ABSTRACT

Title of dissertation: INFORMATION AND INCREMENTALITY IN SYNTACTIC BOOTSTRAPPING

Aaron Steven White, Doctor of Philosophy, 2015

Dissertation directed by: Professor Valentine Hacquard
Department of Linguistics

Some words are harder to learn than others. For instance, action verbs like *run* and *hit* are learned earlier than propositional attitude verbs like *think* and *want*. One reason *think* and *want* might be learned later is that, whereas we can see and hear running and hitting, we can't see or hear thinking and wanting. Children nevertheless learn these verbs, so a route other than the senses must exist. There is mounting evidence that this route involves, in large part, inferences based on the distribution of syntactic contexts a propositional attitude verb occurs in—a process known as *syntactic bootstrapping*. This fact makes the domain of propositional attitude verbs a prime proving ground for models of syntactic bootstrapping.

With this in mind, this dissertation has two goals: on the one hand, it aims to construct a computational model of syntactic bootstrapping; on the other, it aims to use this model to investigate the limits on the amount of information about propositional attitude verb meanings that can be gleaned from syntactic distributions. I show throughout the dissertation that these goals are mutually supportive.

In Chapter 1, I set out the main problems that drive the investigation. In

Chapters 2 and 3, I use both psycholinguistic experiments and computational modeling to establish that there is a significant amount of semantic information carried in both participants' syntactic acceptability judgments and syntactic distributions in corpora. To investigate the nature of this relationship I develop two computational models: (i) a nonnegative model of (semantic-to-syntactic) projection and (ii) a nonnegative model of syntactic bootstrapping. In Chapter 4, I use a novel variant of the Human Simulation Paradigm to show that the information carried in syntactic distribution is actually utilized by (simulated) learners. In Chapter 5, I present a proposal for how to solve a standing problem in how syntactic bootstrapping accounts for certain kinds of cross-linguistic variation. And in Chapter 6, I conclude with some future directions for this work.

INFORMATION AND INCREMENTALITY IN SYNTACTIC BOOTSTRAPPING

by

Aaron Steven White

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Valentine Hacquard, Chair/Advisor
Professor Jeffrey Lidz, Co-Advisor
Professor Philip Resnik
Professor Naomi Feldman
Professor Hal Daumé III

Dedication

To my grandparents Sally and Ken White, who instilled in me from a young age a deep curiosity about the natural world. To my parents Annette and Steve White, who unerringly supported me through the course of my career as a student. And to my partner Bonnie O'Keefe, who made these six years at Maryland possible.

Acknowledgments

Finishing this dissertation is like waking from a dream *in media res*. Its beginnings are hard to discern; its course left no room for a waking world; and its narrative threads—only partially tied off—imply a thousand branching paths I still hope to realize. Here, I'd like to thank some of the many people that played a role in winding those narrative threads.

First and foremost, my sincere gratitude goes to my committee members: Valentine Hacquard, Jeff Lidz, Philip Resnik, Hal Daumé III, and Naomi Feldman. This dissertation is an exercise in interdisciplinary bridge-building, and these five—each a master engineer in their own right—were gracious enough to inspect my work and tell me when a span was structurally unsound.

Valentine and Jeff deserve my deepest thanks in this right. It is very, very hard to do insightful research. Not only do both do what I consider to be some of the most insightful in semantics and language acquisition, but they are both able to demonstrate *how* to do it. I can only hope that this dissertation does some justice to their demonstration.

Beyond their insight, Valentine and Jeff—and particularly Valentine—deserve thanks for their patience and generosity. Many a time did I bring them half-baked ideas, which they would patiently deconstruct. Most of the time very little was left after this deconstruction, but when there was something left, they were able to extract the kernel of the idea clean of the cruft.

I'd similarly like to thank Philip for his patience and generosity. This disser-

tation is, to some extent, about bringing ideas from linguistic theory down to earth. For this to be interesting, there needs to be a reason to bring those ideas down to earth in the first place. Philip has a unique ability to see when there is reason and when there isn't.

In meetings with colloquium speakers and other visitors, I've often heard that it seems like Maryland students are advised by the entire faculty. While Valentine and Jeff are certainly the most central figures in my graduate career, this really is true. In this respect, Alexander Williams and Norbert Hornstein deserve to be singled out. Alexander brings a deep subtlety of thought to every problem I have seen him tackle; he also has great taste in music. Besides Valentine and Jeff, Norbert did the most to show me what the important questions in the field are—in some cases, not without much convincing. I'll miss my semi-weekly impromptu meetings with him, in which he would wander into my office, pose some deep theoretical question—often one that would soon be posted to his blog Faculty of Language—and then discuss it with me and whoever else was in my office (often Dan Parker or Dustin Alfonso Chacón) for an hour or more.

Outside of the Maryland faculty, my deepest gratitude goes to Pranav Anand. This dissertation wouldn't exist without Pranav's mentorship. Pranav introduced me to both the syntactic bootstrapping literature—and thus Jeff's work—and the attitude verbs literature—and thus Valentine's work—in my final year as an undergraduate at UC Santa Cruz (2008-2009)—the former in his Foundations of Linguistic Theory class (Fall 2008) and the latter in his Semantics 3 class (Winter 2009). It was Pranav's mentorship that led me to apply to Maryland in the first place.

Many of the ideas in this dissertation were developed in conversation with the attitude verbs group—Valentine Hacquard, Jeff Lidz, Erin Eaker, Shevaun Lewis, Kate Harrigan, Rachel Dudley, Naho Orita, Morgan Moyer, and Tom Roberts—as well as various visitors who were gracious enough to have discussions with us: Jane Grimshaw, Pranav Anand, Jill de Villiers, John Grinstead, Chris Kennedy, Keir Moulton, Paul Portner, Aynat Rubinstein, Meredith Rowe, Florian Schwarz, Noah Goodman, Jesse Snedeker, and Elizabeth Bogal-Allbritten.

My cohort—Kate Harrigan, Dustin Alfonso Chacón, Naho Orita, and Sayaka Funakoshi-Goto—were also integral in developing these ideas in less direct ways. Kate has helped me remember that there are aspects of life outside academia. I'm not sure how she put up with me for five years, but I'm grateful that she did, because I can't imagine having made it through graduate school without her friendship. Dustin is a brilliant linguist and a good friend who was somehow able to share an office with me while dissertating, tolerating my random outbursts and often joining in my silliness.

Prior to Dustin, I shared an office with Dan Parker—off whom many an idea was bounced, by whom many a chillwave track was bumped, and with whom many a guitar jam was perpetrated—as well as Sol Lago—whose favorite pastime was to be exasperated by said chillwave tracks (or maybe just Dan and me more generally). Dan and I were perpetually blasting off into the brotosphere, and Sol would sometimes join as honorary brodette. I really loved sharing an office with both Dan and Sol.

For a very long time, I was convinced that Shevaun Lewis hated me. I am

now convinced that is not true and realize that what made me think that is exactly what I love about her personality. Shevaun is an extremely incisive thinker, and I'm happy that I'll be spending a year in her company at Johns Hopkins. My first experience of homebrewing was at Shevaun and Brad Larson's place, and though it took a few years to finally get going, Quincy St. Brewery owes its existence to their influence, as it does to Assistant Brewmaster, Peter Enns, and Chief Packaging Officer, Alix Kowalski. (The other position in the company, head of I-don't-drink-beer-so-why-would-I-drink-homebrew, is held by Zoe Schlueter.)

Mike Fetters is a damn cool dude with some damn cool tattoos (and a bad influence in this regard). He also knows a ton about ellipsis, and some day, we may finally get one of the papers we've been working on together out.

As far as I can tell, it's customary for acknowledgment of non-academics to come near the end (despite the often non-academic nature of the preceding thanks). I've followed convention here, but this should not be taken to imply depth of gratitude. I'm deeply grateful to Bonnie O'Keefe, who has seen me through all six years of my time at Maryland. I can't imagine that dating an academic is the most pleasant experience, but Bonnie has taken it in stride.

My final thanks (to flesh and blood people) go to my parents—Annette White and Steve White—my brother—Austin White—and my grandparents—Sally White, Ken White, Fred Roznowski, and Mary Roznowski. None of this would be possible without their unerring support.

Much of this dissertation would not have been possible without support from the National Science Foundation (NSF). The experiments in Chapter 2 were funded

by an NSF BCS grant (1124338); the work in Chapters 3 and 4 was funded by an NSF doctoral dissertation improvement grant (1456013); and I was funded by an NSF DGE IGERT grant (0801465) from September 2011-August 2013. Qualia Coffee, Three Stars Brewing Company, and Kevin's Printing funded my sanity.

Table of Contents

List of Tables xvi

List of Figures xviii

1 Introduction 1

1.1 Contextual information 1

1.1.1 The problem of observability 2

1.1.2 The problem of multi-faceted meanings 4

1.1.3 Solving the two problems 6

1.1.3.1 Syntactic context and two problems 13

1.2 Propositional attitude verb syntax and semantics 17

1.2.1 Representationality 17

1.2.2 Factivity 20

1.2.3 Assertivity 22

1.2.4 Communicativity 24

1.3 Beyond pen and paper . . . 26
1.4 Discussion and roadmap . . . 29

2 A computational model of projection 34
2.1 The meaning-syntax relationship . . . 36
2.1.1 Linguistically relevant meaning . . . 36
2.1.2 Discussion . . . 46
2.2 Experiment 1: verb-frame acceptability . . . 48
2.2.1 Design . . . 48
2.2.1.1 Features of interest . . . 49
2.2.1.2 Stimulus construction . . . 52
2.2.2 Participants . . . 53
2.2.3 Data validation . . . 53
2.2.4 Results . . . 55
2.2.4.1 A bird's eye view . . . 57
2.2.4.2 Recasting the Overlap Problem . . . 61
2.2.5 Principal Component Analysis . . . 63
2.2.6 Non-negative projection model . . . 77
2.2.6.1 Noise model . . . 77
2.2.6.2 Stopping criterion . . . 78
2.2.6.3 Model fitting . . . 80
2.2.6.4 Results . . . 80
2.2.7 Discussion . . . 85

2.3 Experiments 2 & 3: verb similarity 89
2.3.1 Experiment 2: generalized semantic discrimination task 91
2.3.1.1 Design 91
2.3.1.2 Participants 91
2.3.1.3 Data validation 92
2.3.1.4 Results 93
2.3.2 Experiment 3: ordinal similarity 97
2.3.2.1 Design 97
2.3.2.2 Participants 98
2.3.2.3 Data validation 98
2.3.2.4 Results 98
2.3.3 Comparison of (dis)similarity datasets 101
2.3.4 Discussion 104
2.4 Quantifying the syntax-semantic connection 105
2.4.1 Basic correlational analysis 107
2.4.2 Distributional features and similarity 109
2.4.3 Multinomial logit mixed model 113
2.4.3.1 Model comparison 114
2.4.3.2 Feature weights 115
2.4.3.3 Frame informativity 115
2.4.4 Ordinal logit mixed model 122
2.4.4.1 Model comparison 123
2.4.4.2 Feature weights 126

2.4.4.3 Frame informativity . . . 127
2.4.5 Discussion . . . 128
2.5 General discussion . . . 131

3 A computational model of syntactic bootstrapping 133
3.1 Computational models and syntactic distribution . . . 136
3.1.1 Category models . . . 136
3.1.1.1 Prior approaches . . . 136
3.1.1.2 Representational assumptions . . . 140
3.1.2 Vector space models . . . 145
3.2 The model . . . 149
3.2.1 Batch learner . . . 154
3.2.1.1 A useful conditional conjugacy . . . 154
3.2.2 Factor analysis-based smoothing . . . 161
3.3 Experiment . . . 163
3.3.1 Data . . . 163
3.3.1.1 Filtering . . . 167
3.3.2 Hybrid sampler/optimizer design . . . 169
3.3.2.1 MLE pre-training (**D** only) . . . 171
3.3.2.2 MAP pre-training (**S** and **P**) . . . 172
3.3.2.3 Training (**D**, **P**, and **S**) . . . 175
3.3.3 Results . . . 176
3.3.3.1 Basic correlational analysis . . . 176

3.3.3.2 Model analysis . . . 177
3.4 Discussion . . . 179
4 Incrementality in syntactic bootstrapping 180
4.1 The human simulation paradigm . . . 183
4.1.1 The classic paradigm . . . 183
4.1.2 Norming HSP with HSP . . . 185
4.1.3 Spatial HSP . . . 187
4.2 Norming tasks . . . 189
4.2.1 Human simulation norming . . . 190
4.2.1.1 Design . . . 190
4.2.1.2 Materials . . . 191
4.2.1.3 Participants . . . 196
4.2.1.4 Data validation . . . 197
4.2.1.5 Results . . . 199
4.2.1.6 Discussion . . . 216
4.2.2 Similarity norming . . . 216
4.2.2.1 Design . . . 217
4.2.2.2 Participants . . . 217
4.2.2.3 Data validation . . . 218
4.2.2.4 Results . . . 219
4.2.2.5 Discussion . . . 226
4.3 Spatial human simulation . . . 226

4.3.1 Design . . . 226
4.3.2 Materials . . . 227
4.3.3 Participants . . . 228
4.3.4 Data validation . . . 229
4.3.5 Results . . . 232
4.3.5.1 Fixed effects . . . 235
4.3.5.2 Random effects . . . 238
4.3.5.3 Relationship to known words . . . 239
4.4 Discussion . . . 240

5 A strategy for solving the labeling problem 242
5.1 The representational-preferential distinction . . . 244
5.1.1 Representationals and preferentials in English . . . 244
5.1.2 Representationals and preferentials outside English . . . 246
5.1.3 Main clause syntax . . . 248
5.2 Leveraging main clause syntax . . . 251
5.3 Experiment . . . 253
5.3.1 Data . . . 253
5.3.2 Model fitting . . . 254
5.4 Results . . . 254
5.5 Discussion . . . 256

6 Conclusion 258
6.1 Future directions . . . 262

6.1.1 Quantifying meanings . . . 262
6.1.2 Mapping from syntax to meaning . . . 264
6.1.3 Finer-grained incremental conjectures . . . 266
6.1.4 Main clause syntax and beyond . . . 267

A Appendix A 268
A.1 Figures and tables . . . 268
A.2 Non-negative projection model . . . 269
A.2.1 Parametric binary feature model . . . 269
A.2.2 Nonparametric binary feature model . . . 269
A.2.3 Projection principles (feature loading) model . . . 270
A.3 Response models . . . 270
A.3.1 Ordinal logit mixed model . . . 270
A.3.2 Multinomial logit mixed model . . . 273
A.3.2.1 Distribution of bias . . . 273

B Appendix B 275
B.1 Generative story for IBP prior . . . 275
B.2 Sampler derivation . . . 276
B.2.1 Inference equations . . . 279
B.2.1.1 Inferring **D** . . . 279
B.2.1.2 Inferring **P** . . . 284
B.2.1.3 Inferring **S** . . . 287

Bibliography 296

List of Tables

2.1 Pairs rated more highly in the generalized semantic discrimination task than in the likert scale task. 104

2.2 Model comparison measures for multinomial logit mixed model. The minimum values for AIC and BIC are bolded. 115

2.3 Model comparison measures for ordinal logit mixed model. The minimum values for AIC and BIC are bolded. 124

3.1 Spearman rank correlation between Jensen-Shannon divergence derived from three different datasets and similarity judgments. P-values derived from Mantel (permutation) test with 10000 iterations. 176

4.1 Fixed effects for mixed effects logistic regression accuracy model. . . . 202

4.2 Fixed effects for mixed effects logistic regression accuracy model with the addition of the true word's log frequency as a predictor. 205

