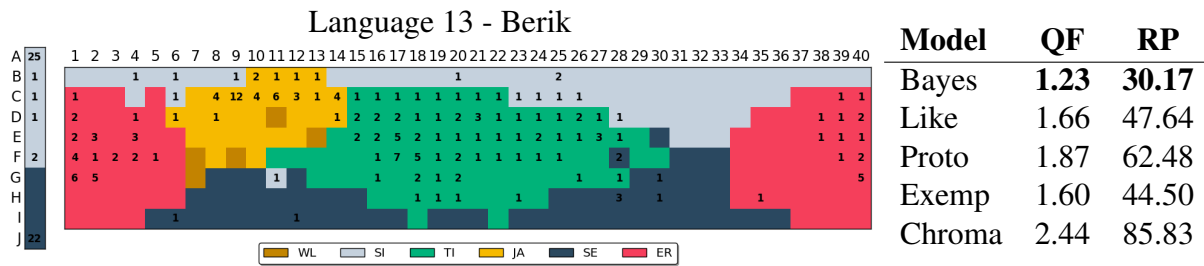
B.1 Language-level analyses

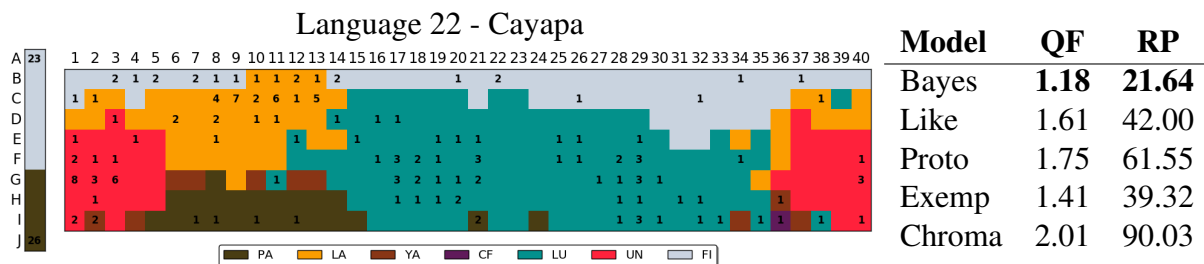
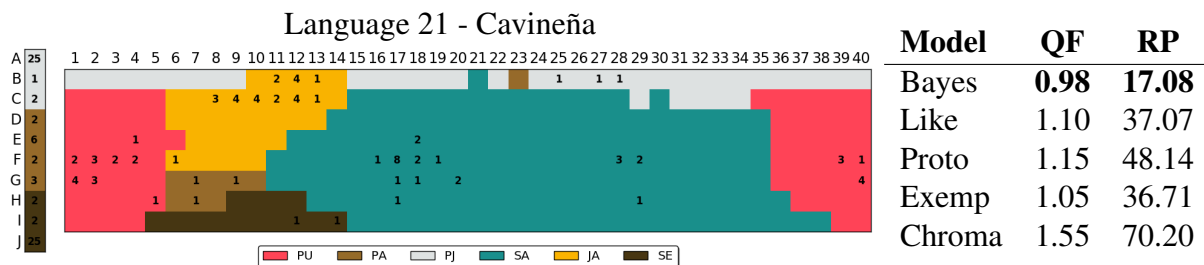
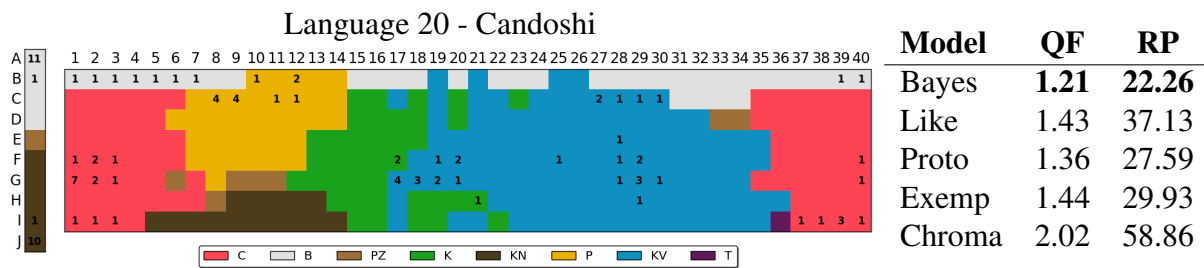
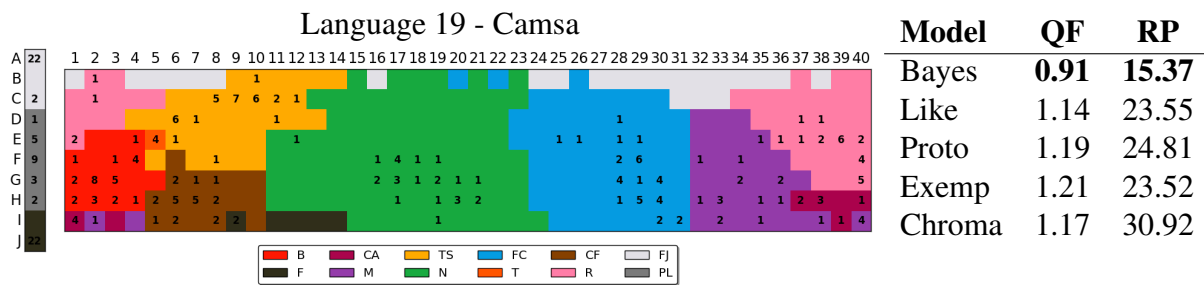
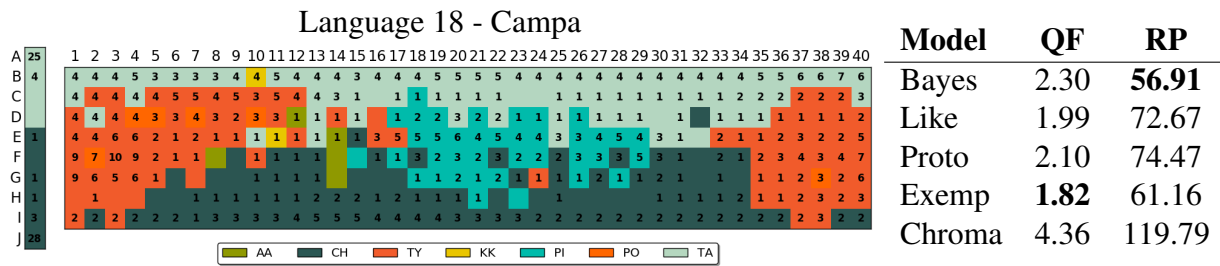
Here we present language-level analyses of the WCS data, along with model performance for each. We present the naming data as mode maps, displaying terms used by a plurality of the speakers. The number of focus hits per color chip, aggregated over speakers in the language, are overlaid on top, in line with our treatment of Dani and Berinmo from Chapter 4. Although the use of mode maps for visualization provides only a partial view of a language's color naming system, one can clearly see both similarities and differences in color naming patterns across languages.

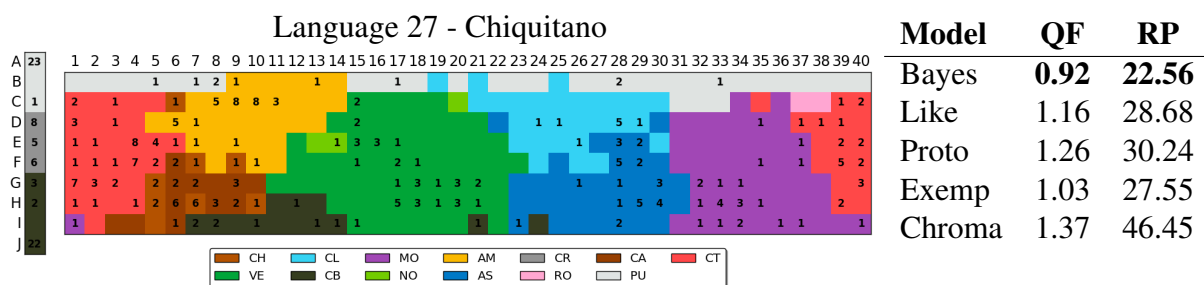
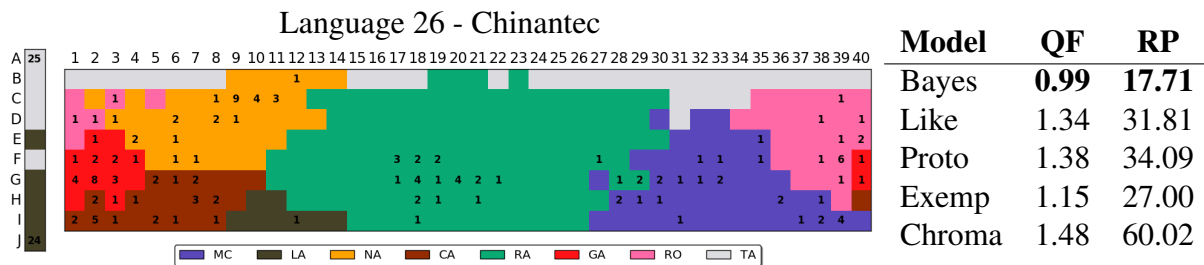
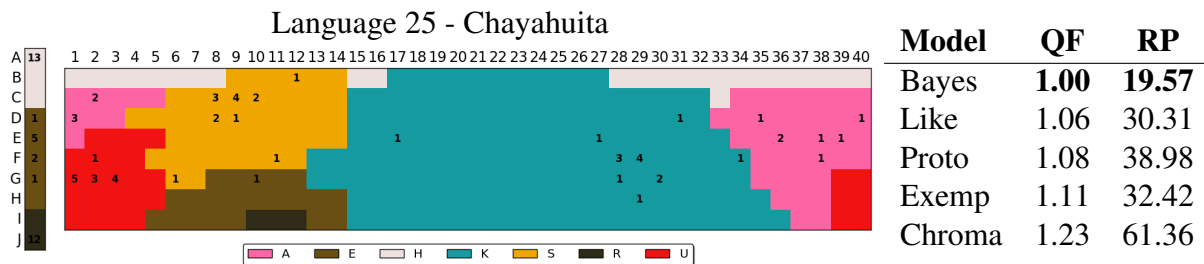
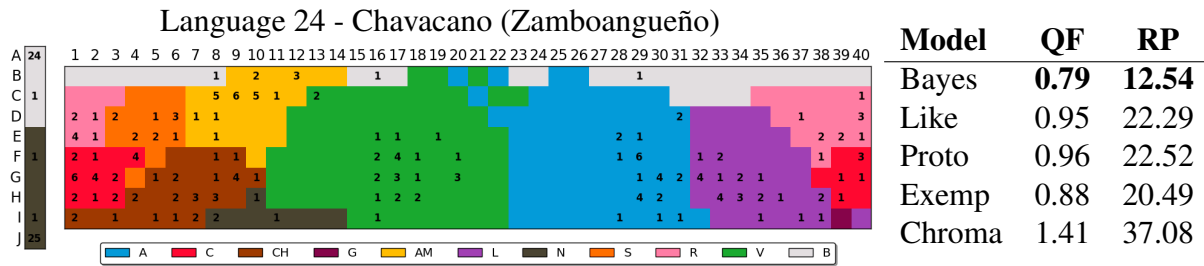
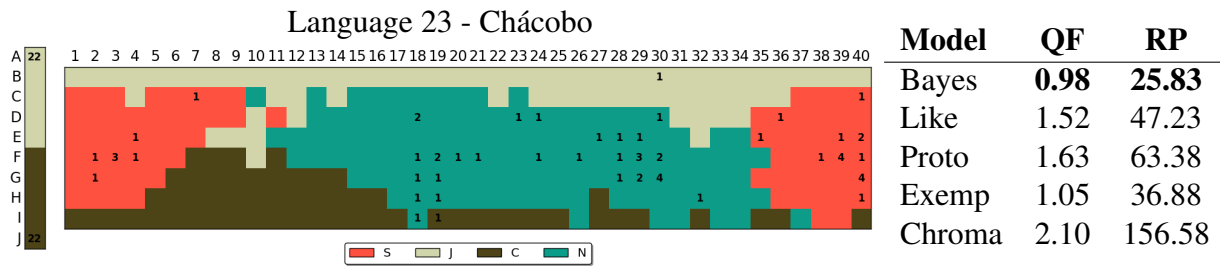


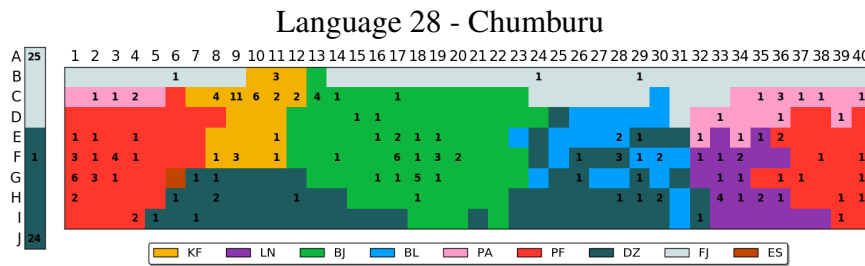




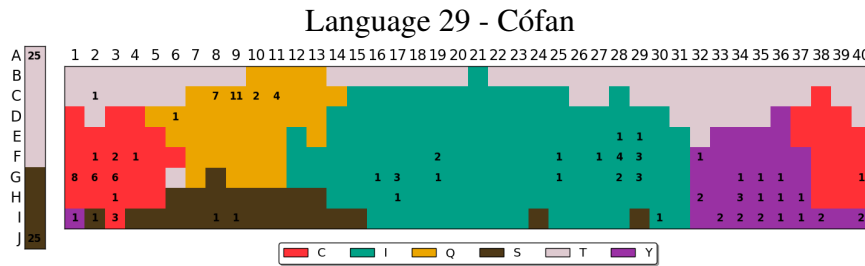




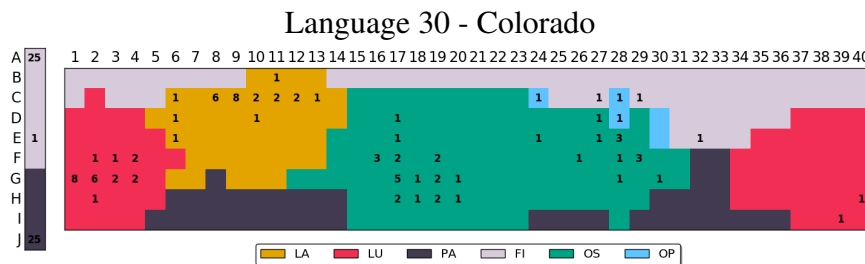




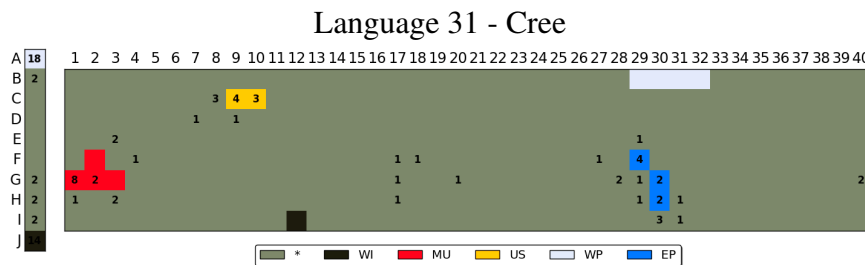
Model	QF	RP
Bayes	1.03	17.77
Like	1.33	31.86
Proto	1.37	34.61
Exemp	1.07	26.94
Chroma	1.56	69.44