4.3 Fixed effects for mixed effects logistic regression nonverb model. . . . 208

4.4 Variable importance as measured by mean decrease in Gini 210

4.5 Fixed effects for linear mixed effects similarity model. 223

4.6 Variable importance as measured by increase in node purity predicting inaccurate verb response similarity judgments 225

4.7 Fixed effects of mixed effects logistic regression with random intercepts for participant and verb. The reference level is LEXICAL CONTEXT:*nonce* x INFORMATIVITY:*low* x TRAINING SIZE:*small*. 231

4.8 Fixed effects of mixed effects regression with random intercepts for participant and verb. The reference level is LEXICAL CONTEXT:*nonce* × INFORMATIVITY:*low* × TRAINING SIZE:*small*. (As is standard for continuous variables, the intercept represents TRUE WORD SIMILARITY: 0) 236

A.1 Pairs rated more highly in the likert scale task than in the generalized discrimination task. 269

List of Figures

2.1 graphical model for generative model corresponding to $S \xrightarrow{\mathrm{P}} D \rightarrow X$. 47

2.2 Mean rating for each verb-frame pair ordered by hierarchical clustering. Darker shades represent higher mean ratings. 56

2.3 Hierarchical clustering of verbs based on data in Figure 2.2. 58

2.4 Verb embeddings on first and second principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$. Four verbs—*amaze*, *bother*, *worry*, and *tell*—are missing from this diagram due to their extreme values on these components. They lie far to the upper left. 64

2.5 Frame loadings on first and second principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$. 65

2.6 Verb embeddings on third and fourth principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$. 69

2.7 Frame loadings on third and fourth principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$. 70

2.8 Graphical model corresponding to $S \xrightarrow{\text{P}} D \rightarrow X$. (Same as Figure 2.1.) 72

2.9 Graphical model corresponding to $S \xrightarrow{\text{P}} D \rightarrow X$ with the addition of the ordinal logit mixed model parameters $\mathbf{g}$. 78

2.10 Log Pointwise Predictive Density (LPPD) and Watanabe Akaike (Widely Applicable) Information Criterion—both scaled by -2—for models with different numbers of features. The gap between the LPPD and WAIC lines is proportional to the effective number of parameters as computed by Gelman et al.'s (2013) second method (p_{WAIC2}). 81

2.11 Verb features ($\mathbf{S}$) inferred by non-negative projection model. Black cells represent 1s. 85

2.12 Relationship between features and syntactic frames ($\mathbf{P}^\top$) inferred by non-negative projection model. Darker cells represent larger values. . 86

2.13 Similarity rating for each verb-verb pair from generalized semantic discrimination experiment. Darker shades represent more times chosen similar. Note that the diagonal elements are not observed and are set to the maximum over all other cells. 93

2.14 Embedding derived by two-dimensional nonmetric multidimensional scaling applied to the generalized semantic discrimination judgments represented in Figure 2.13. 95

2.15 Similarity rating for each verb-verb pair from ordinal scale experiment. Darker shades represent higher mean ratings. Note that the diagonal elements are not observed and are set to the maximum over all other cells. 99

2.16 Embedding derived by two-dimensional nonmetric multidimensional scaling applied to the ordinal scale judgments represented in Figure 2.15. 101

2.17 Relationship between generalized semantic discrimination similarity responses and ordinal scale similarity responses. Only low outlier pairs are labeled. (See Table A.1 in Appendix A for high outlier pairs.) 102

2.18 Relationship between frame distances based on acceptability judgment data present in Section 2.2 and dissimilarities based on the similarity judgment data presented in Section 2.3. Lines show local regression fits. 107

2.19 Feature weights for multinomial logit mixed model with weighted diffusion kernel. Features correspond to those in Figure 2.11. 117

2.20 Unweighted (black) and weighted (black+grey) Gini computed using feature weights from multinomial logit mixed model with weighted diffusion kernel (Figure 2.19) and projection principles inferred from non-negative projection model (Figure 2.12). 120

2.21 Weighted frame combination Gini residualized by beta regression on weighted Gini of each frame in combination (full bars in Figure 2.20). Grey represents a positive residual and orange, a negative. 121

2.22 Feature weights for ordinal logit mixed model with weighted linear kernel. 126

2.23 Unweighted (black) and weighted (black+grey) Gini computed using feature weights from ordinal logit mixed model with weighted linear kernel (Figure 2.22) and projection principles inferred from non-negative projection model (Figure 2.12). 128

2.24 Weighted frame combination Gini residualized by beta regression on weighted Gini of each frame in combination (full bars in Figure 2.20). Grey represents a positive residual and orange, a negative. 129

3.1 Plate diagram for Latent Dirichlet Allocation (LDA) 139

3.2 Plate diagram for non-negative projection model. 153

3.3 Likelihood of d given x and δ. The value of δ is given as the title of each facet. 158

3.4 Count v. rank of particular verbs. Black points show verbs that were kept. 169

3.5 Distribution of normalized pointwise mutual information of verb and embedded clause across verbs. Black bars show verbs that were kept at embedders. 170

3.6 Log of verb-subcategorization frame counts plus 1. White represents 0 and grey is scaled with the log count. 173

3.7 Log of $\mathbf{D}$. White represents values closer to $-\infty$ and darkest grey represents least negative values. 173

3.8 Features extracted from PukWaC dataset using nonnegative projection model of syntactic bootstrapping. 178

4.1 Distribution of by-item accuracy given verb, lexical context, and context of utterance. Each box represents the distribution of accuracy over the 20 items in that condition. 201

4.2 Distribution of by-item proportion of nonverb responses. Verbs are ordered as in Figure 4.1 (by median accuracy). 207

4.3 Distribution of by-item response root relative frequencies given verb, lexical context, and context of utterance. Each bar+error bar represents the distribution of relative frequency for the labeled response root over the 20 items in that condition. 213

4.4 Distribution of ridit-scored similarity across items with accurate items set to 1. 220

4.5 Distribution of ridit-scored similarity across items with accurate items set to 1. 221

A.1 Hierarchical clustering of frames based on data in Figure 2.2. 268

A.2 Distribution over participants of size of acceptability interval mapped to each rating. 272

A.3 Distribution over participants of size of similarity interval mapped to each rating. 272

A.4 Distribution over participants of bias for response based on placement in response list. 274

Chapter 1: Introduction

1.1 Contextual information

Words have meanings. Those meanings must be learned. Learning the meaning of some words seems like it could be quite easy. For the moment, assume that learning a word-meaning—e.g the meaning of the word *dog*—involves pairing some concept or set—e.g. the dog concept or set of dogs, call either DOG—with some linguistic symbol: *dog*. How one goes about doing this, the intuitive story goes, is by noticing that utterances involving the word *dog* cooccur with instantiations of the concept/set DOG quite often and thus an association between the word and the concept is built. And just so, the meaning of *dog* is learned. In this scenario, the learner's ability to discover correlations between the language and the *nonlinguistic context*—the perceivable objects and events surrounding the hearer—is paramount. Left unelaborated, this story has well-known problems (cf. Goodman, 1955; Quine, 1960; Kripke, 1982): why not consider subparts of the dog (TAIL, HEAD)? Superordinate categories properly containing the dogs (MAMMAL, ANIMAL)? Or DOG at the time of utterance, CAT every other time?

Nonetheless, few would deny that nonlinguistic context plays a major role in learning the meanings of at least some—maybe most—words. How else *would* a

learner figure out that *dog* means DOG? It has become clear that solving these problems requires understanding both the nature of human conceptual understanding—how the learner conceptualizes the nonlinguistic context—and the structure of the mechanism that learners use to link words with concepts—how the learner extracts information from nonlinguistic context. Within the latter vein, there have been many interesting proposals: some involving empirically motivated learning biases that might direct the learner toward the correct concepts (cf. Carey and Bartlett, 1978; Markman and Hutchinson, 1984; Markman and Wachtel, 1988; Merriman and Bowman, 1989; Markman, 1990, a.o.) and others that rely on more general properties of inductive reasoning that might direct the learner toward the correct concepts (cf. Xu and Tenenbaum, 2007; Frank et al., 2009, a.o.). For instance, maybe, as Markman and Wachtel (1988) suggest, children prefer to map words to whole objects—DOG is more salient as a meaning for *dog* than TAIL or HEAD—or maybe they assume that concepts with smaller extensions should be preferred to ones with larger extensions (Tenenbaum and Griffiths, 2001)—a sort of weighted Subset Principle (Berwick, 1985).

1.1.1 The problem of observability

But even equipped with these kinds of proposals, learning the meanings of other words seems like it is probably quite a bit harder. For example, how do learners acquire those words whose meanings are not obviously linked with features of the nonlinguistic context—or more precisely, participants conceptualizations of

these contexts? The parade case of such words—what Gleitman (1990) refers to as words with meanings "closed to observation" and which Gleitman et al. (2005) dub the *hard words*—are those that refer to abstract objects/concepts (*liberty*, *tyranny*), mental states (*think*, *know*), preferences (*want*, *prefer*), authorizations (*allow*, *forbid*), etc. One property that binds many of these words together is that many are verbs involving *propositional attitudes*, which express relations to ways the world might be, in fact is, would be best if it were, etc. It is these hard words that this dissertation focuses in on.

The problem with these hard words is that one can't very well see, hear, or feel propositional attitudes like thinkings or wantings, so it is quite unclear how the learner pairs up words for these attitudes—*think* or *want*—with the appropriate concepts—for now, call them THINK and WANT—under an account where correlations between particular words and nonlinguistic context are the primary (or only) data (Landau and Gleitman, 1985; Gleitman, 1990).

There is now a wealth of experimental results evidencing the magnitude of this problem. One particular instance of this can be found in work within the Human Simulation Paradigm (HSP; Gillette et al. 1999; Snedeker and Gleitman 2004)—discussed at length and deployed in Chapter 4. In one instantiation of this paradigm, adult participants are given videos of parents playing with their children. In these videos the sound has been removed, with the idea that this partially replicates the learner's nonlinguistic context. A beep is then placed where a target word was uttered, and participants are asked to say what the word is. Accuracy is quite high in recovering concrete nouns, like *dog*, but essentially zero in recovering mental state

verbs.

What this suggests is that—even for adults, who are constantly talking about desires and beliefs—desires and beliefs are just not salient as potential word meanings from the nonlinguistic context alone. And indeed, further work within this paradigm suggests that, even if scenes are constructed to make propositional attitudes salient, gains from nonlinguistic context alone are only modest (Papafragou et al., 2007).

1.1.2 The problem of multi-faceted meanings

This problem of observability is sharpened by the fact that these words also tend to have meanings that are multifaceted. For instance, one facet of the meanings of both *think* and *know* is that they involve beliefs in some important way.

(1) a. Bo thinks that Jo is out of town.

b. Bo knows that Jo is out of town.

These words clearly don't mean the same thing, though. They have (at least) a second facet to their meaning on which they differ. In saying (1b), a speaker presupposes something very specific about what they and their conversational partners have (typically) already accepted as true—namely, that (2), corresponding to the content of *know*'s subordinate clause, is also true.

(2) Jo is out of town.

This presupposition furthermore *projects* through—i.e. is unaffected by—various semantic operators, such as negation (3a) and questioning (3b) (Kiparsky and Kiparsky, 1970; Karttunen, 1971; Horn, 1972; Karttunen and Peters, 1979). Both (3a) and (3b) show the same behavior as (1b) in requiring the speaker to presuppose the truth of (2).

(3) a. Bo doesn't know that Jo is out of town.

b. Does Bo know that Jo is out of town?

This is certainly not the case with (1a). In uttering (1a), a speaker has no commitments—as far as the meaning of the sentence is concerned—with respect to whether (2) is true. Indeed, one can easily imagine a discourse in which (1a) is uttered as a justification for a behavior that is based on mistaken premises. Maybe Jo is Bo's boss and, incorrectly believing that Jo has left for the day, he is packing up early. A coworker who knows that Jo isn't out of town might well say (1a) to explain Bo's behavior, but it would be very odd to say (1b).

And similarly, the negative (4a) and questioned (4b) versions of (1a) do not show the projection behavior seen with (3a) and (3b).

(4) a. Bo doesn't think that Jo is out of town.

b. Does Bo think that Jo is out of town?

In fact, rather than leaving the content of the subordinate clause alone, as it does when it is attached to *know*, the negation attached to *think* in (4a) seems to *reach down* into that content. A natural interpretation of (4a) is (5), where the negation

has "lowered" into the subordinate clause.[1]

(5) Bo thinks that Jo isn't out of town.

Thus, beyond the fact that states like thinkings and knowings are not salient in the nonlinguistic context, whatever it means to learn the words *think* and *know* (and *want* and *prefer* and *allow* and *forbid*), it doesn't seem so simple, on the face of it, as linking *think* with some concept THINK and *know* with some concept KNOW.

1.1.3 Solving the two problems

But then how *do* learners acquire this constellation of facts about even just these two verbs—*think* and *know*? How do they figure out, on the one hand, that both *think* and *know* share a facet of their meaning in that they both involve beliefs in some crucial way? And on the other hand, how do they figure out that *know* and *think* differ with respect to other facet(s) of their meaning—e.g. that *know* (i) requires the content of its complement to be presupposed and (ii) protects that content from interference by negation and questions, while *think* (i) does not require its content to be presupposed and (ii) allows its content to be interfered with by the likes of negation?

The now standard answer, at least at a broad level, is that learners need to move beyond nonlinguistic context as their only source of evidence for word-learning. They need to furthermore incorporate a word's *linguistic context*. To understand

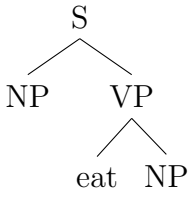
[1] Or perhaps it has *neg-raised* from the subordinate clause to the matrix clause; cf. discussion in Fillmore 1963; Horn 1971, 1975, 1978, 1989; Bartsch 1973; Ross 1973; Prince 1976; Gajewski 2007, a.o.

what this means, it is useful to step back and consider a rough but useful division that exists in the literature between two broad types of linguistic context: *lexical context* and *syntactic context*.

To a first approximation, lexical context encompasses words that cooccur with the one being learned, and syntactic context encompasses the types of abstract structures the word is found in. For instance, suppose a learner received the datum in (6). The lexical context of *eat* might be represented as in (6a). And assuming the learner can parse the string in an adult-like way, its syntactic context might be represented as in (6b).

(6) The men eat apples.

a. {men, the, apples}

b. 

If linguistic context is necessary for learning words like *think*, *know*, and *want*–i.e. those whose meanings are not associated with sensory correlates—which subtype of information—lexical context or syntactic context—might be used? The likely answer to this is that both are necessary in interaction, to different extents, for different kinds of verbs. However, authors differ on whether syntactic context could ever be necessary independent of lexical context. For instance, Pinker (1994) and Grimshaw (1994) both argue that the only sense in which syntactic context might be useful is in interaction with lexical context. Knowing that *eat* tends to take NPs referring

to edibles, like *apples*, in object position could plausibly help a learner figure out that *eat* means EAT. But most of the information in this case is coming from a semantic generalization—commonly occurs with words referring to edibles—and so it is unclear what further work, if any, the syntactic context is doing here. Indeed, if the learner can (i) figure out what role the referent of each noun phrase plays in the event named by the verb—via only the semantics of those noun phrases—and (ii) there are some general constraints regarding where each argument may be syntactically situated given its entailments (cf. Baker, 1988; Grimshaw, 1990; Dowty, 1991), the syntactic context may be doing little work beyond highlighting the lexical material on which the learning mechanism should make its inferences (cf. Connor et al., 2013).

I take this to be a reasonable position for verbs like *eat*, whose syntactic contexts likely contains very little information about its meaning beyond that it involves two participants. (In fact, due to the presence of intransitive uses such as *the men ate*, unilateral reliance on the syntax might even lead learners astray.) Less clear is whether this strategy of seeding inference with only lexical context could be extended to all verbs—especially propositional attitude verbs like *think*, *know*, and *want*. For instance, it seems unlikely that a learner could glean much at all about the meaning of *want* from the distribution of nouns it occurs with, since *want* imposes few to no restrictions on its direct object's meaning (cf. Resnik, 1996, p. 138, Table 1). And this is not specific to *want*; many propositional attitude verbs that allow NP direct objects do not constrain the semantics of those objects.

(7) Bo wants {an apple, a toy, a back rub}.

(8) Bo {knows, remembers, needs, demands} {a doctor, a story, the time}.

This leaves two alternatives: abandon linguistic context as a necessary condition for learning the hard words—maybe it was brash to reject nonlinguistic context so quickly after all—or assume that not all word learning relies chiefly on either lexical or nonlinguistic context. It is quite unclear how the former route could work; but even granting that a learner could learn the meaning of *want* without recourse to some amount of linguistic context, telling a story about how learners go on to distinguish *want* from, e.g., *hope* will likely be difficult.

To see this, note that hopings seem to involve wantings—if (9a) is true, (9b) must also be true—but *hope* seems to have an extra facet of its meaning over and above the component it shares with *want*. If (10) is true, (9a) is very odd, whereas (9b) is fine. Thus *hope* seems to place an extra constraint on the states that it can describe—namely, that the hoper believe that the state of affairs they hope to come about is also possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013; Hacquard, 2014; Harrigan, 2015).

(9) a. Bo hopes to dance with Jo.

b. Bo wants to dance with Jo.

(10) Bo believes he'll never dance with Jo.

This is similar to the previous example involving *think* and *know*. Remember that *think* and *know*, like *want* and *hope*, share a component of their meaning involving

belief, but they differ in the sorts of contexts they are allowed in. This suggests a potentially quite general problem pertaining to propositional attitude verbs: even if the nonlinguistic context is sufficient to learn the meaning of, e.g., *want* or *think*, a learner who also posits the meaning of *want* for the meaning of *hope*, or the meaning *think* for the meaning of *know*, might very well find herself with a subset problem (Wexler and Hamburger, 1973; Baker, 1979; Berwick, 1985; Pinker, 1989); *want* will always be true in the contexts *hope* is true, and *know* will always be true in contexts where *think* is true, so it's hard to see how a learner who posited the meaning *want* for the meaning *hope* might change her mind given the nonlinguistic context alone.

Maybe one of the learning bias accounts could be made to block this from the start. For instance, Markman and Wachtel (1988) show that learners are biased to assume that word meanings are mutually exclusive: if they already have a word associated with a particular meaning, they are biased to not associate a new word with that meaning. So for the sake of argument, assume that a learner has acquired *want* correctly—a plausible assumption given that *want* is almost two orders of magnitude more frequent that *hope*—then she might be disinclined to posit the meaning of *want* for that word, since she already has a word that corresponds to that meaning.

And just so. But the learner is not out of the weeds yet. As noted earlier, since x *hopes* p also seems to entail that x believes that p is possible and there does not seem to be an English word meaning *believe possible* a mutual exclusivity account alone will not solve all possible subset problems here. Indeed, if the learner is biased to posit simpler lexical meanings, the *believe possible* meaning would be

incorrectly preferred absent some further mechanism. Maybe the Size Principle can be leveraged here: *hope*, in having a more complex meaning, plausibly constrains the situations it is compatible with more than *want* or the non-attested *believe-possible*. Then, it might be preferred over these alternatives based on a "suspicious coincidence."

Indeed, this strategy might be generalizable: just as *hope p* entails *want p*, *know p* entails *think p*. And in each case, the former places a further constraint on the states it describes. Maybe, then, learners could use aspects of the conversation at hand—a special sort of context that straddles both nonlinguistic and linguistic context that I refer to as the *discourse context*—to figure out that *hope* is only used when describing situations in which the hoper believes the hoped state of affairs to be possible and that *know* is only used when describing situations in which all conversational participants take the known state of affairs for granted.

As noted by Hacquard (2014) and Dudley et al. (2015), this will likely be a tough row to hoe, at least in the case of *know* v. *think*. *Think* has a property—shared by many attitude verbs that Hooper (1975) dubs assertives (discussed in more detail below)—that it can be (and often is) used to contribute not (just) a report of someone's mental state but also the content of that mental state. For instance, adapting an example from Simons (2007) (cf. Hacquard's 2014 example 4), the main point of Beth's utterance in (11) is not to contribute the mental state report but to proffer a possible answer to Anne's question.[2]

[2]Note that this is not to say that (11) does not involve a mental state report but rather that that mental state report is somehow secondary in this context.

(11) **Anne:** Why isn't Bo at the meeting?

Beth: Jo thinks he's out of town.

In this circumstance, if Beth is being cooperative, she certainly seems committed to Bo's being out of town at least plausibly being true. And indeed, many uses of *think* have this flavor, particularly within child-directed speech and children's production (Diessel and Tomasello, 2001). But if the complement of *think* is meant to be taken as at least plausibly true, how does the learner distinguish it from *know*?

Maybe the learner has some exquisite sensitivity to whether a particular proposition is taken for granted already or not, which is what seems to distinguish the *think* and *know* uses. The problem here is that there are cases of *know* in child-directed speech that patently could not involve all conversational participants taking the content of *know*'s subordinate clause for granted—for instance, cases where the speaker wishes to convey that something is true while also also querying the hearer's mental state, as in (12).

(12) Do you know that Daddy's coming home late tonight?

This suggests that the problem of multi-faceted meanings still rears its head, even once nonlinguistic context is elaborated with discourse context. How does a learner figure out that *want* and *hope* share some facet of their meaning involving desire but that only *hope* requires that the desirer further believe their desire is realizable? Or similarly, how does a learner figure out that *think* and *know* share some facet of their meaning involving belief but that only *know* requires the presupposition of its complement?

1.1.3.1 Syntactic context and two problems

This is all to reiterate that, for at least some distinctions among propositional attitude verbs, neither nonlinguistic context nor lexical context are likely to help in drawing fine-grained distinctions among propositional attitude verbs, and so a final possibility remains within the rough taxonomy given earlier: the sort of distinction I have been discussing must be learned using syntactic context. How might the learner do this?

The answer that Landau and Gleitman (1985) and Gleitman (1990) propose under the heading *syntactic bootstrapping* is that children use the syntactic contexts that a word cooccurs with—its *syntactic distribution*—to deduce its meaning.[3] The proposal moves forward in the following way. Suppose the learner makes the following assumption: the degree to which two verbs overlap in their syntactic contexts correlates with the degree to which their meanings overlap. She then notes that the syntactic contexts of *want* and *hope* only partially overlap. *Want*, but not *hope*, allows noun phrase objects (13a); both *want* and *hope* allow subjectless infinitival complements (13b) (control complements); and *hope*, but not *want*, allows tensed subordinate clause complements (13c).

(13) a. Bo {wants, *hopes} an apple.

[3]This is at least the accepted genealogy of the idea. As a historical note, this was actually proposed earlier in Lasnik 1989, a paper in the proceedings of the 1982 University of Western Ontario Learnability Workshop. The relevant quote: "...there appears to be a tacit assumption that the meaning of, e.g., a verb, can be presented and apprehended in isolation. But this seems implausible. Rather, verbs are presented in grammatical sentences which, therefore, explicitly display subcategorization properties. In fact, one might consider reversing the whole story: subcategorization is explicitly presented, and the child uses that information to deduce central aspects of the meaning of verbs" (p. 195, fn. 12).

b. Bo {wants, hopes} to have an apple.

c. Bo {*wants, hopes} that he will have an apple.

Finally, based on these data and the premise that overlap in syntactic contexts correlates with overlap in meaning, she might then infer that *want* and *hope* may share some facet(s) of their meanings, but not others.

Now, this on its own is insufficient. Knowing that there is an overlap in meaning does not yet indicate what that particular overlap is. Thus, this story needs to be augmented to explain not only how to find out whether there's an overlap, but also to say how that overlap gets labeled with the appropriate facet or feature of the words' meanings. One wants to know not only that *want* and *hope* share a meaning feature, but furthermore what that shared feature is. I refer to the first of these problems—finding out that there is an overlap—as the *clustering problem* because it deals with finding out that particular words cluster together with respect to the syntactic contexts they occur in. I refer to the second of these problems—finding out what facet of the meaning a particular clustering of verbs corresponds to—as the *labeling problem* because it deals with labeling the clusters of words that are found.[4]

The traditional solution to both problems within the syntactic bootstrapping literature is to assume that whatever mechanism solves the clustering problem simul-

[4]The terminological choices *clustering problem* and *labeling problem* belie a particular view of the problem a learner faces—namely, that at base, the learner must discover symbolic relationships among verbs' meanings. This does not preclude a learning model that imputes continuous representations to the learner, as long as those representations are somehow linked to symbolic representations. I do not intend any sleight of hand here, though; the tension between models that (explicitly) traffic in symbols (at some level) and those that do not (explicitly) are discussed at length in Chapters 2 and 3.

taneously solves the labeling problem by associating particular syntactic contexts with particular semantic features. Thus, under a standard syntactic bootstrapping account, the learning mechanism comes *pre-built* with "...grammatical knowledge [that] includes principles that provide a systematic mapping between semantic and syntactic structures" (Lidz et al., 2004, but see also the rich literature on this topic Fillmore 1970; Zwicky 1971; Jackendoff 1972; Grimshaw 1979; Pinker 1989; Levin 1993), and this systematic mapping is deployed to label clusters associated with particular syntactic features. As noted by Kako (1997) as well as Lidz et al. (2004), this might be cashed out in a couple different ways—either by imbuing syntactic contexts themselves with semantic content (the *Frame Semantic Hypothesis*) or by imbuing them with semantic content inherited from the verb's semantic features (the *Lexical Projection Hypothesis*)—but for current purposes what is important to note is that the principles are taken to be hard-coded.

This hard-coding assumption is problematic in the current context because it relies on an assumption that whatever these mapping principles look like, they are cross-linguistically universal. This is a reasonable assumption in the context of verbs that only occur with noun phrase and prepositional phrase arguments, like *hit* or *give*, since these verbs' occurrence in particular frames is quite stable across languages. But it is problematic if one turns to a domain, such as the propositional attitude verbs, where the syntactic contexts that particular subclasses of verbs occur in apparently vary quite wildly. I note where this variability occurs briefly in the next section and then again in Chapter 5.

This, among other considerations I lay out in Chapter 2, is one impetus for

investigating the clustering problem and the labeling problem separately in the domain of propositional attitude verbs, since it could well be that the solutions to these two problems lie in separate mechanisms. In Chapters 2, 3, and 4, I focus in on the clustering problem for attitude verbs, though along the way I note particular successes of my solution to this problem in terms of its ability to find clusters which the analyst might give a coherent labeling to. In Chapter 5, I give a suggestion for a solution to the labeling problem that takes advantage of both a property of the model I propose as well as a novel linguistic insight.

To set the stage for my solutions to these problems, it is useful to review what is already known about the correlation between semantic features and syntactic contexts in English. I carry this review out in the next section. This review has two main purposes: to establish (i) that there appear to be promising correlations between the syntax and the semantics that learners might take advantage of, but (ii) that these correlations are not perfect thus making it unclear to what extent they even could be relied on by a learner.

In the subsequent section, I note that this second point is the result of the imprecise nature of traditional distributional analysis. Results of this methodology are by necessity only suggestive for learning accounts due to the fact that traditional distributional analysis cannot be done *at scale*; it is not possible to ask for precise measures of the relationship between semantics and syntax that might validate or invalidate a syntactic bootstrapping approach to propositional attitude verb learning (or indeed word-learning more generally). Indeed, this is a more general problem for an approach to understanding lexical semantics—a problem that this dissertation

contributes a partial solution to.

1.2 Propositional attitude verb syntax and semantics

In the previous section, I briefly touched on three facets of propositional attitude verb meanings. I noted that *think* and *know* involve beliefs but that they differ with respect to whether their complement is presupposed: *know* is factive, and thus presupposes its complement, whereas *think* is nonfactive, and thus it does not. The other distinction I noted was that between *want* and *hope*, which both involve desires but which differ with respect to whether the referent of their subject must furthermore believe that state of the world is possible. *Hope* requires such a belief, whereas *want* does not. These examples, as one might expect, were not chosen at random: these four verbs exemplify two of four high-level semantic distinctions that appear in the literature to have some amount of correlation with the syntax. In the remainder of this section, I review these four distinctions.

1.2.1 Representationality

Perhaps the most well-known semantic distinction among propositional attitude verbs is that between verbs that express beliefs—or represent "mental pictures" or "judgments of truth" (Bolinger, 1968)—and those that express desires—or more generally, orderings on states of affairs induced by, e.g. commands, laws, preferences, etc. (Bolinger, 1968; Stalnaker, 1984; Farkas, 1985; Heim, 1992; Villalta, 2000, 2008; Anand and Hacquard, 2013, a.o.). Within the first class, which I henceforth refer

to as the representationals, fall verbs like *think* and *know*; and within the second class, which I henceforth refer to as the preferentials, fall verbs like *want* and *order*.

There appear to be various aspects of the syntactic distribution that roughly track this distinction in English. One well-known case is finiteness: representationals tend to allow finite subordinate clauses (1a) but not nonfinite ones (1b); preferentials tend to allow nonfinite subordinate clauses (2b) but not finite ones (2a).

(14) a. John thinks that Mary went to the store.

b. *John thinks Mary to go to the store.

(15) a. *John wants that Mary went to the store.

b. John wants Mary to go to the store.

There are two important things to note about this distinction. First, though the representationality distinction is often talked about as though it were mutually exclusive, some verbs appear to fall into both categories, and suggestively, show up in both frames. For instance, as noted in the last section, *hope p* involves both a desire that p come about and the belief that p is possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013; Hacquard, 2014; Harrigan, 2015, but see also Portner and Rubinstein 2013), and it occurs in both finite (16a) and nonfinite (16b) syntactic contexts.

(16) a. John hopes that Mary went to the store.

b. John hopes to go to the store.

Second, the link between representationality and finiteness is just a tendency. Some

verbs plausibly classed as representationals allow nonfinite subordinate clauses (17a)/(17b), and others plausibly classed as preferentials allow subordinate clauses that look finite (17c).[5]

The roughness of this correlation is perhaps not surprising since not all languages track representationality with tense: for instance, various Romance languages track the distinction with mood—representationals tending to take indicative mood and preferentials tending to take subjunctive mood (Bolinger, 1968; Hooper, 1975; Farkas, 1985; Portner, 1992; Giorgi and Pianesi, 1997; Giannakidou, 1997; Quer, 1998; Villalta, 2000, 2008, a.o.). I return to this cross-linguistic variability in detail in Chapter 5.

(17) a. John believes Mary to be intelligent.

b. John claims to be intelligent.

c. John demanded that Mary go to the store.

But though the correlation between representationality and tense is imperfect, even in English, finiteness does not appear to be the only associated syntactic (distributional) property. Also relevant appears to be a distinction in whether the verb's subordinate clause can be fronted—or in Ross's (1973) terms, S-lifted.[6] At least some representationals' subordinate clauses (18) appear to be able to undergo S-

[5]Whether (17c) involves a finite subordinate clause is to some extent dependent on whether what is often called the English subjunctive involves tense. On the one hand, the complementizer *that* is the same one that occurs with tensed subordinate clauses, but on the other, the verb shows up in its base (untensed) form.