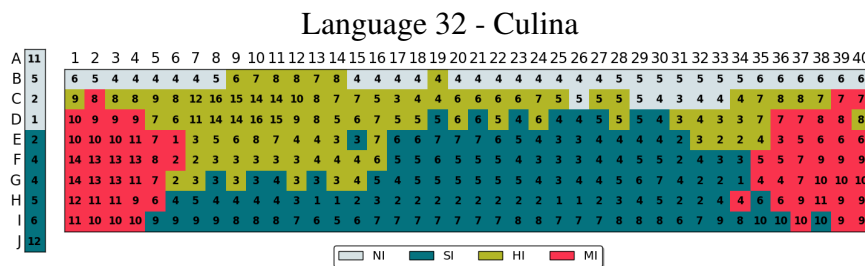
Model	QF	RP
Bayes	1.02	18.44
Like	1.16	30.54
Proto	1.20	37.58
Exemp	1.16	28.65
Chroma	1.51	77.24



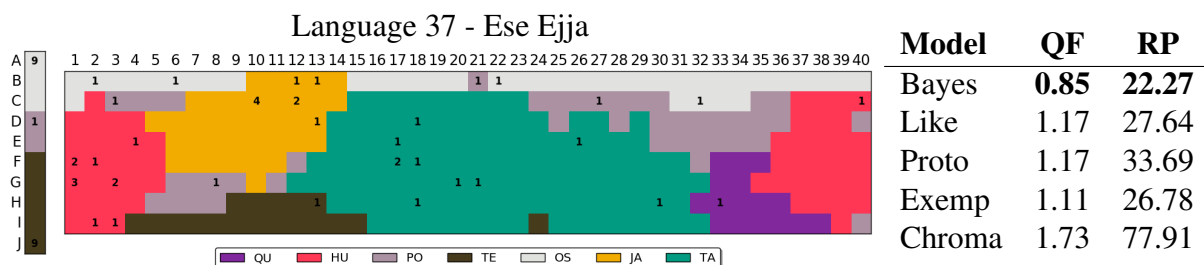
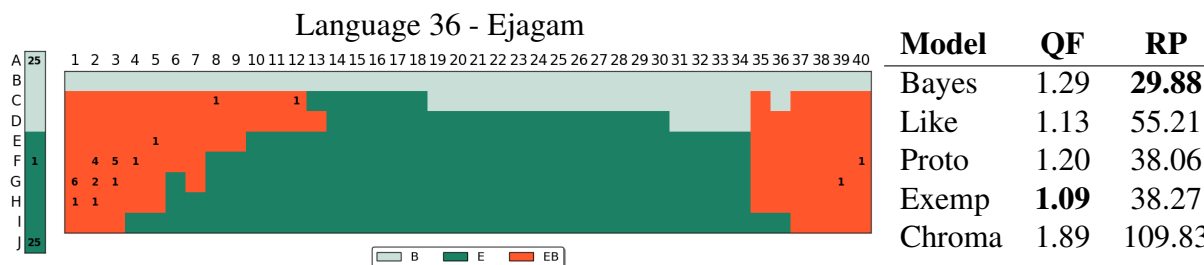
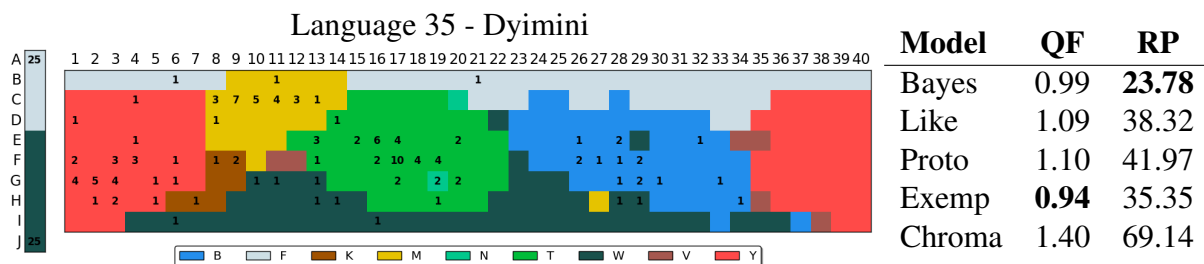
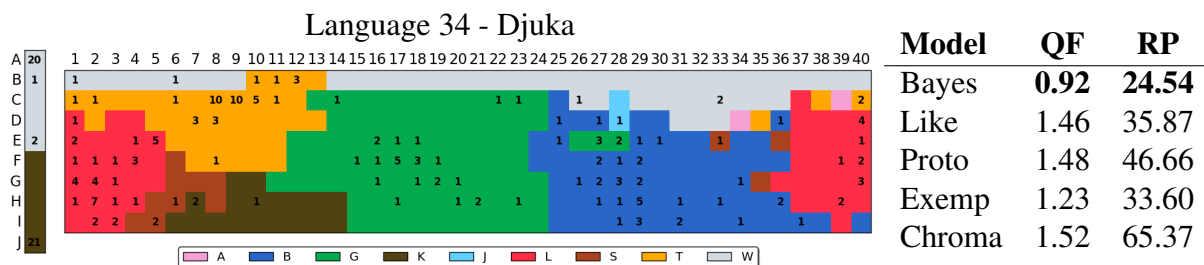
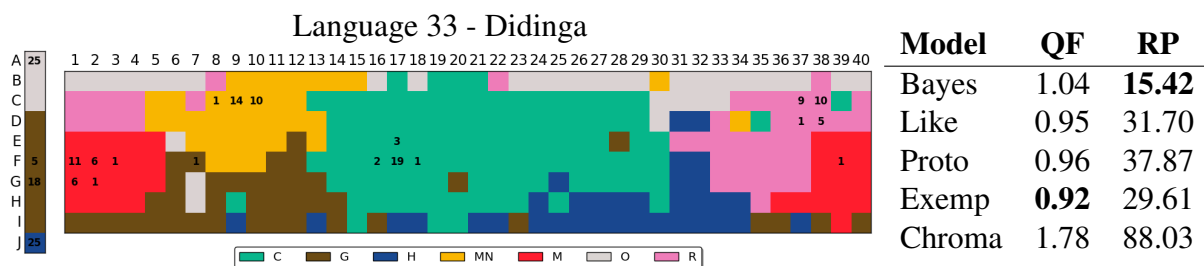
Model	QF	RP
Bayes	1.26	16.73
Like	1.38	37.98
Proto	1.35	39.84
Exemp	1.33	33.70
Chroma	1.57	72.92

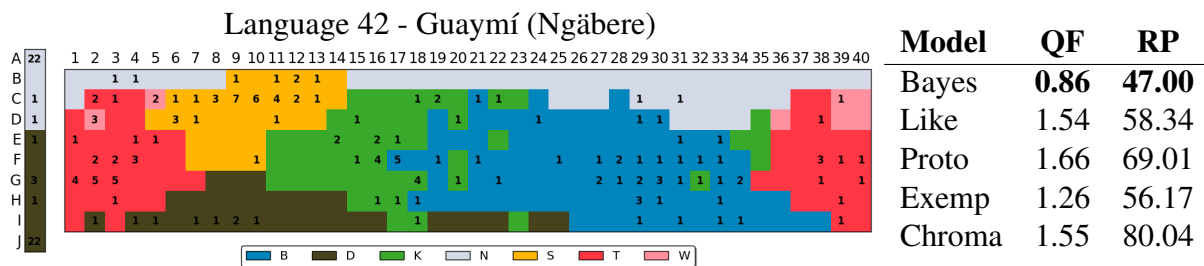
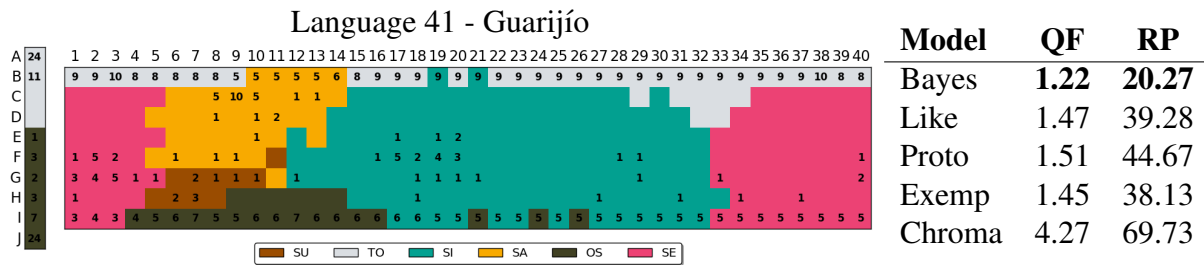
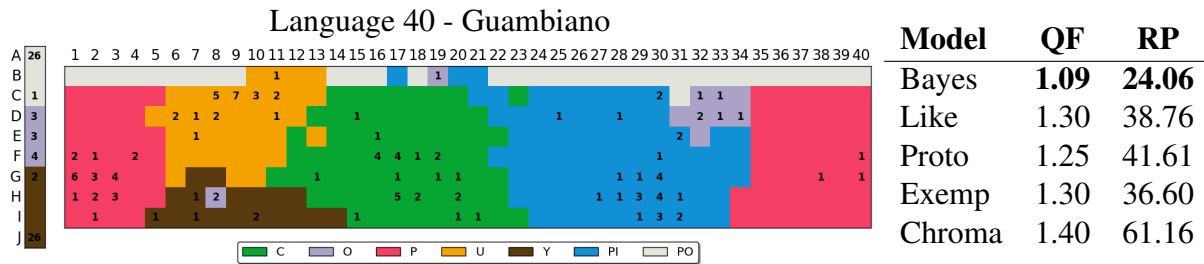
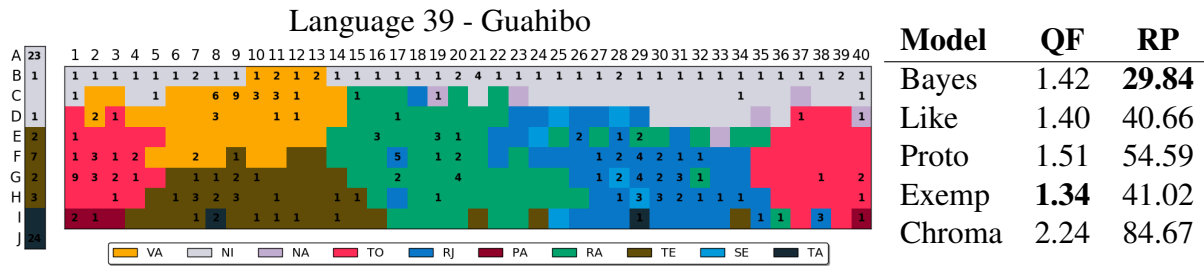
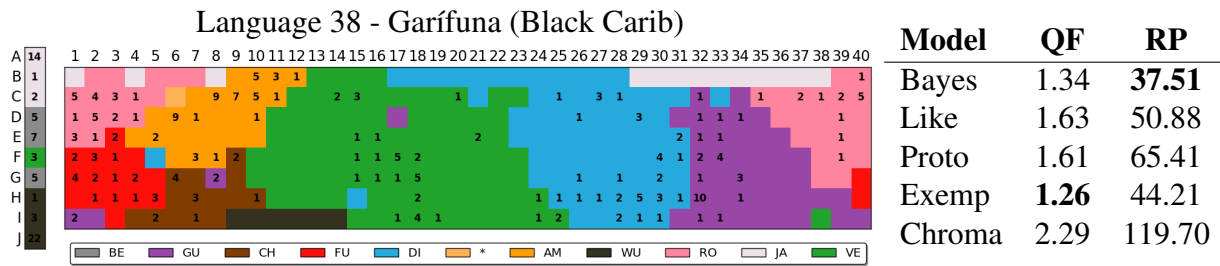