[6]There is a further distinction in the literature made between S-lifts involving first person and third person propositional attitude verb subjects (Reinhart, 1983; Asher, 2000; Rooryck, 2001). I incorporate this first-third distinction into our experiment, but the data regarding this syntactic distinction are murky at best.

lifting, but many preferentials' subordinate clauses (19) cannot (Bolinger, 1968).

(18) Mary already went to the store, I {think, believe, suppose, hear, see}

(19) a. *John already went to the store, I {want, need, demand}.

b. *John to go to the store, I {want, need, order}.

(Not all representationals allow S-lifting. This is likely because the availability of S-lifting for a particular verb is conditioned by other semantic and pragmatic properties it has, so we defer further discussion of which verbs allow it until distinctions beyond representationality have been discussed.)

1.2.2 Factivity

The representationality distinction is cross-cut by another common distinction: factivity (Kiparsky and Kiparsky, 1970; Karttunen, 1971; Horn, 1972; Hooper, 1975). Factivity is defined in terms of its discourse effects. I noted these effects briefly in the last section in contrasting the verbs *think* and *know*, but very roughly, a verb is factive if upon uttering a sentence containing a factive verb with a subordinate clause, a speaker takes the content of the subordinate clause for granted regardless of propositional operators placed around the propositional attitude verb: in particular, negation (21b)/(20b) or questioning (21c)/(20c). For instance, each sentence in (20) commits the speaker to (22) being true, but modulo the context, the sentences in (21) do not. That is, in uttering the sentences in (20), the speaker presupposes (22) (Stalnaker, 1973). This suggests that *know*, *love*, and *hate* are factive, while *think*,

believe, and *say* are not.

(20) a. John {knew, loved, hated} that Mary went to the store.

b. John didn't {know, love, hate} that Mary went to the store.

c. Did John {know, love, hate} that Mary went to the store?

(21) a. John {thought, believed, said} that Mary went to the store.

b. John didn't {think, believe, say} that Mary went to the store.

c. Did John {think, believe, say} that Mary went to the store?

(22) Mary went to the store.

Factivity truly cross-cuts the representationality distinction in that there are verbs representing all four possible combinations: (i) representational (cognitive) factives, like *know*, *realize*, and *understand*, (ii) preferential (emotive) factives, like *love* and *hate*, (iii) representational nonfactives, like *think* and *say*, and (iv) preferential nonfactives, like *want* and *prefer*.[7]

The factivity distinction appears to be tracked most closely by whether the verb allows both question and nonquestion subordinate clauses (Hintikka, 1975; Ginzburg, 1995; Lahiri, 2002; Sæbø, 2007; Egré, 2008; Uegaki, 2012; Spector and Egré, 2014; Anand and Hacquard, 2014). For instance, the factive *know* can occur

[7]One question that arises here is whether, given the existence of representational+preferential verbs like *hope*, there could also be such representational+preferential factives. In a certain sense, this may be the case for the emotive factives, since it seems like sentences containing them imply that the holder of the emotion also believes the subordinate clause to be true. If all preferential factives are emotive (and show this behavior), this might suggest that there are no preferential factives. One must tread carefully here, however, since not all entailments need be encoded in the meaning of the verb—i.e. this belief entailment could plausibly arise via the same sorts of pragmatic processes that give rise to the factive presupposition in the first place. I remain agnostic on this issue here, since it does not bear on the current work.

with both nonquestion (23a) and question (23b) subordinate clauses, while the nonfactive *think* can occur with nonquestion subordinate clauses (24a) but not question subordinate clauses (24b).[8]

(23) a. Mary knows that John went to the store.

b. Mary knows {if, why} John went to the store.

(24) a. Mary thinks that John went to the store.

b. *Mary thinks {if, why} John went to the store.

1.2.3 Assertivity

Further cross-cutting representationality and factivity is the "assertivity" distinction (Hooper, 1975).[9] Like factivity, assertivity is defined in terms of its effects on discourse. Again very roughly, a verb is assertive if it can be used in situations where its subordinate clause is relevant to the main point of the utterance (see Urmson 1952; Simons 2007; Anand and Hacquard 2014 for discussion). For instance, *think* and *say* seem to allow this (25a), but *hate* does not (25b).

(25) a. **A:** Where is Mary?

B: John {thinks, said} that she's in Florida.

b. **A:** Where is Mary?

B: # John hates that she's in Florida.

[8]This paradigm is filled out by what Lahiri (2002) calls rogatives, like *wonder* and (for some speakers) *ask*. *Wonder*, at least, takes only subordinate questions and not nonquestions.

[9]Whether assertivity fully cross-cuts representationality is unclear, since the only verb that both has a preferential component and is plausibly assertive—*hope*—also has a representational component.

Assertivity correlates with the availability of S-lifting and the propositional anaphor object *so*. Assertives, like *think* and *say*, can occur with S-lifted subordinate clauses (26a) and *so* (27a), but *doubt* cannot occur with either S-lifting (26b) or *so* (27b).

(26) a. She's in Florida, John {thought, said}.

b. *She's in Florida, John doubted.

(27) a. John {thinks, said} so.

b. *John doubts so.

(Hooper, 1975) claims that the assertivity distiction cross-cuts the factivity distinction to give rise to a further split between semi-factives (assertive factives), like *know*, and true factives (nonassertive factives), like *love* and *hate* (see Karttunen 1971 for an early description of this distinction).[10] Important for my purposes is that the semi-factive v. true factive distinction appears to correlate (i) with the (semantic) representationality distinction—semi-factives also tend to be cognitive factives and true factives, emotive factives—and (ii) at least two sorts of syntactic distinctions. First, semi-factives tend to allow both polar (28a) and WH (28b) questions, but true factives tend to allow only WH questions (29b), not polar questions (29a).

(28) a. Mary knows if/whether John sliced the bread.

b. Mary knows if/whether John sliced the bread.

[10]The pragmatic effects that distinguish semi-factivity from true factivity are beyond the scope of this dissertation. Much ink has been spilled regarding the nature of semi-factivity in recent years, however, so the interested reader is encouraged to see, e.g., Simons 2001; Abusch 2002; Abbott 2006; Romoli 2011.

(29) a. *Mary {loves, hates} if/whether John sliced the bread.

b. Mary {loves, hates} how John sliced the bread.

Second, semi-factives tend to allow complementizer ommission (30a), but true factives tend not to (30b). This second correlation is less strong and is likely modulated by syntax: expletive subject emotive factives appear to be better with complementizer omission, particularly when they passivize (see Grimshaw 2009 for further recent discussion of complementizer ommission).

(30) a. I {know, realize} (that) Mary already went to the store.

b. I {hate, love} *(that) Mary already went to the store.

(31) a. It {amazed, bothered} me ???(that) Mary already went to the

b. I was {amazed, bothered} ?(that) Mary already went to the

1.2.4 Communicativity

The final distinction I note is communicativity—which, transparent from its name, roughly corresponds to whether a verb refers to a communicative act, or perhaps more generally, (manner of) externalization of linguistic form. This distinction cross-cuts at least the representationality distinction—there are both representational communicatives, like *say* and *tell*, and preferential communicatives, like *demand*—and perhaps other distinctions as well, such as the factive-nonfactive distinction (see Anand and Hacquard 2014 for extensive discussion of whether com-

municativity truly cross-cuts factivity or not).[11]

The syntactic correlates of communicativity seem quite apparent on the surface. Communicative verbs, along with a subordinate clause, tend to take noun phrase (32a) or prepositional phrase (32b) arguments representing their communicatee (Zwicky, 1971).

(32) a. John told me that Mary went to the store.

b. John said to me that Mary went to the store.

But though this is often treated as a clearly marked distinction, there are various reasons to be cautious about it. For instance, note that *demand* and *tell* can occur in string-identical contexts with *want* and *believe*. These string-identical contexts appear to be be distinguished only given some parse of the string. Wanting and believing don't seem to involve anything besides a wanter/believer and a thing wanted/believed. In contrast, telling and demanding seem to require a tellee/demandee.

(33) John {told, demanded, wanted, believed} Mary to be happy.

This is plausibly syntactically encoded. Note that the pleonastic element *there*, which is plausibly an overt cue to the particular syntactic configuration in question, is only allowed with *want* and *believe*, but not *tell* and *demand*. This has been used to suggest that *tell* and *demand* in (33) involve an underlying object while *want* and *believe* do not.

[11]Whether *say* and *tell* are only representational is a question. Both can be used to talk about commands conditional on their taking a nonfinite subordinate clause. In any case, they plausibly have something like a representational use with finite subordinate clause.

(34) John {*told, *demanded, wanted, believed} there to be a raucous party happening outside.

Further, there are some string-identical contexts that both communicative and non-communicative verbs can appear in which plausibly have no syntactic (or perhaps even selectional) distinctions. For instance, the communicative verb *promise* and the verb *deny*, which is plausibly noncommunicative in this syntactic context, both allow constructions with two noun phrases.

(35) John {promised, denied} John a meal.

This is not to say that the semantic distinction has no syntactic correlates, of course; it is just to say that they may not be apparent from the string context.

1.3 Beyond pen and paper

In the last section, I showed various promising results regarding the relationship between semantic features and syntactic features derived from traditional distributional analysis. In the context of propositional attitude verb learning, such results are important in that, it provides a general guide to to where one might look to see whether syntactic bootstrapping is feasible as a strategy for learning these words.

The problem is that this is as far as traditional distributional analysis is likely to take us. Understanding how words are learned involves understanding how contextual cues to a word's meaning are utilized to infer that meaning. Distributional

analysis can tell us which contextual cues (or combinations thereof) might be correlated with which features, but it cannot tell us the strength of this correlation. But if what we are looking for is a mechanism to take advantage of contextual cues, this information is crucial, especially if the correlations traditional distributional analysis provides are not perfect. And as noted above, while possibly quite strong, these correlations are certainly *not* perfect in the domain of propositional attitude verbs. Indeed, as I note in Chapter 5, beyond not being perfect, they are also cross-linguistically unstable.

So how does one go about assessing the correlations between contextual cues—of interest here, syntactic contextual cues—and a word's semantic features? The answer is that one must devise some way of quantifiying the relationship between a word's semantics and its syntactic distribution. This in turn requires some way of *measuring* a word's semantics and its syntactic distribution. In Chapter 2, I address how one can obtain a measure of the semantics (or at least a suitable proxy). For the remainder of this section, I focus on what it means to obtain a measure of syntactic distribution.

There are a couple ways to obtain such a measure, which correspond to two common notions of syntactic distribution. For the syntactician and the semanticist, a word's syntactic distribution is defined modally: which syntactic contexts can a word occur in? For the computational linguist, a word's syntactic distribution might more commonly be defined as actualized: which syntactic contexts does a word occur in—e.g. in a corpus? The two are presumably related, but the latter likely involve aspects of linguistic performance independent of the former. For this

reason, I refer to the former (modal) notion as *competence distribution* and the latter (actualized) notion as *performance distribution*.

To get a sense for how these two notions pull apart, note that a verb might allow a frame according to its competence distribution that rarely, if ever, shows up. For instance, *believe* can occur in the syntactic context *I _ Mary to be intelligent*, but that locution has a register that will make that syntactic context's empirical distribution with *believe* look quite different from the nearly equivalent *I _ that Mary is intelligent*. This is presumably not because *believe* is any worse in the first context as compared to the second; it's just that the second is found in a much wider variety of registers. Further, this is not because the first syntactic context is unlikely overall, since the highly frequent verb *want* shows up in contexts like *I _ Mary to be intelligent* quite frequently.[12]

Pulling these two notions apart yields two kinds of syntactic distribution that might be measured. Measurement of the first kind, the competence distribution, corresponds most closely to the methodology employed in traditional distributional analysis—grammaticality/acceptability judgments—and thus gathering such a measure provides a way of validating those traditional methodologies' findings while

[12]A related distinction arises in the literature on selectional preferences/semantic plausibility (Katz and Fodor, 1963; Johnson-Laird, 1983; Trueswell et al., 1993, 1994; Grimshaw, 1994; Pinker, 1994; Resnik, 1996, among many others). Frequency is only partially correlated with the plausibility of a description. For instance, in a search of the PukWaC corpus (Baroni et al., 2009), the six most frequent content words heading objects of the verb *eat* (log relative frequency in parentheses) are *food* (-3.25), *meat* (-4.16), *meal* (-4.35), *diet* (-4.51), *fish* (-4.59), and *lunch* (-4.83). In contrast, words for offal, such as *heart* (-7.74) or *liver* (-8.15), occur much less frequently; and some, such as *kidney*, do not occur at all in the corpus. This does not seem to be a fact about the coherence of the description *eat kidney*, e.g., as compared to a description like *eat idea*. Rather, as in the case of *I believe Mary to be intelligent*, some performance factor(s), broadly construed—e.g. the frequency with which one has cause to talk about eating offal as opposed to eating lunch—conspire to make it less frequent. The relationship between this notion of coherence and the notion of acceptability in a frame is fleshed out in Chapter 3.

augmenting them with an explicit methods for assessing correlation between particular semantic features and syntactic distribution. I carry this out in Chapter 2.

Measurement of the second kind of syntactic distribution, performance distribution, hews more closely to the sort of information that learners actually have access to—a corpus of utterances—and thus gathering such a measure provides a way of showing what may actually be learnable from the input. I carry this out in Chapter 3 by building on the methods utilized in Chapter 2.

In the case of both of these measures, one can only say whether information that correlates with the semantics lies in that measure's notion of syntactic distribution. However, for a syntactic bootstrapping story to go through, it must be further shown that learners can utilize the information in syntactic distribution in a setting where they receive that information incrementally. In Chapter 4, I present a methodology for assessing this.

1.4 Discussion and roadmap

In this chapter, I laid out the central problems of learning what Gleitman et al. (2005) dub the *hard words*, focusing in particular on the *propositional attitude verbs* like *think*, *know*, and *want*. I noted two main problems for learning these verbs: (i) the eventualities they describe tend not to have sensory correlates, and (ii) their meanings are both fine-grained and multi-faceted, thus presenting problems for accounts based on learning from nonlinguistic context (or even discourse context)

alone.

I then turned to a discussion of learning from linguistic context, noting two particular kinds of linguistic contexts that have been discussed as possible learning cues: lexical context and syntactic context. I noted that, while lexical context is likely useful for certain distinction among verbs—indeed, it may be useful even for some distinctions among propositional attitude verbs—it likely does not track other distinctions of central interest. This led me to turn to the use syntactic context as a word-learning cue—a strategy exemplified most notably in *syntactic bootstrapping* approaches to word learning.

I noted two problems that any syntactic bootstrapping approach must solve: (i) it must explain how learners cluster verbs based on the syntactic contexts they occur with—the *clustering problem*—and (ii) it must explain how learners label these clusters with the facets of meaning they correspond to—the *labeling problem*. The ability of a syntactic bootstrapping account to solve either of these problems for any particular type of verb is dependent on the (i) the granularity with which that particular verb type's semantics is mirrored by the syntactic distribution and (ii) the availability of principles that would allow a learner to label the semantic features. I raised doubts about this second prospect having to do with the cross-linguistic stability of the mapping principles, particularly in the attitude domain, arguing that the labeling problem quite plausibly could be solved via other means, and so the first problem should be attacked first in isolation.

I then turned to an overview of what is known about this relationship in the domain of propositional attitude verb. I showed that the results are quite promising

but also that the correlations are not perfect. This raises the need for a more fine-grained investigation of these correlations, which I carry out in this dissertation.