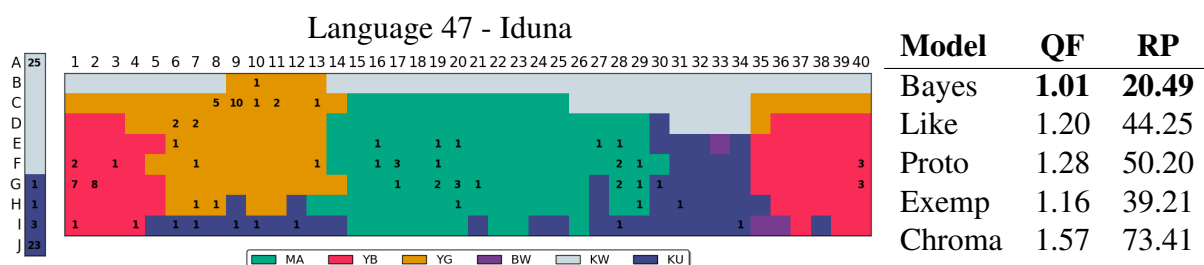
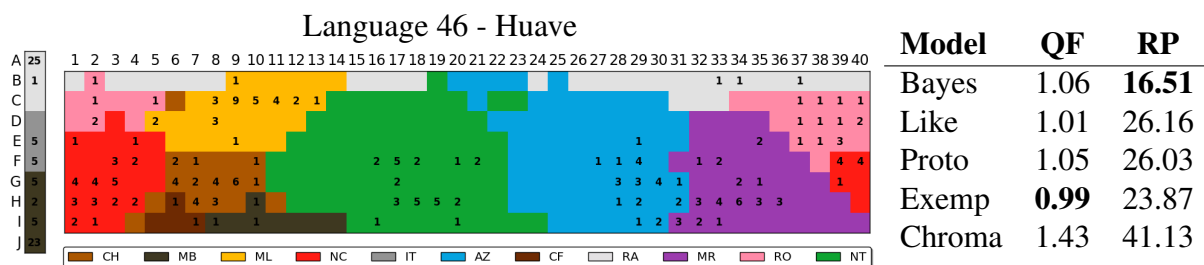
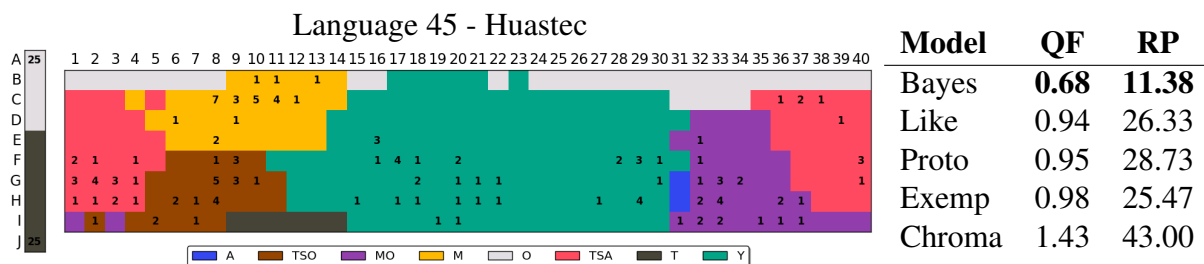
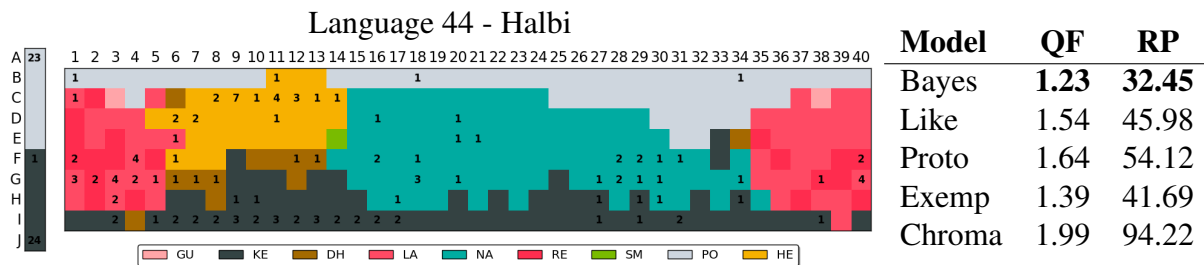
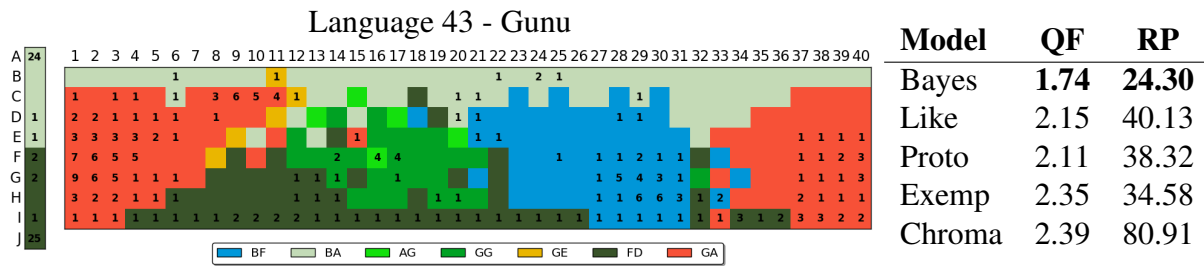
Model	QF	RP
Bayes	1.13	15.45
Like	1.13	16.06
Proto	1.13	20.21
Exemp	1.05	14.99
Chroma	0.93	21.95

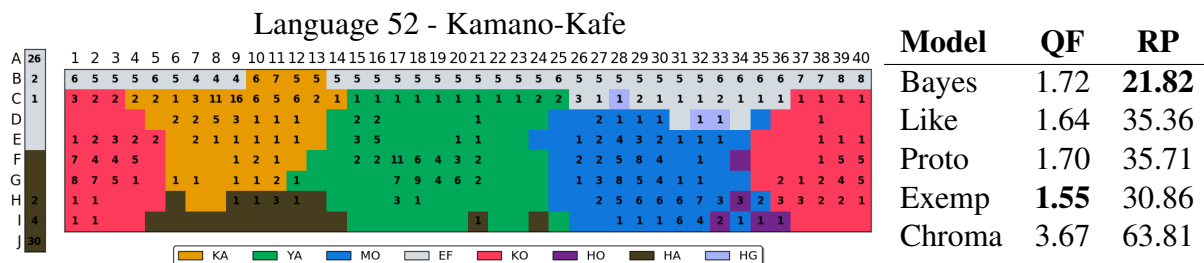
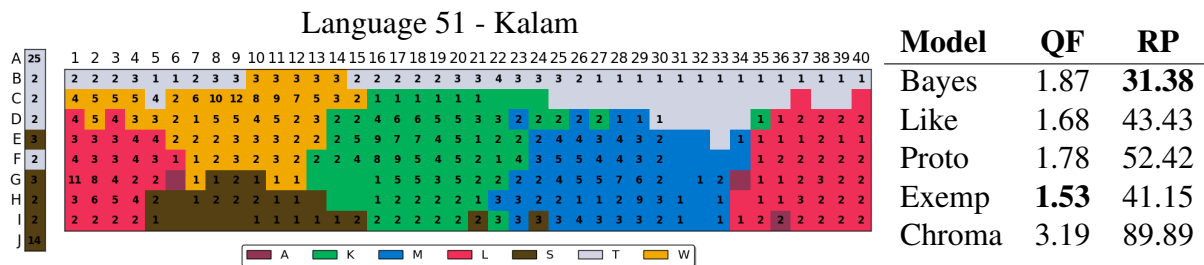
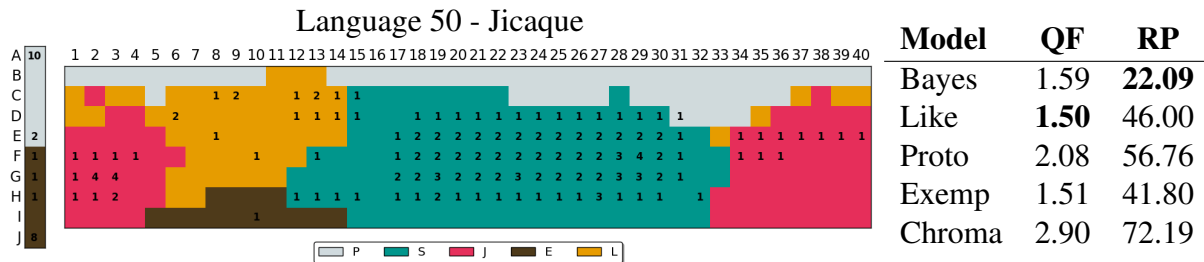
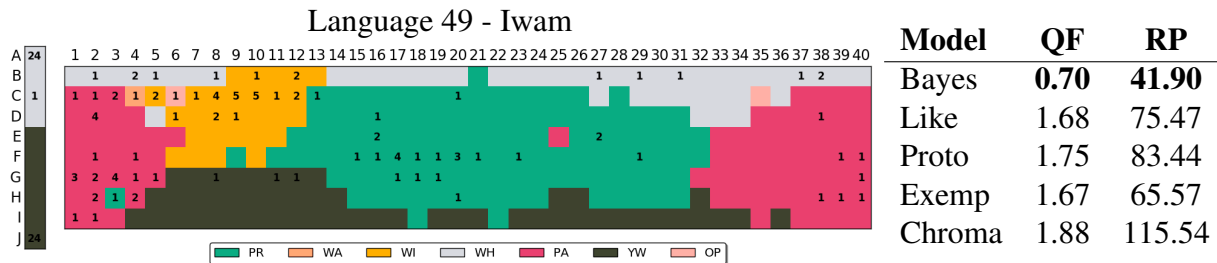
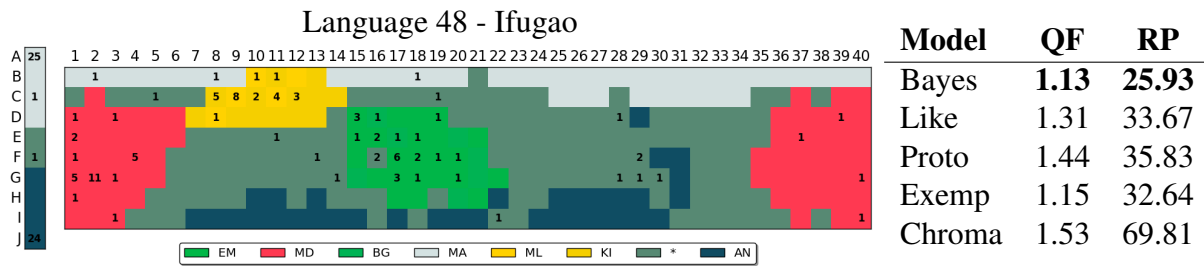


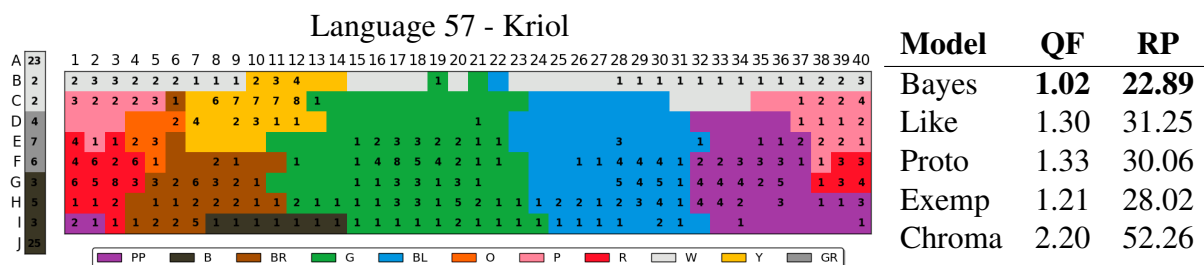
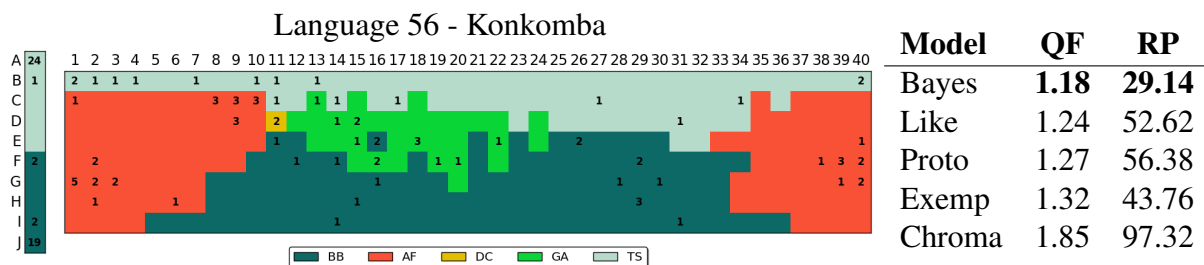
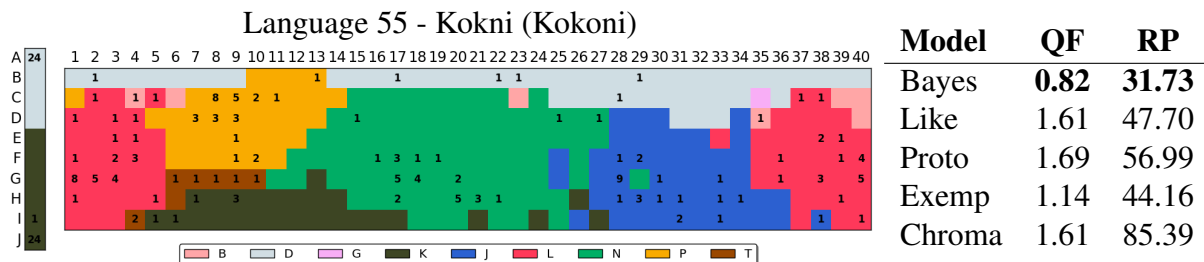
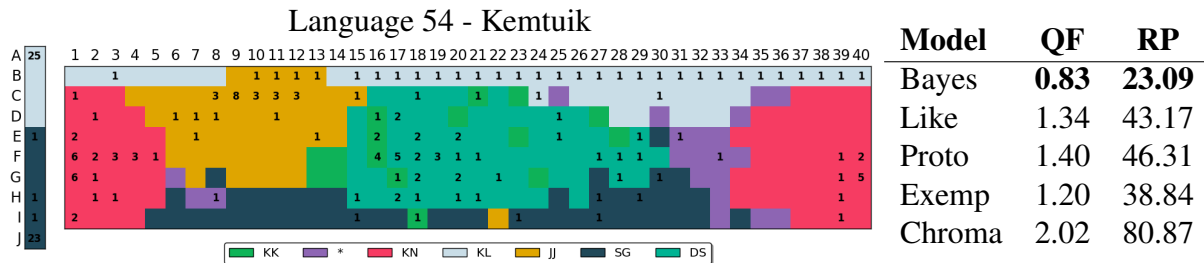
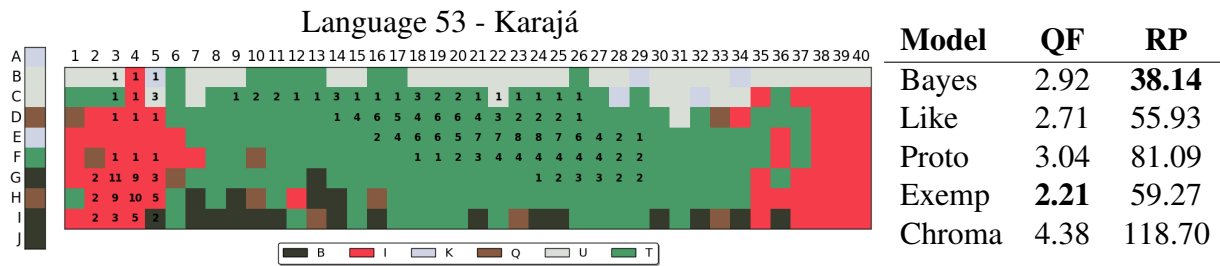
Model	QF	RP
Bayes	1.32	82.84
Like	1.75	81.94
Proto	2.04	89.68
Exemp	1.99	81.20
Chroma	3.67	113.37