In **Chapter 2**, I begin this investigation by showing how to quantify the relationship between naïve speakers' knowledge of the syntactic contexts a propositional attitude verb can occur in—what I refer to as the *competence distribution*—and their knowledge of that verb's semantics. To do this, I deploy a methodology that Fisher et al. (1991) used to probe such relationships as they obtain for verbs across the lexicon, here focusing in on the propositional attitude verb domain in order to test the limits of this relationship. The main result of this chapter is that there is a significant correlation between the syntax measure and the semantics measure. This omnibus result, however, tells us little about the relationship between particular syntactic contexts and particular facets or features of the meaning. To delve into this, I develop a model, which I dub the *nonnegative model of projection*, to investigate this relationship. The benefit of this model is that it furthermore implements part of a solution to the clustering problem. I show that this model discovers the sorts of fine-grained features discussed above.

In **Chapter 3**, I investigate to what extent the same sort of relationship found between verbs' competence distributions and their semantics also obtains between the distribution of syntactic contexts a propositional attitude verb occurs in in a corpus, what I refer to as its *performance distribution*, and participants' knowledge of those same verb's semantics. To do this, I develop a model that augments the nonnegative model of projection presented in the previous chapter with a model of corpus count data. This model simultaneously discovers competence

distributions using the corpus distributions, while at the same time solving the clustering problem. The main result of this chapter is that performance distributions also carry a significant amount of information about propositional attitude verb semantics and that this information is comparable with that found in the direct measures of competence distribution employed in Chapter 2.

In **Chapter 4**, I investigate whether the information in performance distributions is in fact accessible to learners, and if so, how robustly represented this information is. To do this, I adapt recently developed methodologies related to the Human Simulation Paradigm (HSP) to (i) measure the informativity of particular items in the performance distribution about the semantics of the word that occurs in them and (ii) measure the informativity of the distribution itself. The main result of this chapter is that, even if items are manipulated in such a way to give participants as little information as possible, inference to all propositional attitude verbs meanings are extremely robust, even down to extremely fine-grained facets of those verbs' meanings.

In **Chapter 5**, having focused for the majority of the dissertation on solving the clustering problem, I present a novel proposal for how to approach the labeling problem. This proposal starts with the observations that, particularly in the propositional attitude verb domain, the relationship between particular aspects of the semantics and particular syntactic contexts seems to be cross-linguistically unstable. This does not raise problems for the model presented in previous section necessarily, since as long as those languages exhibit roughly the same patterns of correlations between meaning and syntactic context, this model should similarly

succeed in solving the clustering problem. The problem arises if labels are somehow associated *a priori* with particular syntactic contexts—for instance, if tense were somehow associated with the representationality distinction—since not all languages show this correlation. The proposal presented in this chapter is that, while not all languages associate particular facets of the semantics with particular syntactic contexts, at least some particular facets may be associated with families of syntactic contexts and that the learner's job is to select the appropriate syntactic context to associate with that facet using the data. I then show how this might be encoded in a model like the one I develop in the previous chapters.

In **Chapter 6**, I conclude the dissertation with future directions for this work.

Chapter 2: A computational model of projection

In Chapter 1, I laid out the central motivations for the syntactic bootstrapping approach. Propositional attitude verbs are prime candidates for words whose meanings are learned via syntactic bootstrapping. This raises the question: how much information about a propositional attitude verb's meaning lies in its syntactic distribution?

I begin to give an answer to this question in the current chapter by quantitatively assessing how much information about a word's meaning lies in that word's competence distribution using both experimental and computational methods. As I noted in the last chapter, such a quantitative assessment provides a way of assessing the viability of results from traditional distributional analysis. I show that (i) there is significant agreement between the competence distributions and the measure of semantics employed and (ii) these agreements largely corroborate the results of the traditional distributional analysis methods.

Besides providing such an assessment, this chapter also contributes a novel methodology for inducing semantic features from the sorts of competence distribution measures employed here. I show that this methodology is furthermore linguistically interesting in that it quite naturally models the linguist's traditional notion

of projection.

To lay the groundwork for this contribution, I begin the chapter with a broad overview of the notion of projection. This leads naturally into a discussion of the experimental methodology I utilize in this chapter to measure aspects of a word's meaning and its competence distribution. This methodology is the direct application of one developed by Fisher et al. (1991) and extended by Lederer et al. (1995). As it will be important for grounding discussion throughout the dissertation, I review the logic of this methodology as it relates to the notion of projection.

Subsequently, I present three experiments that focus in on propositional attitude verbs: one that aims at quantifying these words' competence distributions and two that aim at assessing their semantic properties. In analyzing the competence distribution data, I explore various ways of extracting information from the measure of the competence distribution that fall broadly in the domain of factor analysis. I relate these factor analysis methods back to the earlier discussion of projection, showing that the assumptions these methods make about the process that generates competence distributions directly maps onto the linguist's notion of projection. I argue that in particular non-negative matrix factorization methods hew very closely to the sorts of projection architectures linguists conceptualize.

In the next section, I turn to an analysis of the two semantic measures. I show that, on the whole, these measures agree, but that they also show interesting areas of disagreement. Despite this disagreement these measures both correlate reliably with the competence distribution measure. This establishes that there is a significant amount of information shared between the syntax and the semantics. I then delve

into these what drives these correlations by asking how well particular features extracted using the factor analysis methods employed to analyze the competence distribution data fare in predicting the two measures of semantics. I then conclude.

2.1 The meaning-syntax relationship

In this section, I present an abstract characterization of the traditional methodology employed by linguists to study the relationship between meaning and syntactic distribution. I then review a critique of this traditional methodology brought forward by Fisher et al. (1991) along with their methodological solution.

2.1.1 Linguistically relevant meaning

Linguists of all stripes have a standing interest in the relationship between word meaning and syntactic distribution—an interest that Zwicky (1971) distills quite elegantly in the introduction to his classic squib on manner-of-speech verbs.

> To what extent is it possible to predict certain properties of words (syntactic, semantic, or phonological), given others? [And] insofar as there are such dependencies among properties, what general principles explain them? (ibid., p. 223)

Indeed, it has long been recognized that questions regarding which semantic distinctions are morphosyntactically relevant are the only ones linguists can claim propriety over; distinctions in meaning beyond those predictable from other linguistic properties fall equally well into the domain of the lexicographer (Fillmore,

1970)—or in modern times, the computer scientist (cf. Mikolov et al., 2013). Embedded in this view is the idea that an item's linguistic contexts are responsive to only some conceivable contrasts in meaning and that a linguistic theory of the link should speak to exactly which these are and why other conceivable contrasts are excluded (cf. Jackendoff, 1972; Grimshaw, 1979; Pinker, 1989; Levin, 1993). As Zwicky puts it, the question for the linguist is "what sorts of word classes are there, and why these and not others?" (ibid., p. 223)

An example of the distinction between linguistically relevant and linguistically irrelevant semantic distinctions comes from Pesetsky (1991). Following Zwicky, he notes that, though "verbs of manner of speaking"—e.g. *holler* and *whisper*—and "verbs of content of speaking"—e.g. *say* and *propose*—are distributionally distinguishable, "verbs of loud speech"—e.g. *holler* and *shout*—and "verbs of soft speech"—e.g. *whisper* and *murmur*—do not seem to be. (For example, verbs of content of speaking "resist adjunct extraction and allow complementizer deletion" (Pesetsky, 1991, p. 14).) That is, the manner-content distinction has consequences for the syntax, whereas the loud-soft contrast does not.

In fact, the generalization extends beyond predicates that refer to speech sounds to predicates that refer to sounds in general. The volume, pitch, resonance, and duration of the relevant sound do not seem to have bearing on its distributional properties, but the mode of generation (internally v. externally caused) does (Levin and Hovav, 2005). This suggests that, whatever constitutes the nature of the connection between a word's semantic properties and its syntactic distribution, it is blind to certain possible conceptual distinctions—in this case, sonic properties.

Thus, though nonlinguistic meanings—i.e. concepts—may be distinguishable to a very fine grain-size, linguistic meanings may not be. In this sense, the linguistic system can be conceived of as a filter on the properties of the conceptual system's objects, retaining some properties wholesale while discarding others. To introduce a convention I use throughout the dissertation, suppose $\mathbf{c}_i$ is some representation of concept i, then $\mathbf{s}_i$ is some representation that encodes all and only the linguistically relevant features of $\mathbf{c}_i$.[1] At a high level of abstraction, then, (part of) the interface between language and other areas of cognition—in particular, the Conceptual-Intentional (CI) interface—might be viewed as an information-preserving, or homomorphic,[2] mapping CI from objects in the concept space C to their syntactically relevant features in the semantic feature space S.[3]

$$C \xrightarrow{\text{CI}} S$$

[1] I use the following conventions throughout the remainder of the dissertation: italicized capital letters stand for representational spaces—i.e. possible representations; normal capital letters refer to mappings between these spaces; bolded lower-case letters, which will tend to be subscripted, refer to particular instantiations of the corresponding space, and bolded upper-case letter refer to collections of these specific instantiations. The bolding convention in particular is used because representational instantiations are cashed out as vectors and their collections as matrices or tensors, and bolding is standard in linear algebra and related disciplines for representing vectors and their generalizations.

[2] This way of speaking assumes that the distinctions made among linguistic meanings is a subset of those made between nonlinguistic meanings. Such a containment relationship is not conceptually necessary. I have nothing to say about this possibility.

[3] This presupposes a contentious point about the complexity of lexical items (cf. Fodor and Lepore, 1998, 1999). While Fodor and Lepore's arguments are serious, I believe that a reader who abides by the dictum that lexical items be represented atomically might still find use in this chapter. This is one main reason that I stress the distinction between discovering distributional regularities and discovering semantic content throughout the chapter. I take it as *prima facie* reasonable that representations of a word's distributional regularities may be complex in a way that the representation of its content may not be, since of course, one needs to explain C-selection somehow. One might then wonder whether I have just put a new name—*distributional regularity*—on an old concept—semantic decomposition. I think there is reason to believe that I have not, given the way I link distributional regularities to similarity judgments, but I will have to leave this question open.

It is worth stressing the following implication: $\mathbf{s}_i$ may not exhaust the semantic representation of word i; indeed, $\mathbf{s}_i$ may not even be semantic in any important sense. For instance, it might be conceived of as a (structured) index into subsets/subspaces of concepts—hence the importance of specifying that the mapping is homomorphic, not necessarily isomorphic. In this sense, $\mathbf{s}_i$ would be purely formal, though depending on its structure, it might imperfectly mirror relationships in the conceptual space. Thus, knowing $\mathbf{s}_i$ for a word i would be insufficient for fixing that word's corresponding concept $\mathbf{c}_i$.

By definition, however, $\mathbf{s}_i$ would be sufficient for determining various linguistic properties of word i, such as its syntactic distribution $\mathbf{d}_i$. To say that the syntactic distribution $\mathbf{d}_i$ of word i can be determined from its linguistically relevant semantic features $\mathbf{s}_i$ is to say that there is some mapping from the space of possible semantic representations S to the space of syntactic distributions D. In standard models of the syntax-semantics interface, this mapping, call it P, is determined by a set of *projection principles* (cf. Gruber, 1965; Carter, 1976; Chomsky, 1981; Pinker, 1989; Grimshaw, 1990; Levin, 1993; Hale and Keyser, 2002).[4]

$$S \xrightarrow{\mathrm{P}} D$$

To take a concrete example: in reviewing the literature on the representational (*think*, *say*, *know*) v. preferential (*want*, *order*, *prefer*) distinction among propositional attitude verbs, I noted the apparent correlation (in English) between repre-

[4]In the remainder of the dissertation, I overload the term *projection principles* to refer to either the mapping **P** itself or the principles that make it up, allowing context to disambiguate where possible. In general, the term will be used to refer to the function **P** itself.

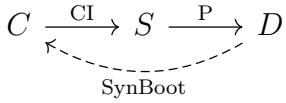
sentationality and tense: representationals tend to take finite subordinate clauses, whereas preferentials tend to take nonfinite subordinate clauses. If this correlation holds, this would suggest (i) that representationality and preferentiality are encoded in $\mathbf{s}_i$; and (ii) that the principles map the encoding of representationality to some representation of the distribution that encodes finite complementation and the encoding of preferentiality to some representation of the distribution that encodes nonfinite complementation.

Putting these two components together—the mapping from the conceptual space C to the (linguistically relevant) semantic feature space S and the mapping P from the semantic feature space S to the syntactic distribution space D—the following abstraction over the relationship between meanings and distributions results.

$$C \xrightarrow{\text{CI}} S \xrightarrow{\text{P}} D$$

If this model is correct, the upshot for a theory of verb-learning that relies on syntactic context—e.g. syntactic bootstrapping—is that there is likely a limit on the meaning properties that syntactic context could be used to learn even in principle. Why? Suppose the learner has access to the syntactic distribution $\mathbf{d}_i$ for some word i and that their job is to infer the concept $\mathbf{c}_i$ association with word i. That is, they need to "reverse" both the projection rules P and the mapping from the conceptual space to semantic features CI.[5]

[5]In the remainder of this chapter, I use the following convention: solid lines represent theoretical, or computational-level (Marr, 1982), relationships; dashed lines represent algorithms that map between representations—as in the case of syntactic bootstrapping, possibly utilizing the computational-level relationships.



A learner's ability to perform this reversal from the syntax alone will necessarily be bound by the information lost due to the mapping P from semantic features to syntactic distributions—i.e. the projection principles—and the information lost due to the mapping CI from concepts to semantic features. This gives Zwicky's question new force. Reformulating it relative to the above abstraction, this question has two parts: (i) what kinds of things constitute the possible semantic representations S? And (ii) what kind of relation do the principles P instantiate?

How does one approach this question? The traditional methodology—e.g. the one that produced many of the results discussed in the last chapter—is to assume knowledge of both the concepts $\mathbf{C}$ and the syntactic distributions $\mathbf{D}$ associated with various words and then to attempt to infer both the space of (linguistically) relevant semantic representations S and the projection principles P. Thus, making the simplifying assumption that learners have access to all objects in the conceptual space C and the syntactic distributions of some words $\mathbf{D}$, the linguist and the learner look very much alike. The only difference between the two under this model, modulo the simplifying assumptions, is that the learner does not have access to the pairing of the concept $\mathbf{c}_i$ and syntactic distribution $\mathbf{d}_i$ for word i; rather, they have the syntactic distribution $\mathbf{d}_i$ of word i and a set of possible concepts $\{\mathbf{c}_j\}$.[6]

[6]From a machine learning perspective, the linguist carries out some supervised learning algorithm to inferm S and the learner carries out an unsupervised learning algorithm.

This method has been quite successful, uncovering regularities in many disparate areas of the lexicon (see Levin and Hovav 2005 and Williams 2015 for broad overviews of this work). However, Fisher et al. (1991) note that this methodology has its limits in the fact that

> ...only those semantic generalizations that can be readily labeled by the investigator are likely to be discerned. It may well be that there are semantic abstractions which, while correlated with the syntax, are not so easy to puzzle out and name. (p. 342)

In the propositional attitude verb domain, for instance, one possibility is that representationality, factivity, assertivity, and communicativity have been been posited as syntactically relevant, in part, because they are readily labeled by investigators. This methodological problem arises, they argue, as a consequence of confounding isolation of a property and labeling of that property, since "...disagreements over labels for semantic features can get in the way of deciding whether those features are marked in the syntax" (ibid, p. 342). Note that this is analogous to the labeling problem, discussed in Chapter 1: even assuming syntactic distribution is attended to, how does a learner link the appropriate features of that distribution (syntactic contexts) to the appropriate meaning components?