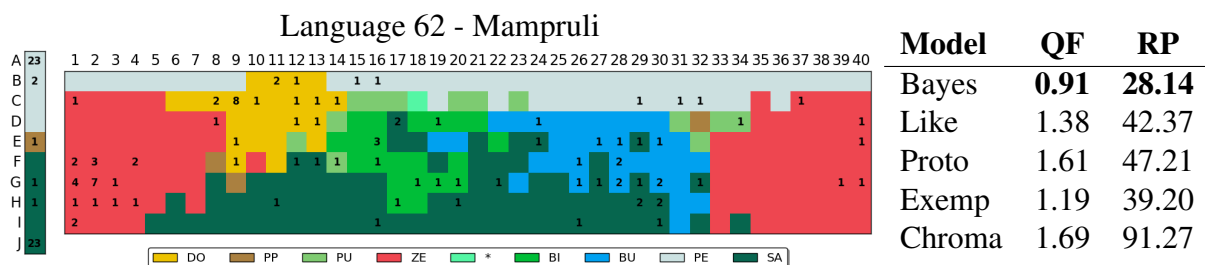
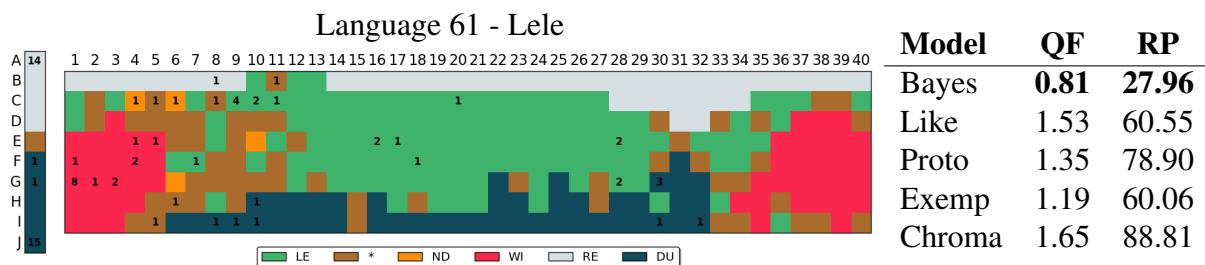
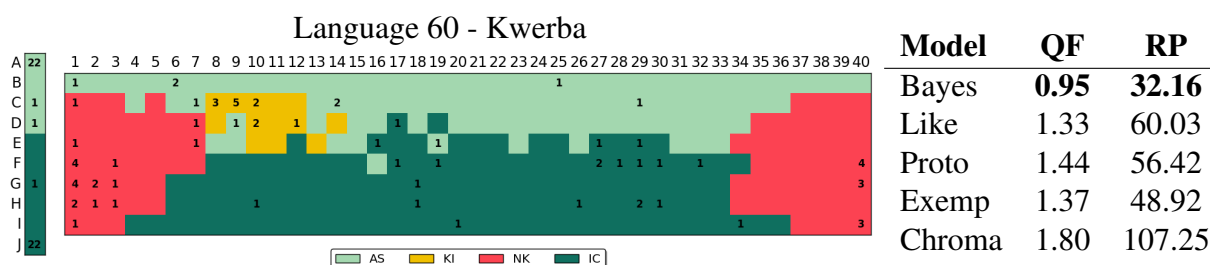
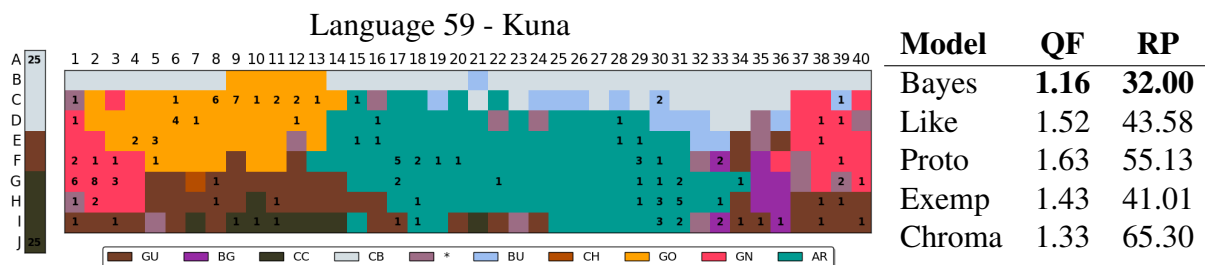
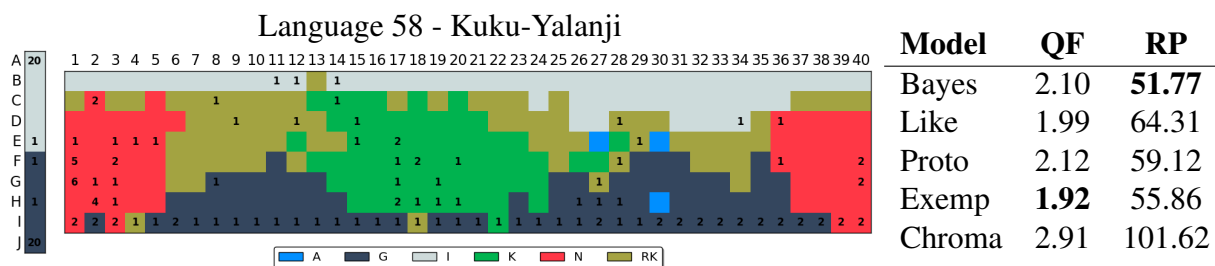


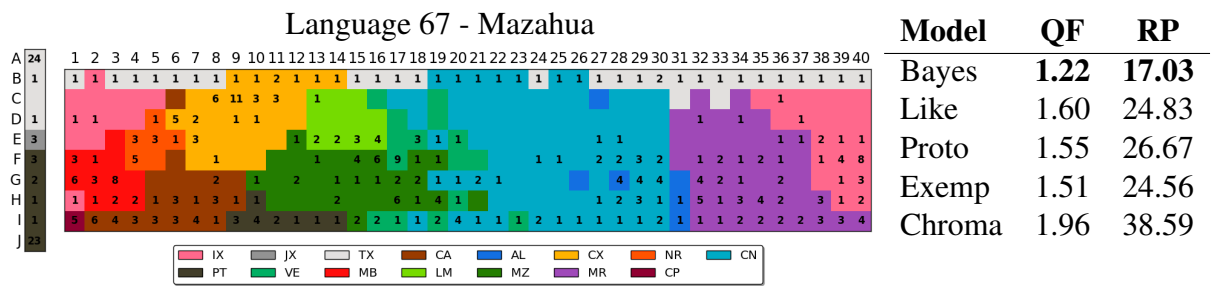
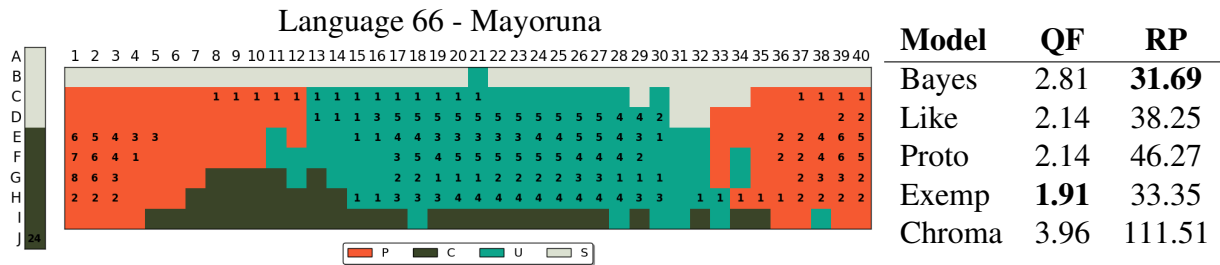
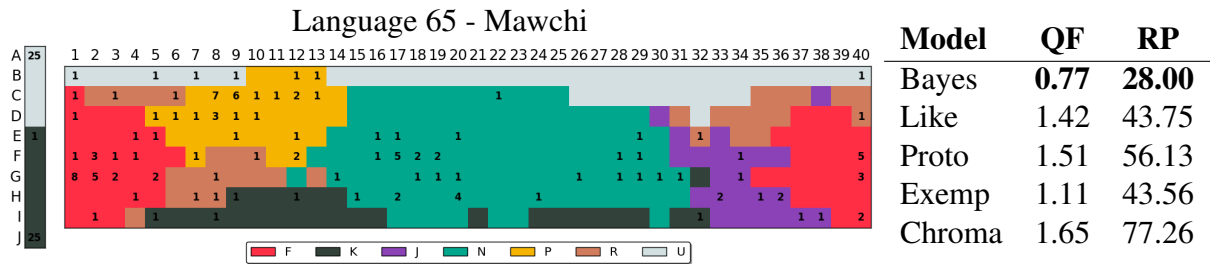
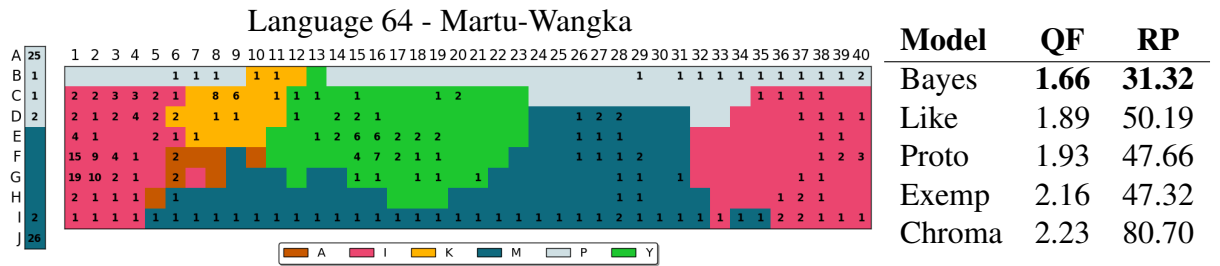
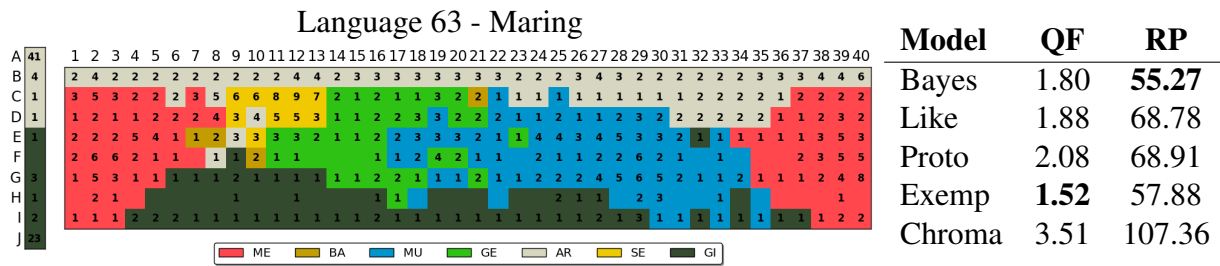




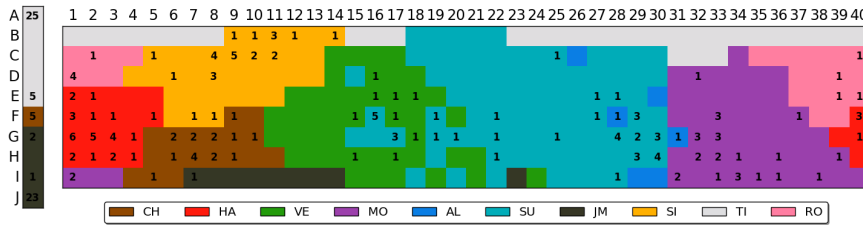






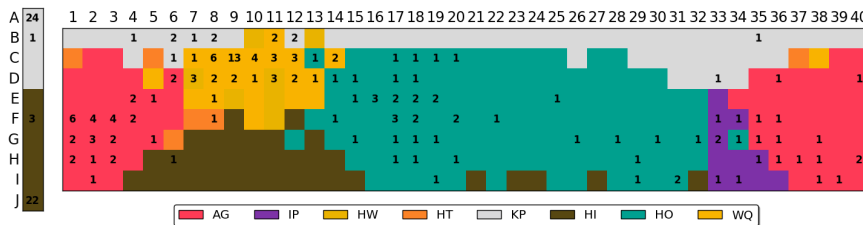


Language 68 - Mazatec



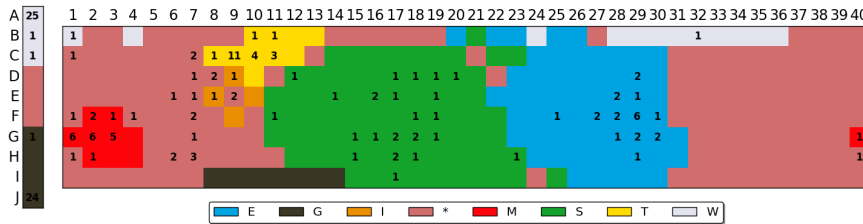
Model	QF	RP
Bayes	0.85	20.38
Like	1.18	28.92
Proto	1.30	32.47
Exemp	1.23	28.77
Chroma	1.42	52.61

Language 69 - Menye



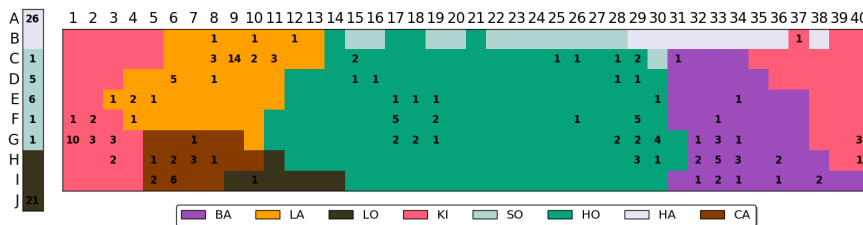
Model	QF	RP
Bayes	0.87	29.62
Like	1.62	48.40
Proto	1.90	69.59
Exemp	1.29	44.54
Chroma	1.58	90.57

Language 70 - Micmac



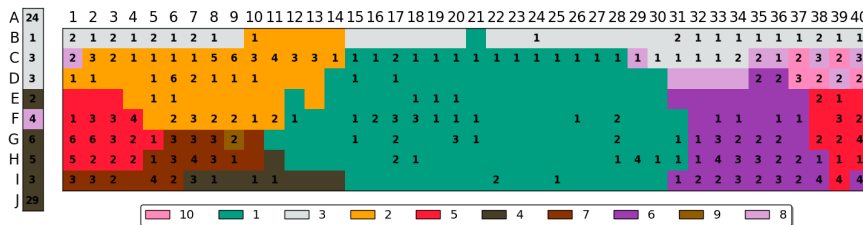
Model	QF	RP
Bayes	1.15	17.76
Like	0.97	21.71
Proto	0.88	22.98
Exemp	1.03	19.56
Chroma	1.32	42.31

Language 71 - Mikasuki

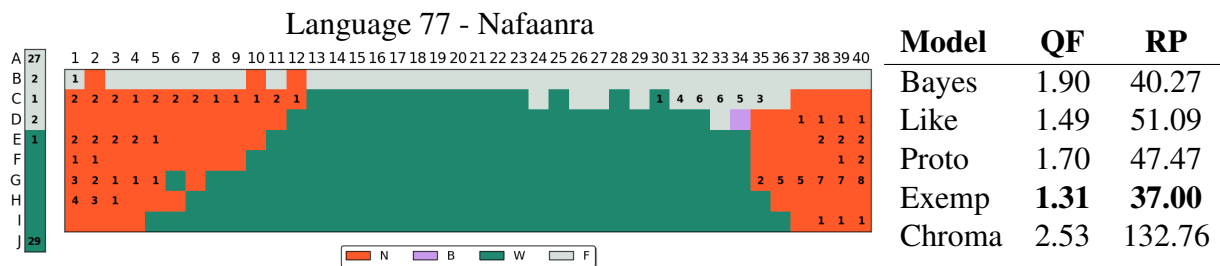
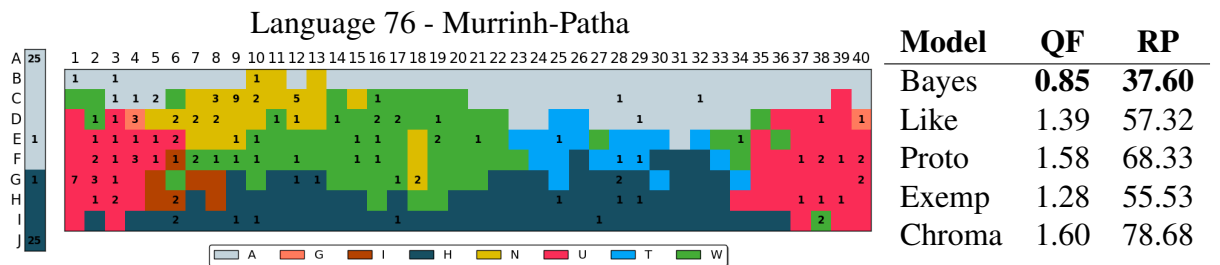
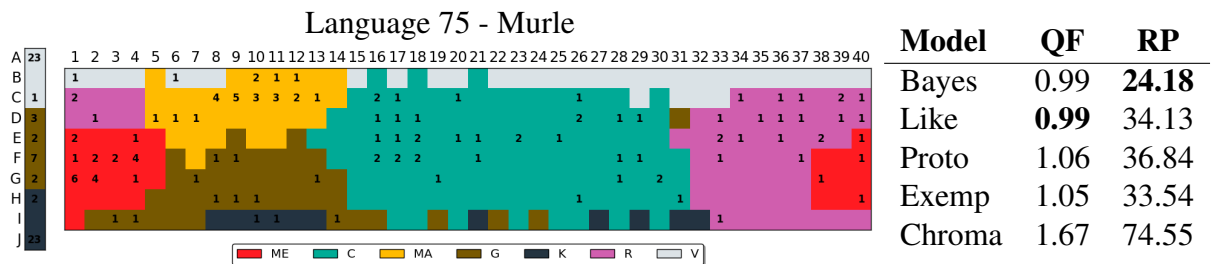
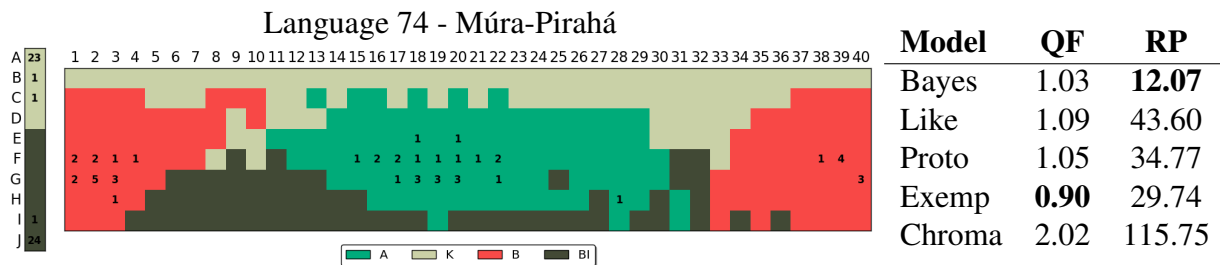
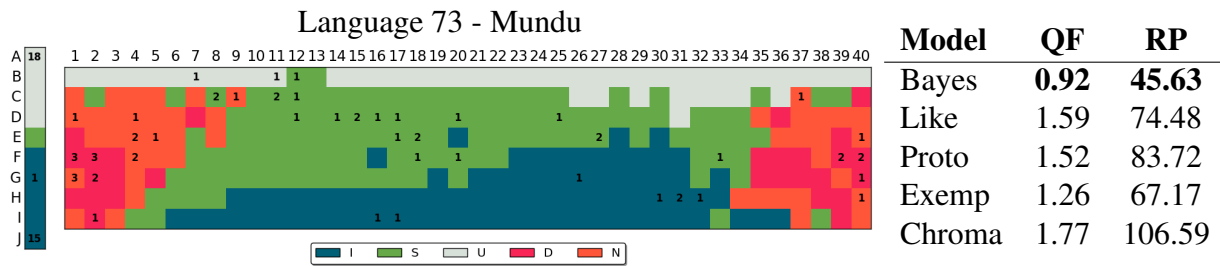


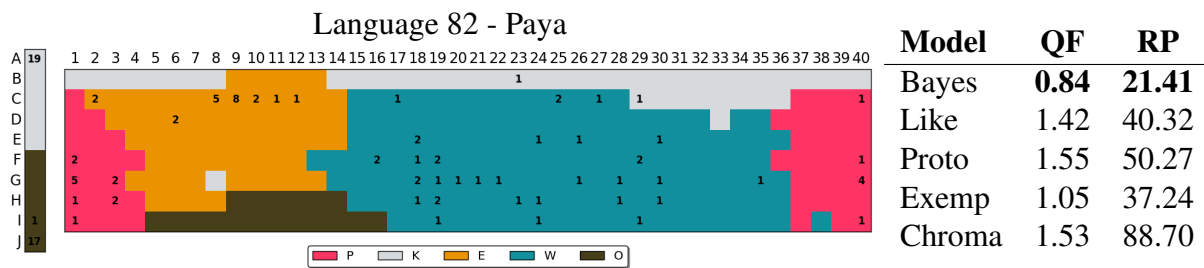
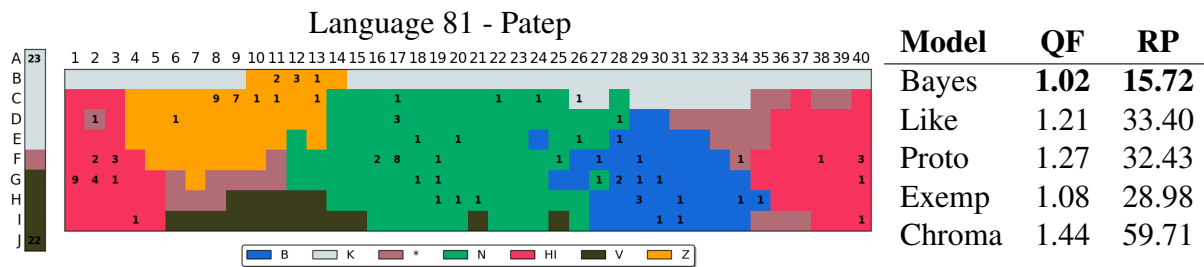
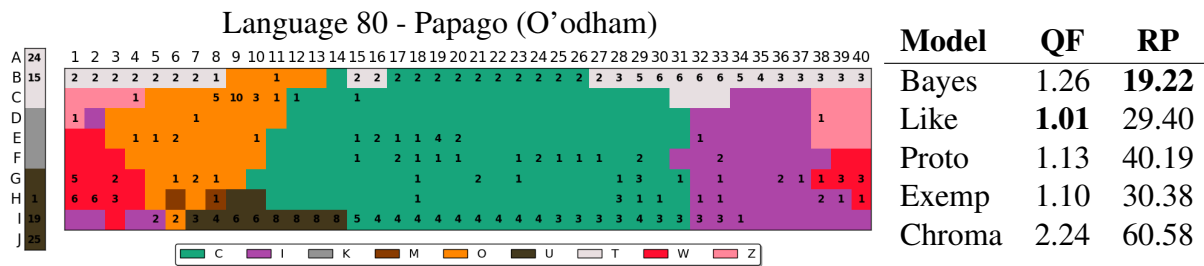
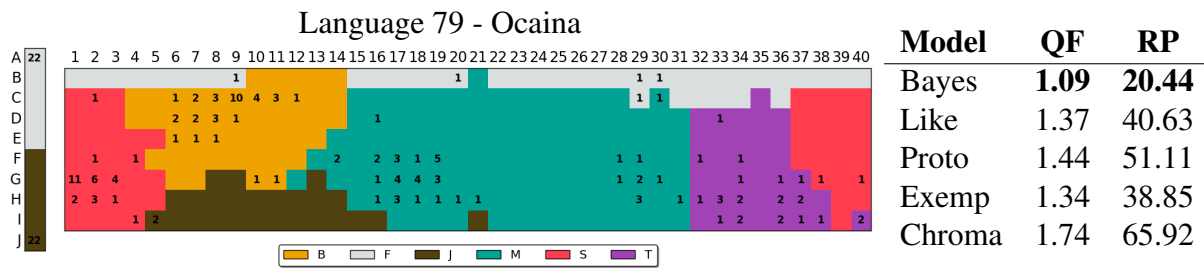
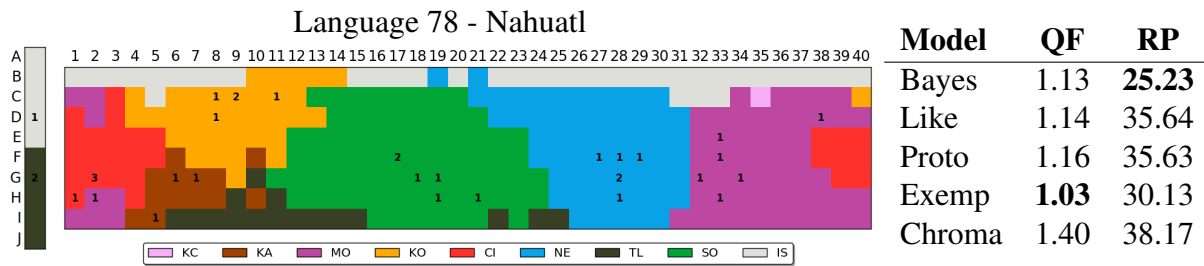
Model	QF	RP
Bayes	0.87	14.79
Like	1.51	32.31
Proto	1.51	42.61
Exemp	1.38	33.34
Chroma	1.28	54.15