Their methodological solution has three components: (i) independent measures of both the array of syntactic contexts a verb i can occur in (its syntactic distribution) and that verb's meaning; (ii) some way of extracting regularities from meaning measure; and (iii) some way of stochastically mapping these regularities

into the semantic measure. Within the above abstraction, (i) involves measuring $\mathbf{d}_i$ and $\mathbf{c}_i$, while (ii) and (iii) involve constructing a mechanism that carries out the same sort of reversal as syntactic bootstrapping.

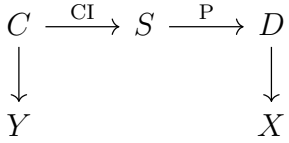
To implement these components, Fisher et al. begin by attaining, for a set of verbs spanning the lexicon, semantic similarity judgments for those verbs—call the resulting data $\mathbf{Y}$, an approximation to the true concepts $\mathbf{C}$. The idea here is that such quantitative representations allow one to bypass the sort of explicit labeling inherent to the traditional method, since distinctions among features salient to the participants are not explicitly invoked.[7]

$$\begin{array}{ccccc} C & \xrightarrow{\text{CI}} & S & \xrightarrow{\text{P}} & D \\ \downarrow & & & & \\ Y & & & & \end{array}$$

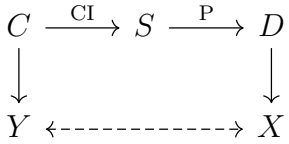
Their goal is to then compare this proxy $\mathbf{Y}$ with a quantitative representation of those verbs' syntactic distributions gathered using an acceptability judgment task.[8] I refer to these sorts of quantitative representations as $\mathbf{X}$, an approximation of $\mathbf{D}$.

[7]There is a question here to what extent the $\mathbf{C}_i$ is solely dependent on $\mathbf{C}$ and not, e.g., $\mathbf{D}$ itself. Fisher et al. give various arguments that $\mathbf{Y}$ is plausibly the product of participants utilizing some aspects of the meaning of the words in the task independently of the correspond syntactic distributions $\mathbf{D}$. As far as I can discern, it would be nearly impossible to tell whether the similarity judgments $\mathbf{Y}$ are a product (to some extent) of comparing of verbs' syntactic distributions $\mathbf{D}$ or whether they are a product of conceptual feature correlated with those distributions.

[8]Lederer et al. (1995) took a similar tack, using the same sort of semantic similarity judgment task but replacing acceptability judgments with syntactic distributions extracted from a corpus.

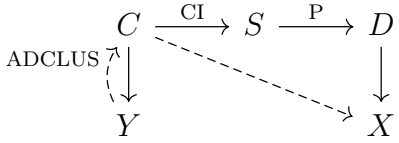


Fisher et al.'s question is then how well the semantic similarity judgments **Y** and acceptability judgments **X** match up and in what ways do they match up. The first method they use for doing this is direct comparison of the similarity judgments and (similarities derived from) the acceptability judgments.

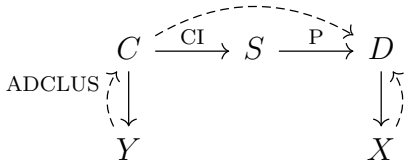


The second method they use is to explicitly extract features from the semantic similarity judgments. The specific algorithm they use for extracting these features is a form of *additive clustering* (ADCLUS; Shepard and Arabie 1979), which can in turn be viewed as doing inference over a generative model of similarity judgments, where similarity judgments are the number of matching binary features, weighted by feature and with noise (Tenenbaum, 1996; Tenenbaum and Griffiths, 2001).[9] As such, I denote it with a a dashed arrow back up to the conceptual space. The features extracted using this procedure are then mapped into the syntactic distribution proxy using a linear map.

[9]Fisher et al. actually use a related algorithm called OVERCLUS (Sarle, 1979). I have been unable to track down the original reference for OVERCLUS, which is an unpublished dissertation proposal, and an implementation no longer ships with the statistical package Fisher et al. report using (SAS). As far as available secondary references go in describing OVERCLUS, however, it appears to produce the same basic kind of representations as ADCLUS—binary features—and can similarly be encoded as doing inference over a generative model.



Another way of thinking about this mapping is as going from C to D directly. This depends to some extent on whether there is some procedure for reconstructing the distributions **D** from their acceptability judgment proxy **X**. For instance, averaging acceptability judgments for a particular verb-subcategorization frame pair, as Fisher et al. do, might qualify.[10] This is discussed extensively later in this and the next chapter.



One thing worth noting about these last two steps is that the methods that Fisher et al. utilize to extract syntactic regularities and compare them against the semantic similarities make potentially substantive assumptions about the nature of the syntactic regularities and the nature of their relationship to the semantics—in the case of the regularity extraction procedure, that these representations are discrete/symbolic.

[10] As I note in the next section, however, averaging is not a particularly good way of analyzing the sort of acceptability judgment data—ordinal data—Fisher et al. collect.

2.1.2 Discussion

In this section, I presented an abstract characterization of the traditional methodology employed by linguists to study the relationship between meaning and syntactic distribution. The basic architecture of the system can be described by two mappings: one from the conceptual space to a space of linguistically relevant semantic features (CI) and another from the linguistically relevant semantic features to syntactic distributions (P)—the projection principles.

$$C \xrightarrow{\text{CI}} S \xrightarrow{\text{P}} D$$

I casted the traditional methodology for discovering the projection principles P and linguistically relevant semantic features **S** as involving analysis of concept ($\mathbf{c}_i$)-syntactic distribution ($\mathbf{d}_i$) pairs. I then reviewed a methodological critique of the traditional methodology brought forward by Fisher et al. (1991) along with their methodological solution. In the course of this review, I reified the logic of their methodology pseudoformally as using quantitative proxies of the semantics **Y** and the syntactic distributions **X** to discover these relationships.

$$\begin{array}{ccccc} C & \xrightarrow{\text{CI}} & S & \xrightarrow{\text{P}} & D \\ \downarrow & & & & \downarrow \\ Y & & & & X \end{array}$$

The remainder of this chapter follows Fisher et al.'s experimental methodologies and analytical logic quite closely. I diverge from them in two ways, however.

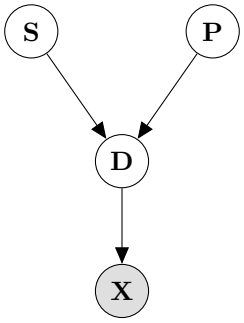


Figure 2.1: graphical model for generative model corresponding to $S \xrightarrow{\mathrm{P}} D \rightarrow X$.

First, I focus here on a much smaller piece of the lexicon than Fisher et al., who look at much wider swaths of the lexicon and similarly sized swaths with subcategorization frames that were not particularly fine-grained. The idea of looking at a smaller swath and using fine-grained frames is that this can help uncover the limits of semantic information in syntactic distribution. Second, in contrast to Fisher et al., who are not concerned with the regularities that can be extracted from the syntax directly, I present an analytical innovation that takes the above structure seriously by reifying it into a generative model. This generative model gives the chapter its title in that it naturally captures the traditional notion of projection.

Figure 2.1 shows an abbreviated representation of this model in the form of a graphical model, where shaded circles represent observed variables—here, the syntactic proxy **X**—and unshaded circles represent variables that must be inferred. Following the logic laid out above, all but the syntactic proxy is observed, and thus the true competence distribution **D**, the linguistically relevant semantic features **S**, and the projection rules **P** must be inferred.

In the next section, I show two ways this model can be cashed out—Principal

Component Analysis and Nonnegative Matrix Factorization—contrasting the output of these methods with another: hierarchical clustering.

2.2 Experiment 1: verb-frame acceptability

In this section, I present an experiment aimed at getting a measure of how acceptable a variety of propositional attitude verbs are in different syntactic contexts. My goal is two-fold. First, I assess how closely the sorts of regularities found in these data correspond to the attitude verb distinctions discussed above. I carry this out by using two standard exploratory analyses—hierarchical clustering and principal component analysis—and one novel analysis (at least in this domain)—nonnegative matrix factorization. Second, I assess the strengths and weaknesses of each exploratory analysis with respect to how well they satisfy various methodological considerations. This assessment is driven in part by the sorts of regularities that are discovered in the data, along with theoretical considerations. I suggest that the nonnegative matrix factorization approach most closely fits with the traditional notion of projection.

2.2.1 Design

This experiment aims to get a measure of how acceptable a variety of propositional attitude verbs are in different syntactic contexts. To do this, 30 propositional attitude verbs were selected in such a way that they evenly spanned the classes in Hacquard and Wellwood's (2012) semantic classification. This classification is

essentially a more elaborated version of the classification presented in Section 2.1.

19 syntactic features whose distribution has been claimed to be sensitive to attitude verb lexical semantics were then selected. These features consist in five[11] broad types: clausal complement features, noun phrase (NP) complements, prepositional phrase (PP) complements, expletive arguments, and anaphoric arguments. (Note that I break these features into types for expository purposes only. No special status is afforded to these groupings in the analysis.)

2.2.1.1 Features of interest

Six types of clausal complement features were selected: finiteness, complementizer overtness, subordinate subject overtness, subordinate question type, S-lifting, and small clause type. Finiteness had two values: finite (1a) and nonfinite (1b).

(1) a. Mary thought that John went to the store.

b. Mary wanted John to go to the store.

Complementizer presence had two values: present (2a) and absent (2b).

(2) a. Mary thought that John went to the store.

b. Mary thought John went to the store.

Embedded subject presence had two values: present (3a) and absent (3b) and is relevant only when the clause is finite and has no overt complementizer.

[11] A sixth feature—degree modification—was also selected for investigation. I exclude this from the analyses since the information degree modification carries is likely purely—or at least mostly—semantic in nature.

(3) a. Mary wanted John to go to the store.

b. Mary wanted to go to the store.

Embedded question type had three values: nonquestion (4a), polar question (4b), and WH question (4c).

(4) a. Mary knows that John went to he store.

b. Mary knows if John went to he store.

c. Mary knows why John went to he store.[12]

S-lifting had two values: first person (5a) and third person (5b).

(5) a. John went to the store, I think.

b. John went to the store, Mary said.

Small clause type had two values: bare small clause (6a) and gerundive small clause (6b).

(6) a. Mary saw John go to the store.

b. Mary remembered going to the store.

Two NP structures were selected: single (7a) and double objects (7b).[13]

(7) a. Mary wanted a meal.

b. Mary promised John a meal.

[12]Only adjunct questions were used, since constituent questions are ambiguous on the surface between a question and a free relative reading.

[13]NPs were chosen so as not to have an interpretation in which they could be interpreted to have propositional content (Moulton, 2009a,b; Uegaki, 2012; Rawlins, 2013; Anand and Hacquard, 2014).

A third feature relevant to NP complements—passivization—was also included (8).[14]

(8) John was said to be intelligent.

Two types of PP complement were selected: PPs headed by *about* (9a) and PPs headed by *to* (9b).

(9) a. Mary thought about John.

b. Mary said to John that she was happy.

Three types of expletive arguments were selected: expletive *it* matrix subject, expletive *it* matrix object, and expletive *there* matrix object/embedded subject.

(10) a. It amazed John that Mary was so intelligent.[15]

b. John believed it that Mary was top of her class.

c. John wanted there to be food on the table.

Three types of anaphoric complement features were selected: *so* (11a), null complement/intransitive[16] (11b), and nonfinite ellipsis (11c).

(11) a. Mary knew so.

b. Mary remembered.

c. Mary wanted to.

[14]The availability of structures like (8) and the unavailability of structures like (3a), appears to correlate with whether a predicate's eventivity and/or its encoding of manner (Postal, 1974, 1993; Pesetsky, 1991; Moulton, 2009a,b, see also Zwicky 1971 for other syntactic and semantic features that track manner of speech).

[15]It is difficult to force the subject in a sentence like (10a) to be interpreted nonreferentially. As I see in Figure 2.2, this likely affected the judgments for verbs like *tell*, which are fine in this frame if the subject is interpreted referentially.

[16]Note that I cannot be sure that these structures involve null complements in either Williams' (2015, Ch. 5) broad or narrow sense. See Hooper 1975; Hankamer and Sag 1976; Grimshaw 1979; Depiante 2000; Williams 2012 for further discussion of these structures.

2.2.1.2 Stimulus construction

These 19 features were then combined into 30 distinct abstract frames. (Again, note that the features are mentioned for expository purposes only. They do not enter into the analysis in any formal sense.) These abstract frames are listed along the x-axis in Figure 2.2. Each categorial symbol in the frame should be interpreted as follows:

NP	NP constituent (e.g. *Mary*)
WH	(Adjunct) WH word (e.g. *why*)
V	Bare form of verb (e.g. *think*)
VP	Verb phrase with verb in bare form (e.g. *fit the part*)
S	Finite clause without complementizer (e.g. *John fit the part*)

For each abstract frame, three instantiations were generated by inserting lexical items, resulting in 102 frame instantiations. These 102 frame instantiations were then crossed with the 30 verbs to create 3060 total items.

Thirty lists of 102 items each were then constructed subject to the restriction that the list should contain exactly 3 instances of each verb and exactly 3 instances of each frame and that the same verb should never be paired with the same frame twice in the list. (That is, no verb showed up with more than one instantiation of the same frame in a single list.)

These lists were then inserted into an Ibex (version 0.3-beta17) experiment script with each sentence displayed using an unmodified `AcceptabilityJudgment`

controller (Drummond, 2014). This controller displays the sentence above a discrete scale. Participants can use this scale either by typing the associated number on their keyboard or by clicking the number on the scale. A 1-to-7 scale was used with endpoints labeled *awful* (1) and *perfect* (7). All materials, including the instructions participants received, are available on my github.

2.2.2 Participants

Ninety participants (48 females; age: 34.2 [mean], 30.5 [median], 18–68 [range]) were recruited through Amazon Mechanical Turk (AMT) using a standard Human Intelligence Task (HIT) template designed for externally hosted experiments and modified for the specific task. Prior to viewing the HIT, participants were required to score seven or better on a nine question qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $3.50.

2.2.3 Data validation

Even with the stringent requirements listed above—a qualification test, IP restriction, and high HIT acceptance rate—some participants attempt to game the

system. There are two main ways that participants do this: (i) submitting multiple HITs despite being instructed not to and (ii) not actually doing the task—e.g. choosing responses randomly.

The first is easy to detect. When data are submitted in Ibex, the submitting participant's IP address is converted into an MD5 hash, which is in turn associated with the responses they submit. This hash can then be used to check whether participants followed instructions in only submitting a single HIT. Two participants submitted multiple HITs: one participant submitted three and another submitted two. In both of these cases, only the first submission was used.[17]

The second requires more care to detect. Here, I use the fact that multiple participants did the same list. The idea is to compare each participant's responses against those of all other participants that saw the same list. If a participant has low agreement with the other participants that saw the same list and the other participants show high agreement with each other, then I conclude that the disagreeing participant was providing lower quality data and remove them from the analysis.

To implement this, the Spearman rank correlations between each participant's responses and those of every other participant that did the same list were calculated. For instance, if participants x, y, and z all did list 1, the correlation between x's and y's responses, x's and z's, and y's and z's was computed. The distribution of these correlations was then inspected for outliers.

The median Spearman rank correlation between participant responses is 0.64

[17]Note that this method does not distinguish between one participant attempting to submit multiple HITs from the same IP and two participants each submitting a single HIT from the same IP. I err on the side of caution in filtering all the but the first HIT from the same IP.