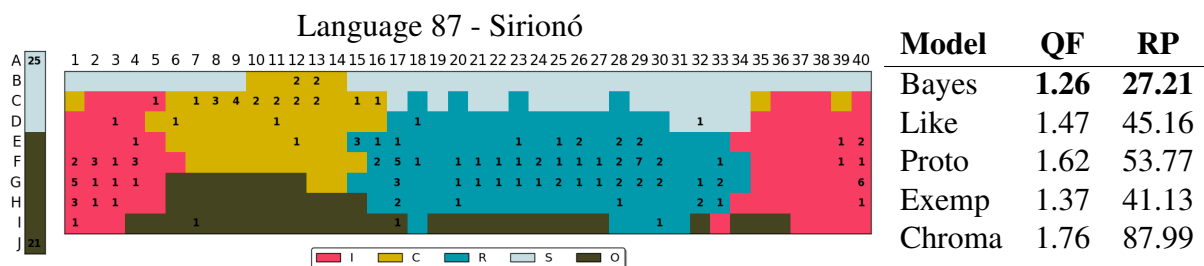
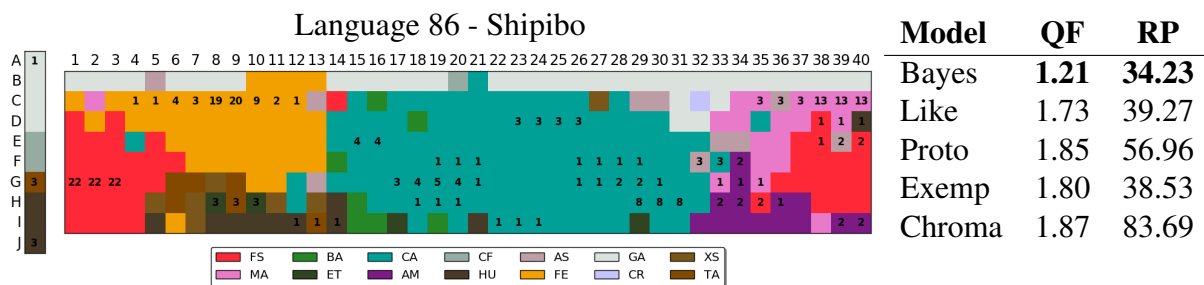
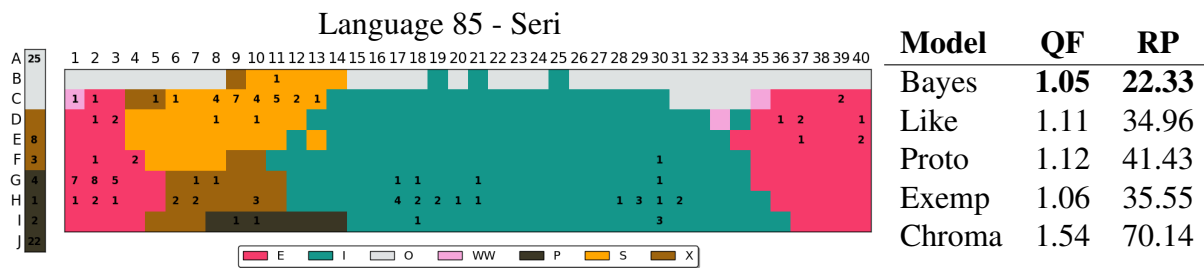
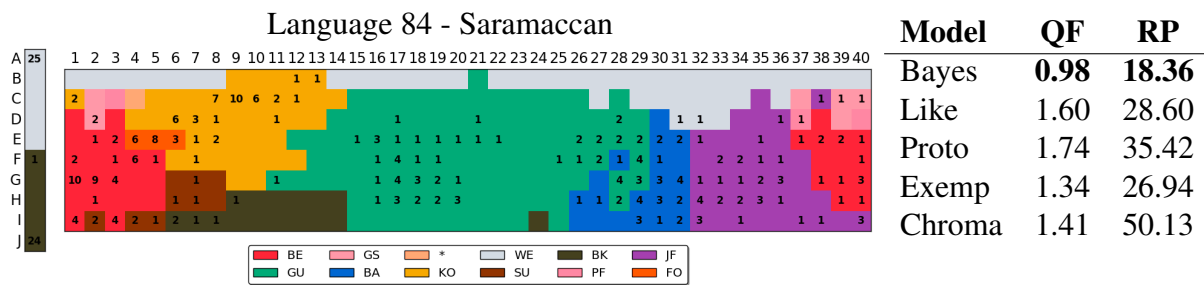
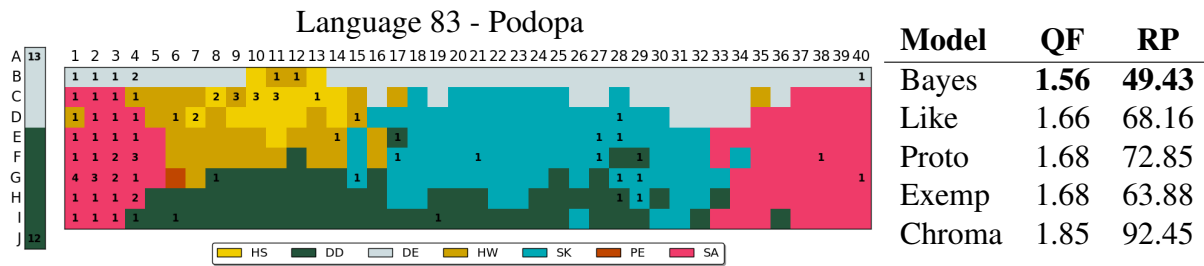
Language 72 - Mixtec

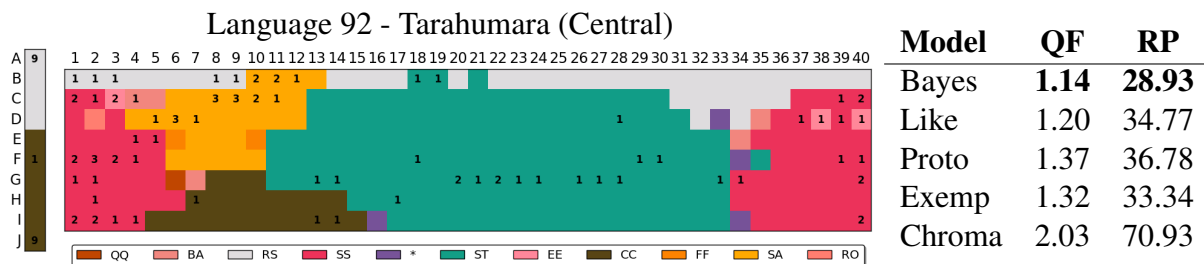
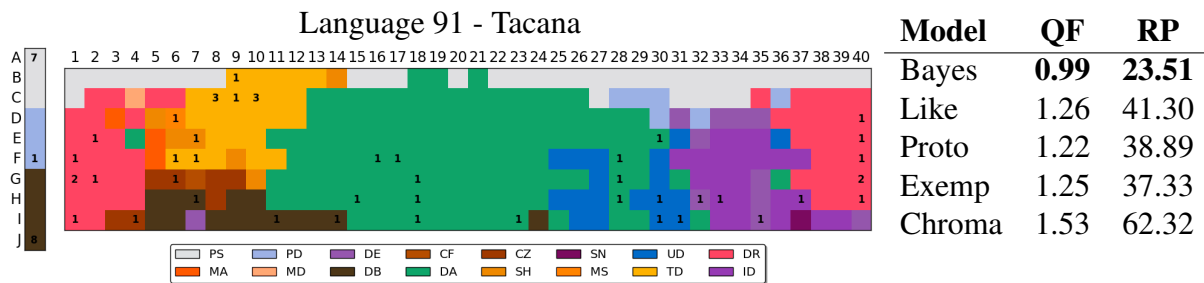
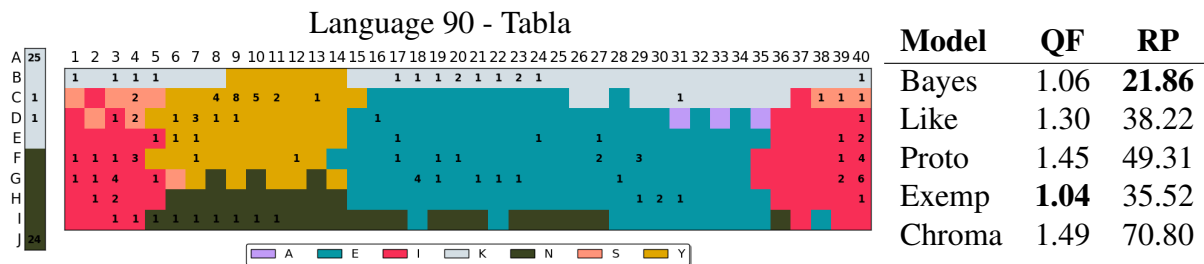
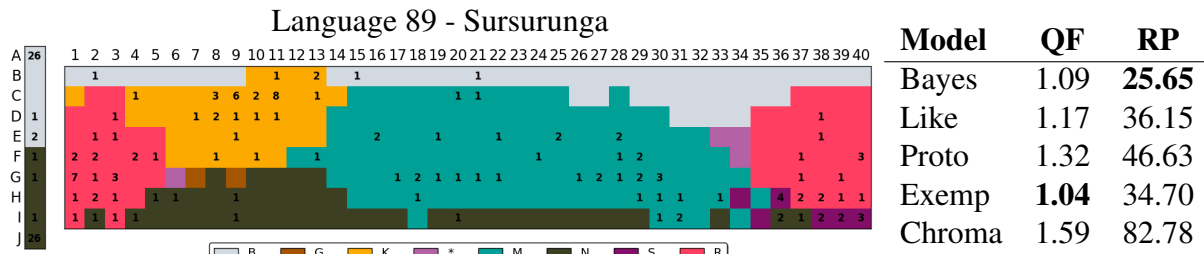
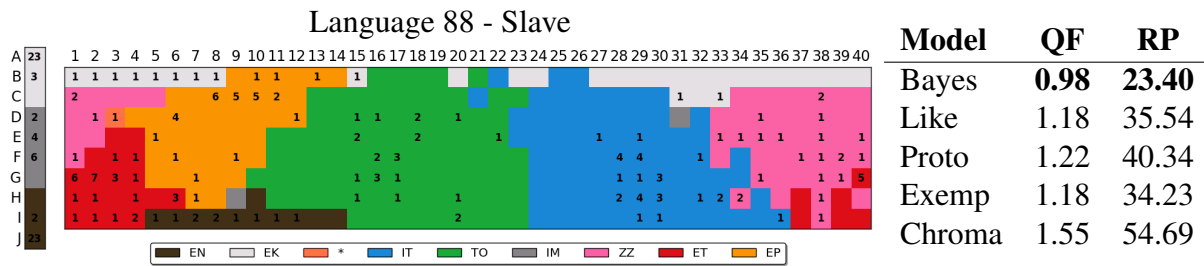


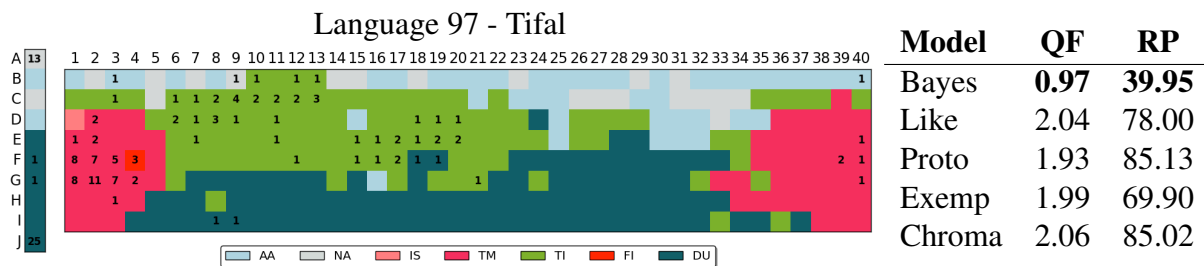
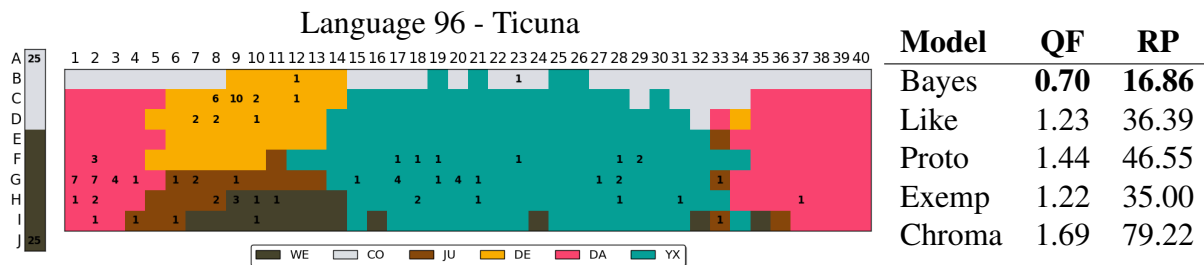
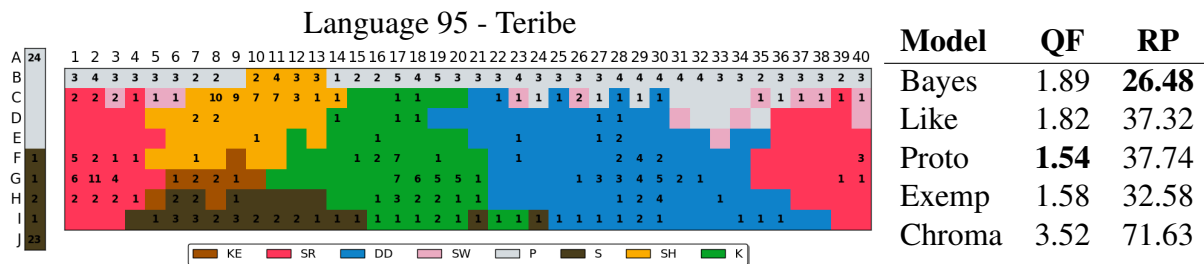
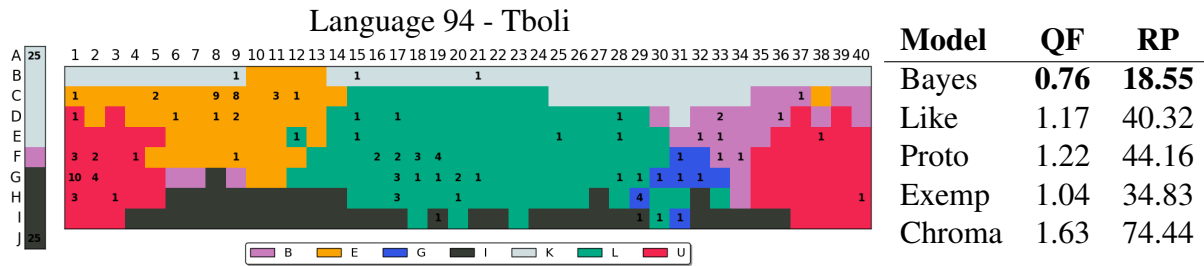
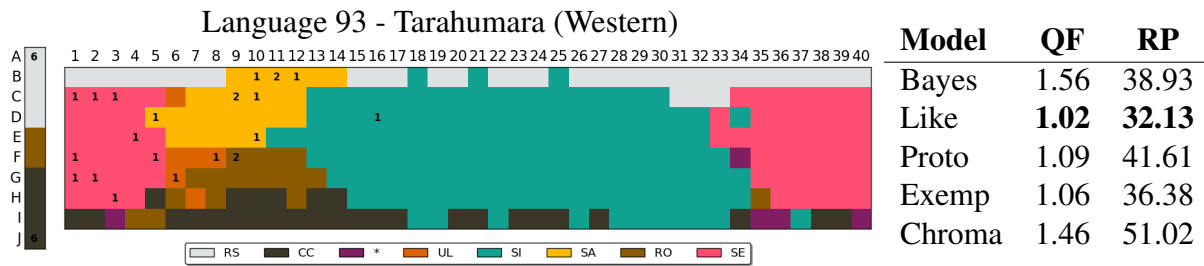
Model	QF	RP
Bayes	1.30	25.51
Like	1.50	35.29
Proto	1.49	37.52
Exemp	1.48	33.93
Chroma	2.12	57.58

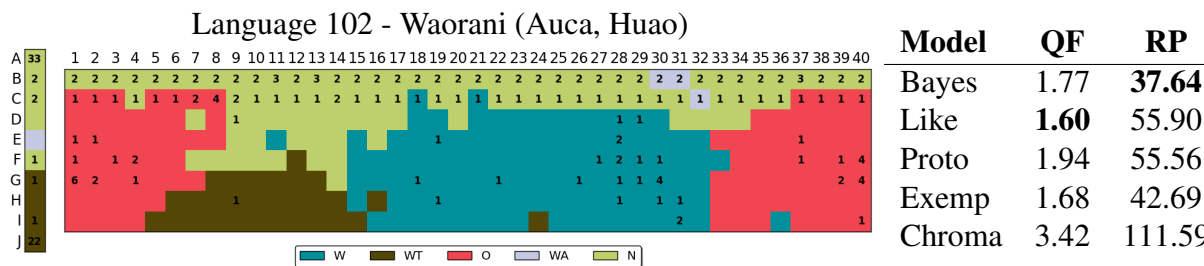
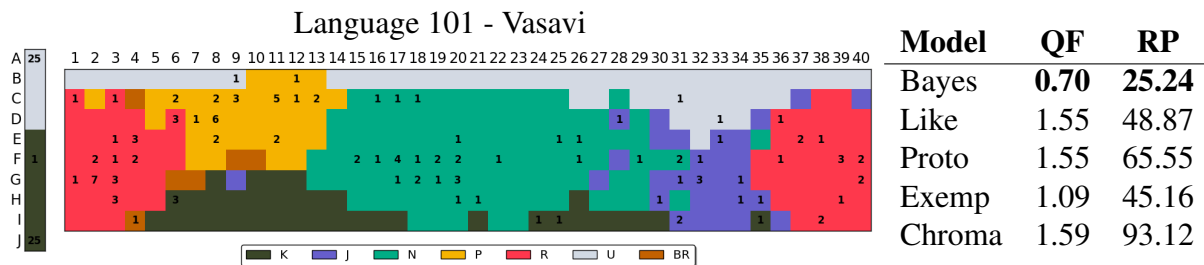
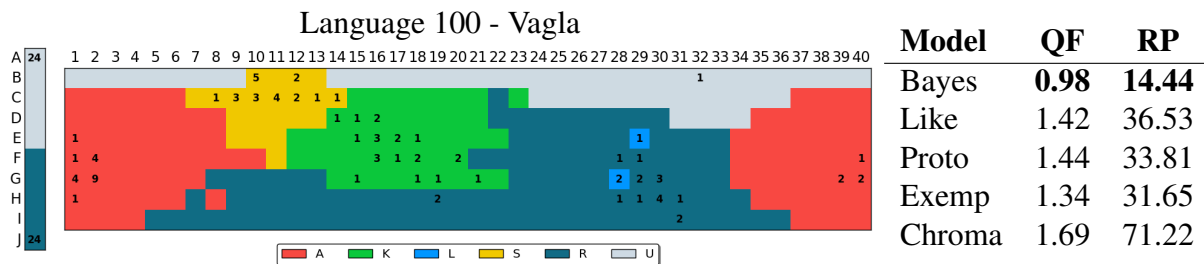
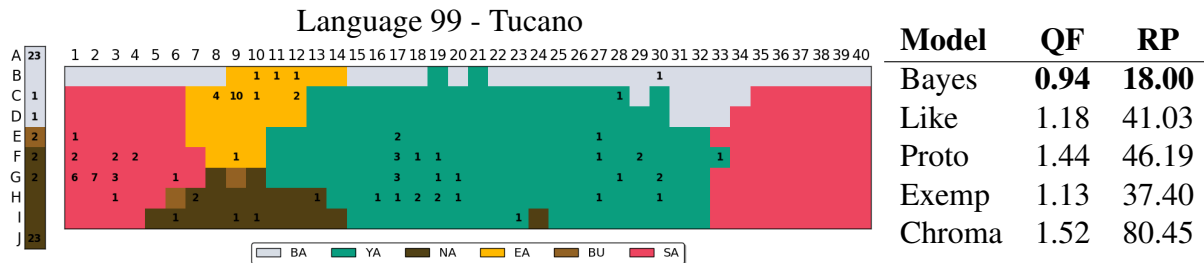
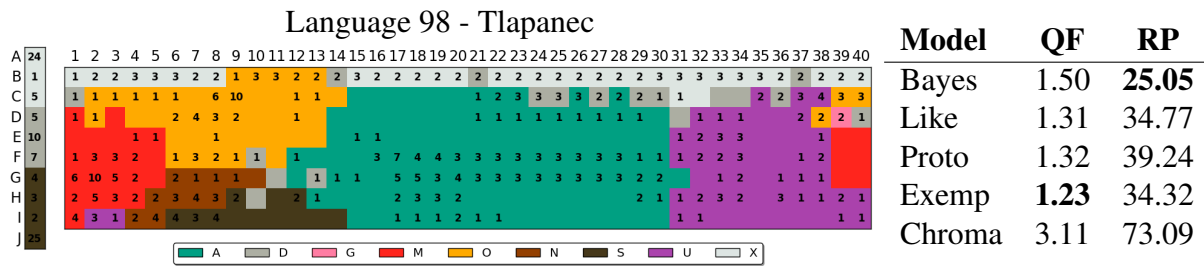


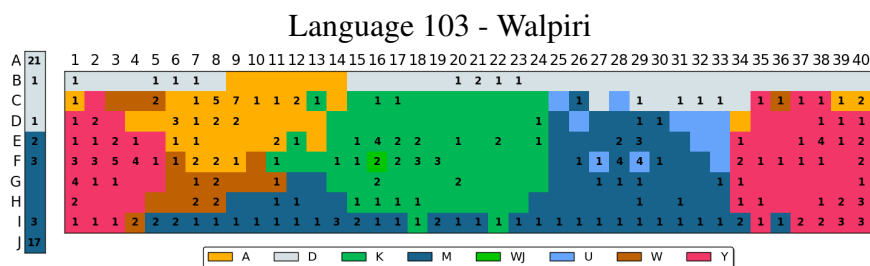




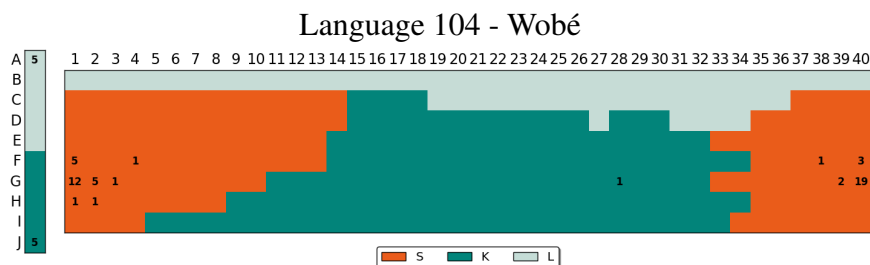




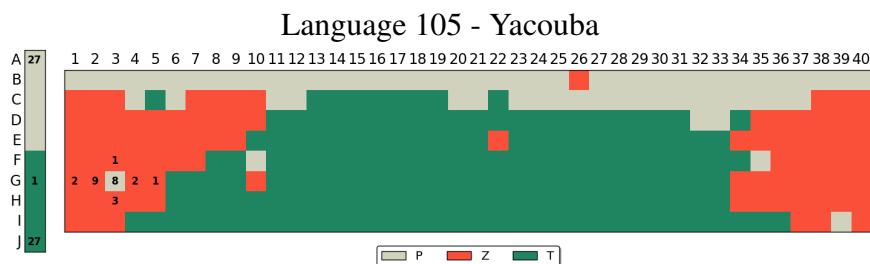




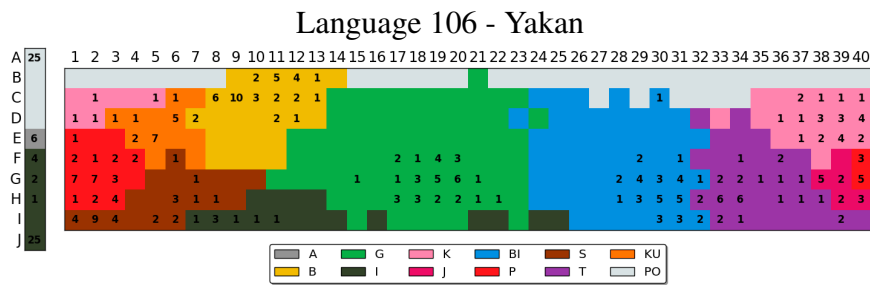
Model	QF	RP
Bayes	1.15	48.09
Like	1.25	52.95
Proto	1.37	56.05
Exemp	1.14	51.34
Chroma	2.39	82.77



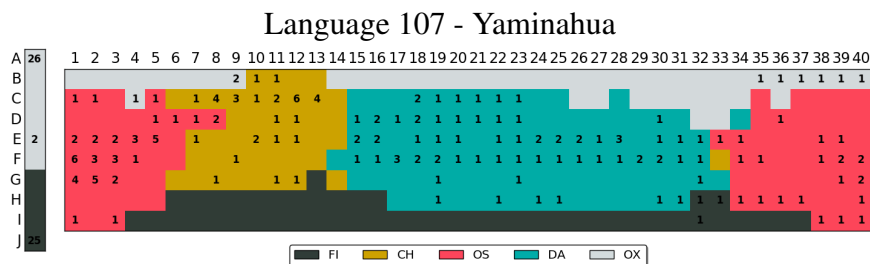
Model	QF	RP
Bayes	2.20	57.67
Like	2.17	61.34
Proto	2.13	57.82
Exemp	1.79	42.31
Chroma	2.30	110.96



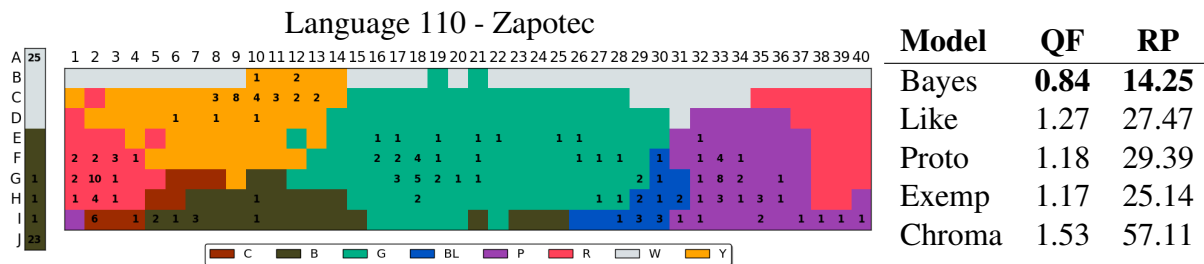
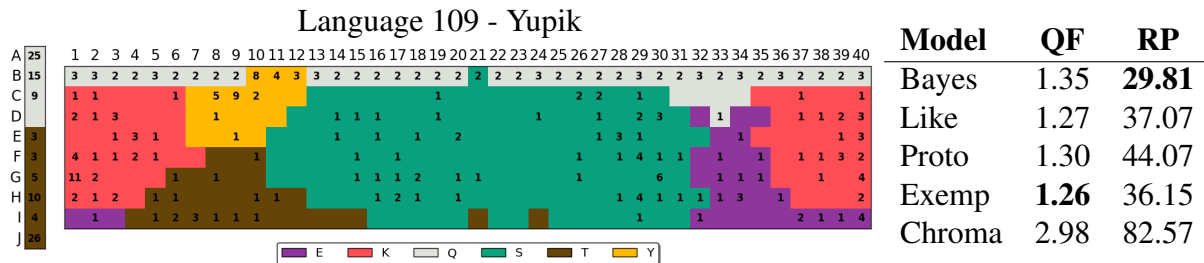
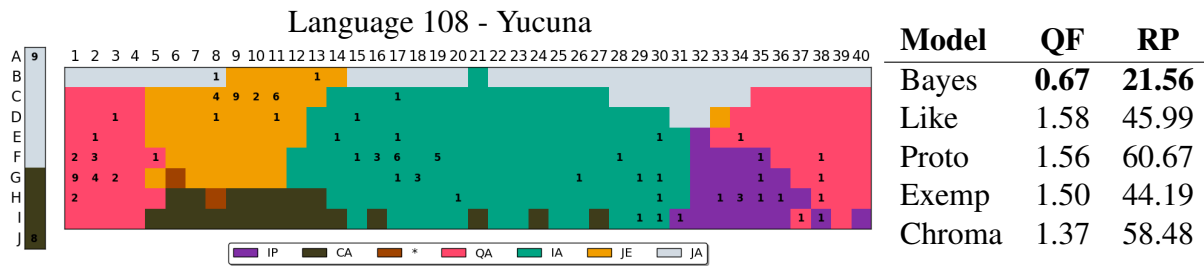
Model	QF	RP
Bayes	1.12	24.77
Like	1.30	65.11
Proto	1.45	53.79
Exemp	1.20	49.73
Chroma	1.83	135.72



Model	QF	RP
Bayes	1.19	15.38
Like	1.42	24.01
Proto	1.47	27.46
Exemp	1.33	23.06
Chroma	1.74	42.57



Model	QF	RP
Bayes	1.00	30.31
Like	1.62	51.68
Proto	1.76	66.49
Exemp	1.44	49.80
Chroma	2.03	90.45



B.2 Category Unusualness

In this section we provide further treatment of the category unusualness analyses from the main text. We present examples of WCS categories ranked by the unusualness measure below.