(mean=0.63, IQR=0.69-0.58). To find outliers, Tukey's method was used. Four comparisons fall below Q1-1.5*IQR and none fall above Q3+1.5*IQR. The four that fall below are due to two participants, each from a different list. Perhaps not coincidentally, those participants were also the ones that submitted multiple HITs. The remainder of the analyses exclude responses from these two participants.

After excluding these participants, the median remains the same (to two significant figures) and the mean shifts upward slightly, from 0.63 to 0.64. (This is to be expected since the mean is more sensitive to outliers.) The IQR becomes slightly smaller, and Q1 shifts slightly upward (IQR=0.69-0.59). These correlations are comparable to those reported by Fisher et al. (1991).

2.2.4 Results

In this section, I provide an exploratory analysis of the acceptability judgment data. The goal here is two-fold: first, to show the general contours of the data set; and second, to develop a model that extracts interpretable distributional features from the acceptability judgment data. I begin with (hard) hierarchical clustering of the verbs as a way of breaking into the data. One problem this method has is that it cannot capture overlapping categories. To capture such overlapping categories, I move to analyzing the data with factor analysis, which can capture overlapping clusters/features. Two factor analysis approaches are explored: Principal Component Analysis (PCA) and Nonnegative Matrix Factorization (NMF). I show that, generally, PCA does well at capturing high-level feature while NMF does well at

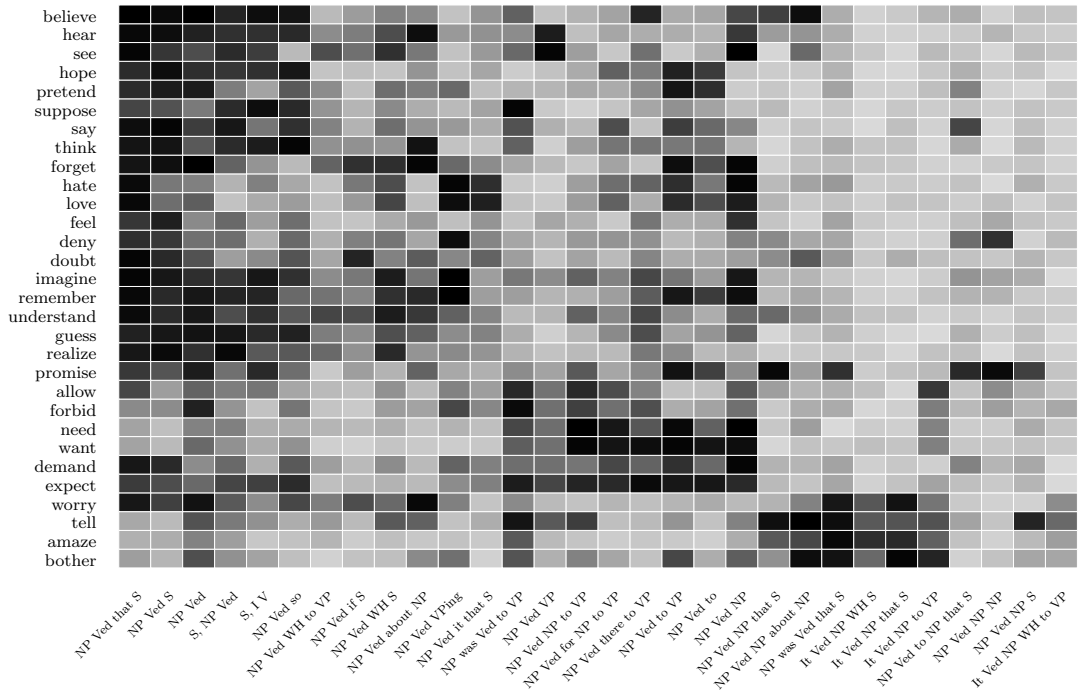


Figure 2.2: Mean rating for each verb-frame pair ordered by hierarchical clustering. Darker shades represent higher mean ratings.

capturing fine-grained features. I argue, however, that NMF does a better job of capturing theoretical intuitions.

2.2.4.1 A bird's eye view

Figure 2.2 displays the mean rating for each verb-frame pair.[18] Darker cells represent higher mean rating. The ordering along each axis is derived from a hierarchical clustering of the verbs (y-axis) and a separate hierarchical clustering of the frames (x-axis).[19] The hierarchical clustering of the verbs can be seen in Figure 2.3 and the hierarchical clustering of the frames, in Figure A.1 in Appendix A. These two display methods provide a bird's eye view of the verb clusters and the syntactic distributions that belie those groupings.

Three clusters of verbs are immediately clear from Figures 2.2 and 2.3. First, a major cluster emerges that tends to be good with finite complements (*believe, hear, see, hope, pretend, suppose, say, think, forget, hate, love, feel, deny, doubt, imagine, remember, understand, guess, realize,* and *promise*). These verbs appear to correspond roughly to Bolinger's (1968) representational class.

[18]Note that averaging in this way implicitly assumes that all points on the ordinal (likert) scale map onto (contiguous) intervals of equal measure on the latent scale (acceptability). This is not a valid assumption in general, since each participant in acceptability judgment experiments appears to use likert scales in slightly different ways—i.e. ordinal scale responses tend to exhibit scaling effects. (The existence of scaling effects in ordinal scale tasks has been well-known since at least Stevens 1946; see Schütze and Sprouse 2014 for a recent discussion of scaling effects in acceptability judgment tasks.) A 6 response for one person could be equivalent to a 7 response for some and a 5 response for others. This is taken into account explicitly in later sections by incorporating an ordinal logit model with participant random effects into the analyses; but for current purposes, it seems unlikely that this violation is problematic.

[19]For both clusterings, Euclidean distance, the default in the R function `dist()`, was used as the metric and complete linkage, the default in the R function `hclust()`, as the agglomerative clustering criterion. Neither choice is principled, but since the analysis in this section is mainly qualitative, I see no apparent problem with either.

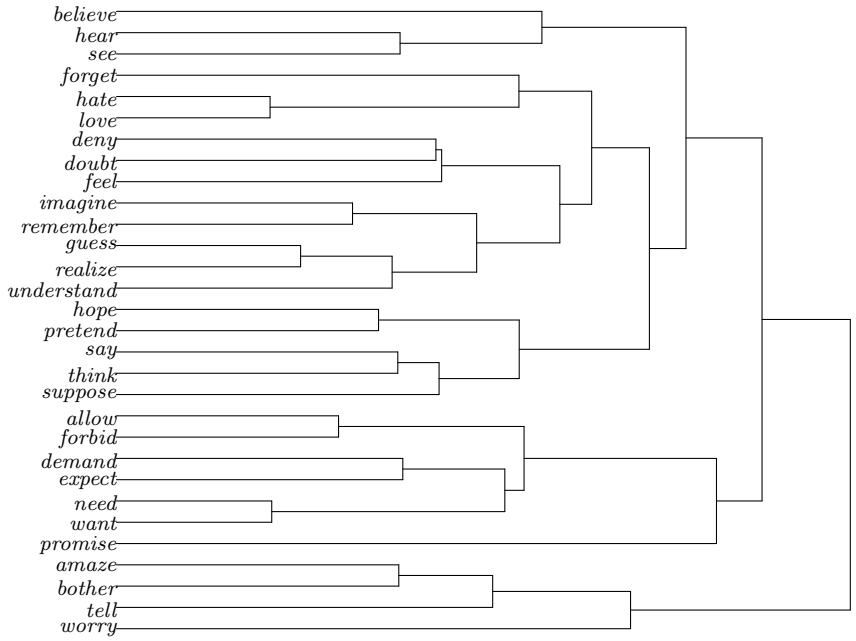


Figure 2.3: Hierarchical clustering of verbs based on data in Figure 2.2.

The second major cluster that arises contains verbs that involve a way of ordering states of affairs for their optimality with respect to some set of constraints: desires/needs (*want*, *need*), permits (*allow*, *forbid*), and commands (*demand*, *expect*). These verbs are also the ones that tend to take nonfinite complements. This second grouping is interesting in that it also turns out to be one of the most cohesive in terms of semantic similarity judgments, as I show in Section 2.3. For the remainder of the chapter, I refer to this cluster as the preferential class.

The final high-level cluster involves some verbs that encode emotion toward a state of affairs or object (*worry*, *amaze*, and *bother*)—but not others (*hate* and *love*)—along with the apparent outlier *tell*.[20] There are likely two reasons *tell* ends up in this cluster. First, since it is impossible in a standard acceptability judgment task to differentiate expletive subjects—e.g. *it*—from their referential counterparts, participants have the option of reading any of the frames that begin with *it Ved*... as though the *it* were referential. Second, since all of the frames constructed with an expletive subject also contained an NP object and only some of the other frames did, *tell* ends up globally more close to the seemingly semantically disparate emotion verbs.

As one digs further into smaller clusters within these three major ones, some regularities emerge, but some unexpected aspects arise as well. Among the regularities are pairs that clearly belong together semantically: *hate* and *love*, *hear* and *see*, *want* and *need*, *forbid* and *allow*, etc. Many of the intermediate groupings are some-

[20]The distinction the clustering finds between *worry*, *amaze*, and *bother*, on the one hand, and *hate* and *love*, on the other, is clearly driven by the fact that the former group takes their EXPERIENCER argument as an object and the latter takes their EXPERIENCER argument as a subject. To what extent this syntactic distinction is mirrored in the semantics is an open question.

what odd, however. In terms of its semantics, why should *believe* be grouped with perceptual verbs like *hear* and *see* and not other belief verbs like *think* and *suppose*. Similarly, why should *forget* be grouped with *hate* and *love* and not *remember* and cognitive verbs, like *realize* and *understand*?

Indeed, even the three high-level clusters are not immune to these sorts of questions. For instance, though it clearly takes multiple frames that involve nonfinite complements, *expect* is something of an outlier among the preferential; though it does seem to have a use involving obligations (12a), it also seems to have a use involving predictions (12b). This second use—which can be drawn out by making the referent of the embedded subject something that could not have obligations (Wurmbrand, 2014)—seems much more akin to the representational (see Portner and Rubinstein 2013 for discussion of *expect* and its relation to *wish* in languages like Spanish).

(12) a. I expect you to be here on time.

b. I expect the pizza to be a little late.

(Cf. I expect that the pizza will be a little late.)

The presence of *expect* in the preferential contrasts with a notable absence: the verb *hope*, which shows up among the representational verbs. That *hope* patterns with the representational on the macro level is interesting since, like *expect*, *hope* seems to share elements of its semantics with both the representational and the preferential (Anand and Hacquard, 2013; Portner and Rubinstein, 2013; Hacquard, 2014; Harrigan, 2015).

The problem here is that, in partitioning verbs into classes—even hierarchical

ones—one misses aspects of the verbs' meanings that seem to cross-cut these partitions; some verbs, like *expect* and *hope*, share aspects of their meaning with both verbs in the representational and the preferential. This results in clusters that look interpretable at the macro level but not at the intermediate and lower levels. I refer to this problem as the Overlap Problem and take it up in the next section.

2.2.4.2 Recasting the Overlap Problem

To approach the Overlap Problem, it will be useful to first recast it. Assume some verb-by-frame matrix $\mathbf{D}$ such that d_{ij} represents (an approximation of) the association between (e.g., acceptability of) verb i in frame j inferred from $\mathbf{X}$. For instance, Figure 2.2 is a visual representation of such a $\mathbf{D}$, estimated by averaging likert scale responses for each verb i and frame j.

The aim is then to find some $\mathbf{S}$ that encodes generalizations about the distributions encoded in $\mathbf{D}$. Suppose that $\mathbf{S}$ is represented as a matrix—one whose cells s_{ik} encode the association between verb i and property k. (For the moment, I leave vague both what these distributional features are and how to interpret the nature of these verb-feature associations, though both questions are addressed in turn.) What is needed, then, is some method f for inferring $\mathbf{S}$ from $\mathbf{D}$.

$$\mathbf{S} \overset{f}{\leftarrow} \mathbf{D}$$

(Hard) hierarchical clustering (HHC) is one such (family of) method(s), in the sense that nonterminal nodes in trees, such as the one in Figure 2.3, can be conceived

of as representing some relevant distributional regularities. These regularities can in turn be encoded in a matrix **S** in the following way: assign an index k to each node of the tree, and let $s_{ik} = 1$ denote that verb i is dominated by node k (associated with property k) and $s_{ik} = 0$ denotes that verb i is not dominated by node k (not associated with property k). The possible **S** produced by f for arbitrary **D** are subject to the following condition.[21]

$$|\{i : r_{im} = r_{in}\}| \in \left\{ 0, \min \begin{pmatrix} |\{i : r_{im} > 0\}| \\ |\{i : r_{in} > 0\}| \end{pmatrix} \right\}, \forall m, n$$

The Overlap Problem then arises in the following way. Taking the example from the last subsection, suppose (distributional) feature m is associated with the (semantic) representational (in some yet-to-be-defined way) and (distributional) feature n is associated with the (semantic) preferential (in some yet-to-be-defined way). If it is correct to characterize verbs like *expect* and *hope* as sharing meaning components with verbs in both the representational class and the preferential class and if this is tracked by the syntax (which it seems to be), both would be marked positively for features m and n. But the above constraint implies that, if features n and m have any overlap, the verbs that are associated with one, must be a subset of the verbs that are associated with the other. This, in turn, means that either all representational verbs would need to have preferential components or that all preferential verbs would need to have representational components. This could be true (cf. Heim's (1992) seminal analysis of *want*), but it begs the current question,

[21] Note that this itself is a generalization of "flat" hard clustering methods—e.g. k-means—which do not allow containment.

since *hope* and *expect* still fall onto either side of the relevant divide.

To solve the Overlap Problem, then, the representational constraint on **S** inherent to the HHC inference procedure f must be loosened. To do this, it will be useful to represent the structure of the procedure directly. Suppose f is the function that right-applies the linear map **Q** to its argument. Thus, **Q** maps from verb distributions to regularities underlying those distributions.

$$\mathbf{S} = f(\mathbf{D}) = \mathbf{DQ}$$

If one knew **Q**, it could be applied to **D** to get **S**. But **Q** is unknown, so it is necessary to infer it. One common way of finding such a **Q** is Principal Component Analysis (PCA). Indeed, the method often used to carry out PCA, Singular Value Decomposition, is the same one that underlies Deerwester et al.'s (1990) Latent Semantic Analysis/Indexing, which was in turn proposed as a model of lexical context-based word-learning by Landauer and Dumais (1997).