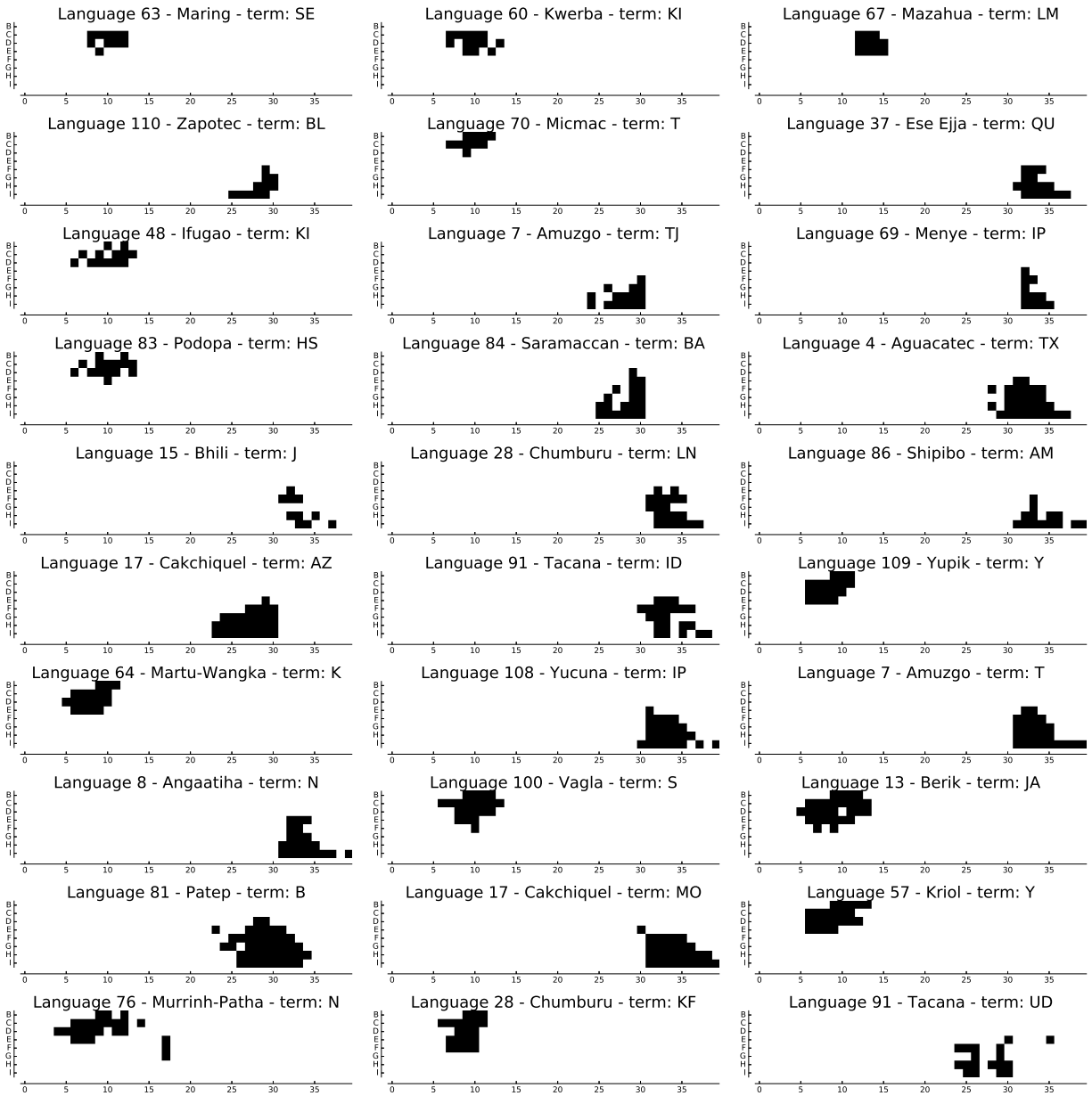


Figure B.1: The 30 most unusual WCS categories (presented in descending order of average Hausdorff distance)

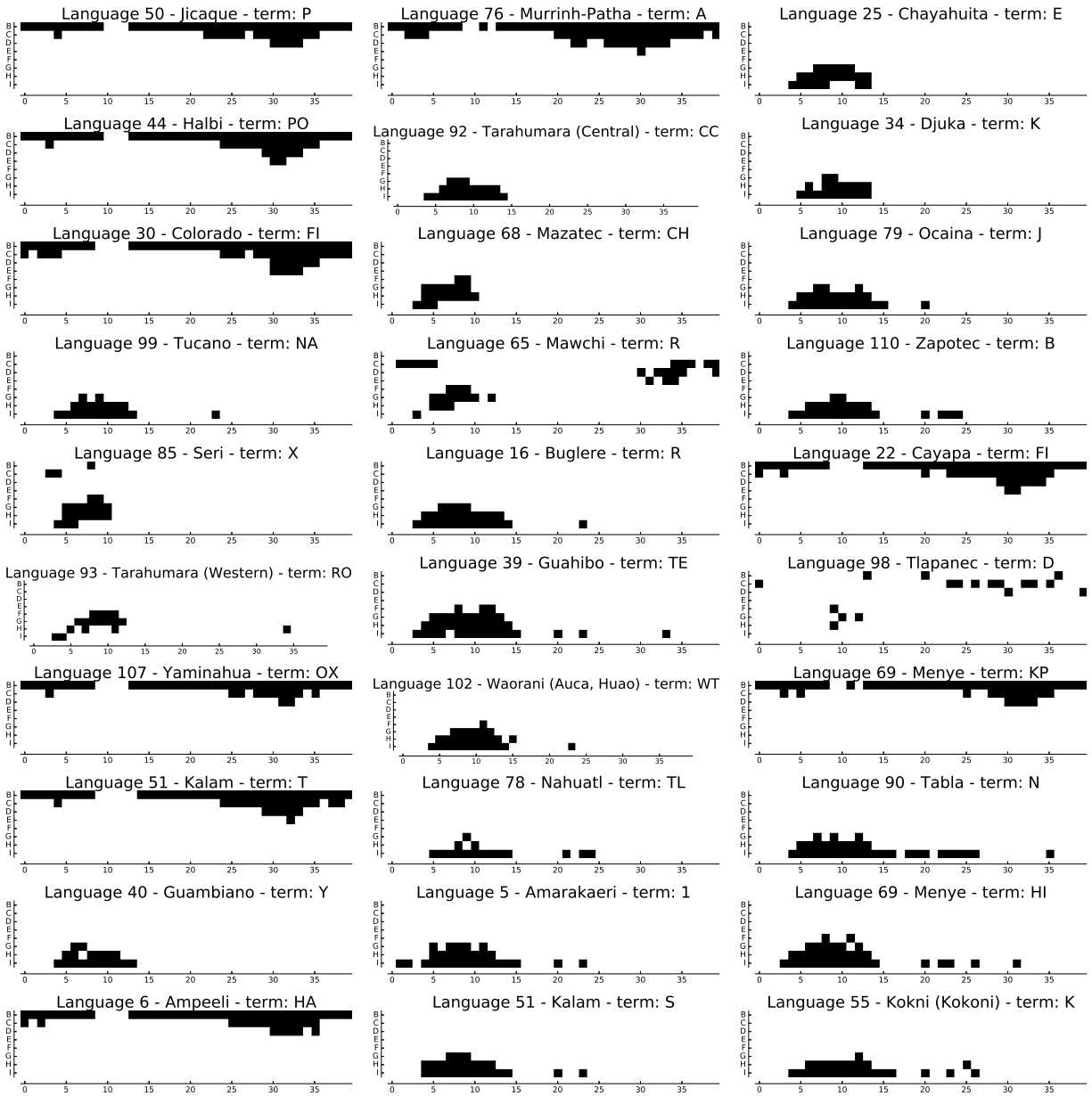


Figure B.2: The 30 least unusual WCS categories (presented in ascending ranked order of average Hausdorff distance)

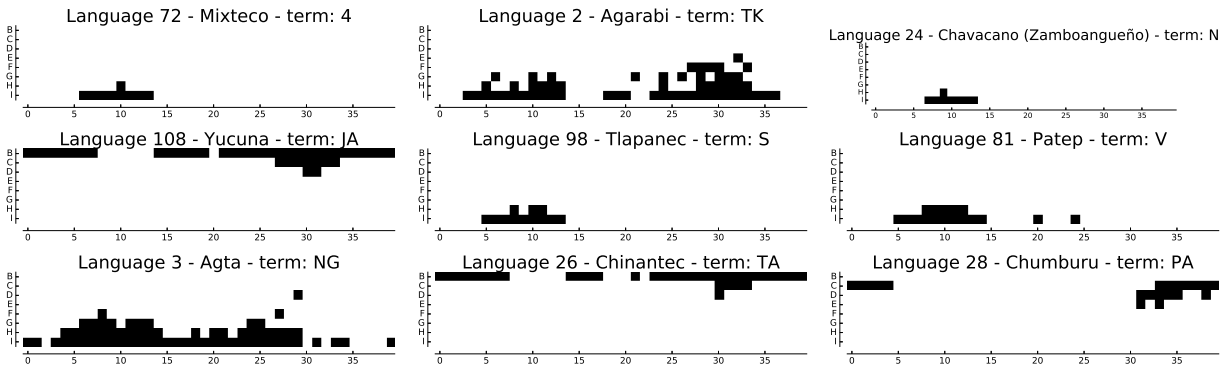


Figure B.3: WCS categories in the 25th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)

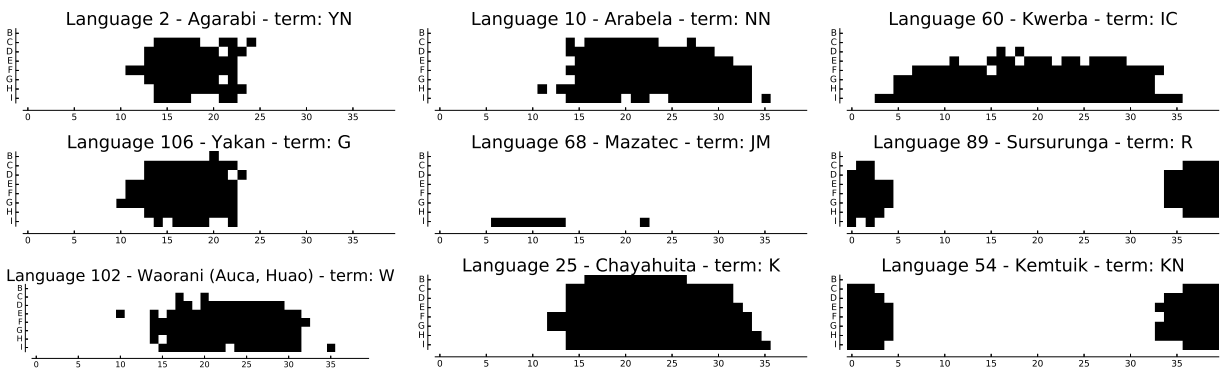


Figure B.4: WCS categories in the 50th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)

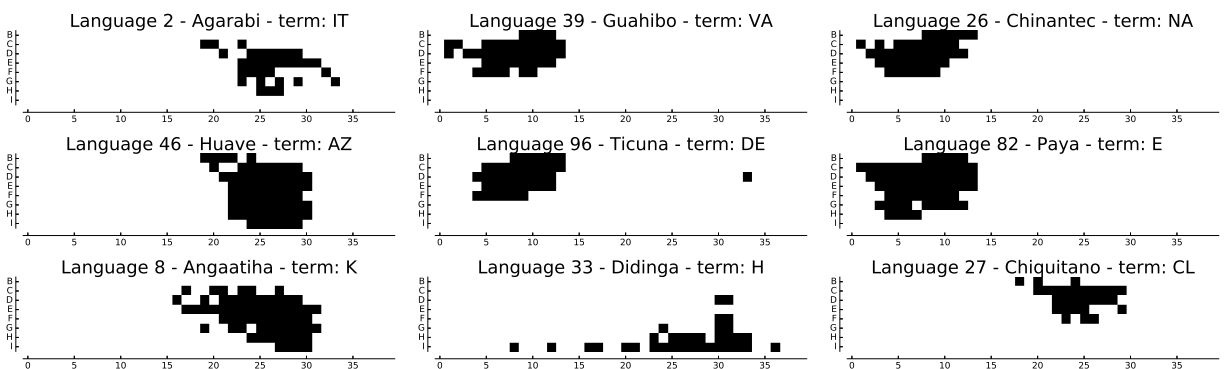


Figure B.5: WCS categories in the 75th percentile of unusual scores (presented in ascending ranked order of average Hausdorff distance)