2.2.5 Principal Component Analysis

As a method for extracting regularities in data, PCA can be viewed as constructing a mapping **Q** that shifts the perspective on the data so that the most salient regularities are laid bare—where salience, here, is defined in terms of variance. It does this by using **Q** (often called the loading matrix) to rigidly rotate the datapoints (verbs), valued on some dimensions (syntactic contexts), using weight-

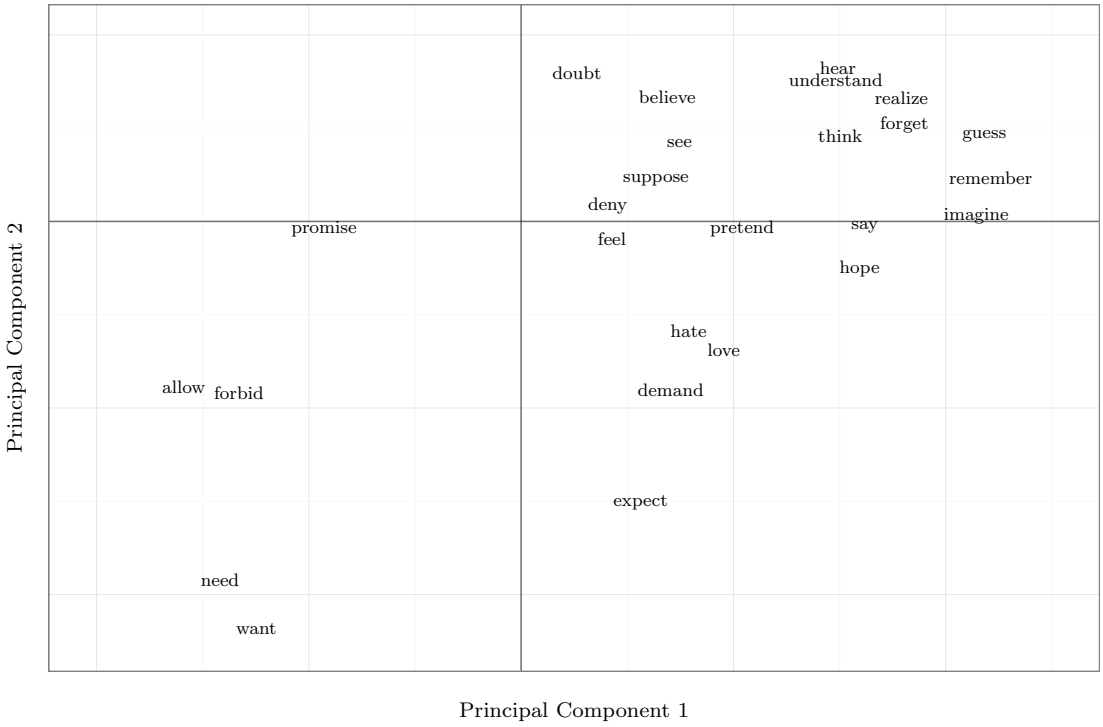


Figure 2.4: Verb embeddings on first and second principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$. Four verbs—*amaze*, *bother*, *worry*, and *tell*—are missing from this diagram due to their extreme values on these components. They lie far to the upper left.

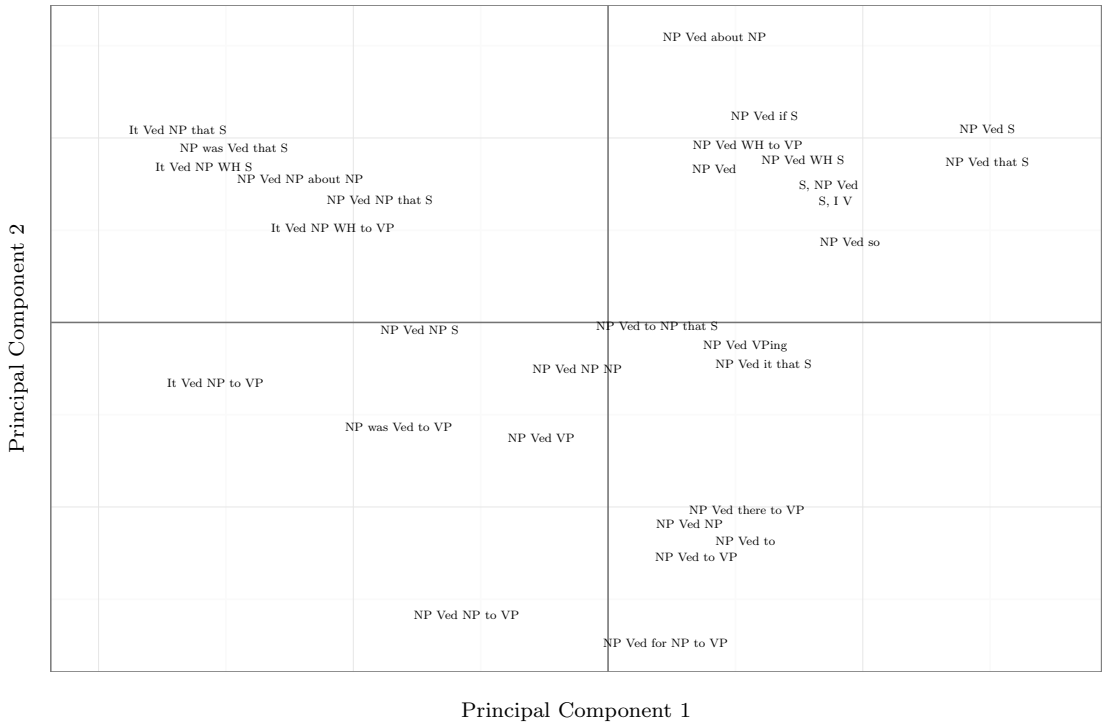


Figure 2.5: Frame loadings on first and second principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$.

ing combinations of those dimensions.[22] These weighted combinations of the original dimensions, derived (roughly) via the correlation between different dimensions (agreement between the columns of Figure 2.2), correspond to underlying dimensions called principal components. For instance, one principal component that PCA finds in the acceptability data positively weights frames like *NP Ved S*, *NP Ved that S*, *S, I V*, and *S, NP Ved* and negatively weights frames like *NP Ved NP S*, *NP Ved NP that S NP Ved NP to VP*, *NP was Ved to VP*, and all of the expletive subject frames. This information is encoded along the x-axis in Figure 2.5.

Each verb is in turn associated with a weight for each principal component, encoded in **S** (the score matrix). With respect to the principal component described above, the representationals (*think*, *believe*, *remember*, *forget*, etc.) tend to be positively weighted, and the preferentials (*allow*, *forbid*, *need*, *want*) tend to be negatively weighted. (I refine this generalization shortly.) This information is encoded along the x-axis in Figure 2.4.

The amount of variance along a dimension further provides a natural way of measuring the salience of a feature in the data, with the convention that the principal components are ordered by the amount of variance they explain in the original dataset. Figure 2.4 shows each verb's embeddings on the first and second principal components according to this ordering, and Figure 2.5 shows the relationship between each component and different syntactic frames. The dark grid lines on each figure represent the zero-intercept for each axis. These lines are useful since the

[22] As per standard practice, each dimension (the columns of Figure 2.2) were mean-centered and standardized prior to applying PCA.

quadrant a verb falls into determines whether it has a preference or dispreference for a specific frame given its value on a certain component.

For instance, *think* is in the positive-positive quadrant in Figure 2.4, and the frames *NP Ved S* and *S, I V* are in the positive-positive quadrant in Figure 2.5. This means that, given only its values on the first two principle components, *think* will prefer these frames. It will further outright disprefer the frames in the negative-negative quadrant. The converse is true of verbs in the negative-negative quadrant in Figure 2.4, such as *want*. These verbs will prefer frames in the negative-negative quadrant of Figure 2.5 and disprefer frames in the positive-positive quadrant.

It is important that these preferences and dispreferences are true given only the values of the verbs and frames on the first two principal components. This is because, even though a verb may show a preference for a frame via its score on one component, its score on other principal components may encode that verb's dispreference for that frame. In fact, this can be seen even on these two figures. Note that the frame *NP Ved to VP* shows up in the positive-negative quadrant in Figure 2.5. The extent to which *think* and *NP Ved to VP* are positive with respect to principal component 1 determines the preference of *think* for *NP Ved to VP* on that dimension of its meaning. Similarly, the extent to which *think* is positive and *NP Ved to VP* is negative with respect to principal component 2 determines the dispreference of *think* for *NP Ved to VP* on that dimension of its meaning. Taking into account just these two dimensions, *think* disprefers *NP Ved to VP*, though on other principal components, it may show a preference for it.

As noted, it seems that the first principal component corresponds to some-

thing like the classification seen earlier: representational verbs tend to be positive along the x-axis and preferential verbs tend to be negative.[23] Note that if this is a correct characterization of the first component, both *expect* and *hope* come out as representational. The second principal component appears to correspond to having an ordering component. The more negative, the more likely that the verb orders states of affairs for their optimality with respect to some set of constraints (desires, commands, etc.). Of particular interest here is the fact that both *expect* and *hope* are negative on this component. This is interesting because it suggests that there is evidence in the syntax that *expect* and *hope* share properties with both the representational and the preferential.

Figures 2.4 and 2.5 show only the two most salient underlying dimensions of the data. And as noted, very much the same split emerges that was seen in the hierarchical clustering in Figure 2.3, with the exception that the mixed status of *expect* and *hope* is now apparent. But PCA yields features beyond these first two, which explain only about half of the variance in the data. Indeed, PCA yields as many components as there are features (frames) in the original matrix. If all components are taken into account, $\mathbf{X}$ itself can be perfectly reconstructed; but since many components will explain little variance—i.e. are not very salient—taking them into account yields little insight into verbs' underlying features. For this reason, many methods for choosing how many components to keep—know as stopping rules—have been developed (see Jackson 1993 for discussion of various stopping rules). One

[23] *Tell* is again a notable exception here in that it is quite negative along this component despite its being a representational. This, again, is likely because it comes out good with the expletive subject frames due to the availability of a referential reading of the expletive *it*.

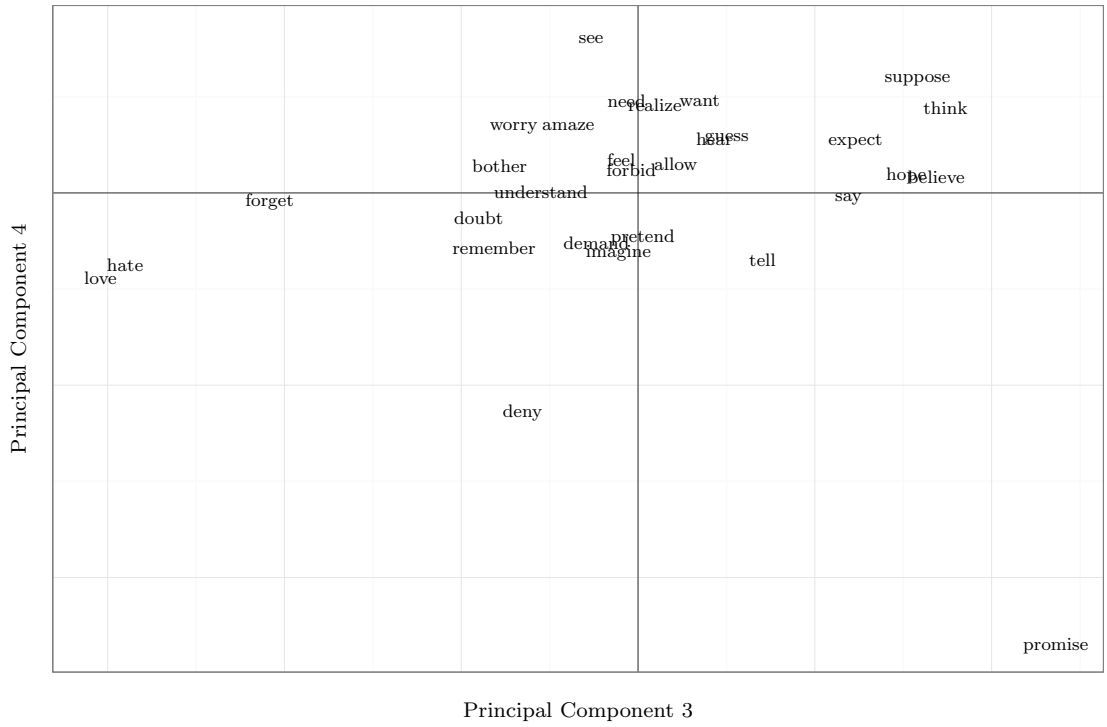


Figure 2.6: Verb embeddings on third and fourth principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$.

common stopping rule, known as the Kaiser-Guttman Criterion (Guttman, 1954), is to take only the components with associated eigenvalues great than 1. For this dataset, this yields 21 significant components. Another common rule is the total variance criterion. With a standard cutoff of 95%, this criterion yields 14 significant components.

Once one delves into these 14-21 later significant components, however, their interpretations, while still somewhat clear, become murkier. For instance, in Figures 2.6 and 2.7 the third principal component corresponds quite well to factivity. The factives *hate*, *love*, *forget*, *remember*, *understand*, *amaze*, *worry*, and *see* are all negative, and the nonfactives *suppose*, *think*, *expect*, *hope*, *believe*, *say*, *tell*, and *guess*

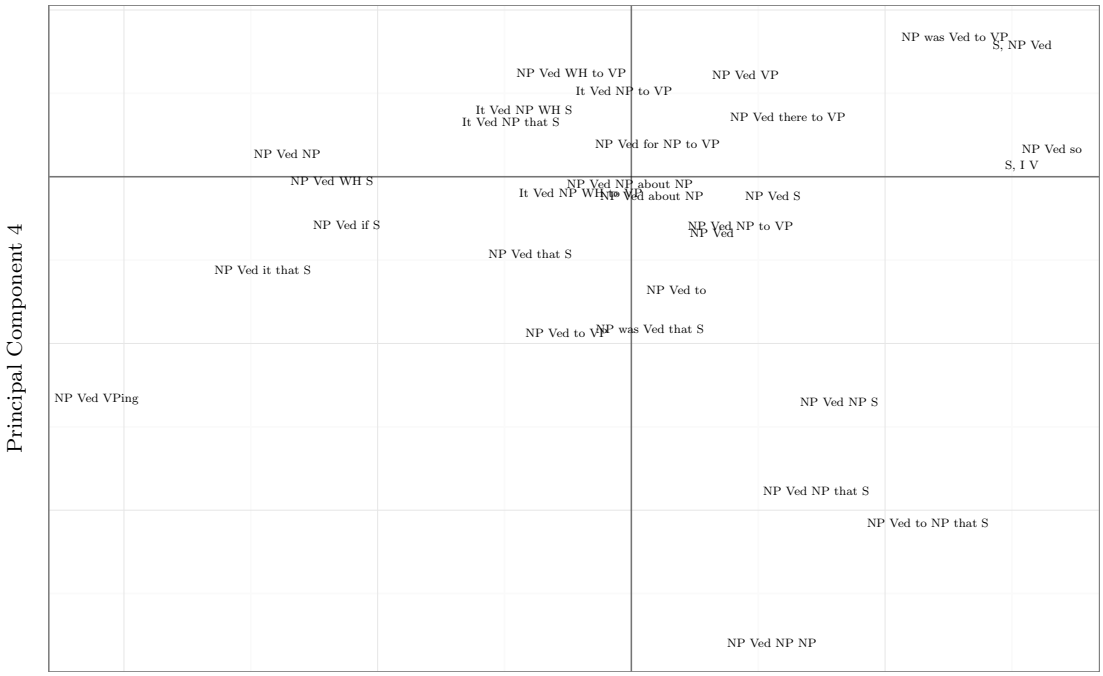


Figure 2.7: Frame loadings on third and fourth principal components of data in Figure 2.2. Dark gridlines are $x = y = 0$.

are all positive. Further, one noted distributional characteristic of factives—-that they occur with both question and nonquestion finite complements (Hintikka, 1975; Zuber, 1983; Ginzburg, 1995; Lahiri, 2002; Sæbø, 2007; Egré, 2008; Uegaki, 2012; Spector and Egré, 2014; Anand and Hacquard, 2014)—also obtains; both question (*NP Ved WH S*, *NP Ved if S*, *It Ved NP WH S*) and nonquestion (*NP Ved that S*, *NP Ved it that S*, *It Ved NP that S*) finite complements are negative on component 3. The murkiness arises when considering other verbs and frames that are negative on component 3. For instance, both *doubt* and *deny* are nonfactive yet negative on component 3,[24] and the frames *NP Ved NP VPing* and *NP Ved NP* show no clear relationship to the question-taking generalization.

This murkiness deepens when considering the fourth principal component. This component corresponds roughly to speech. Verbs like *promise*, *deny*, *tell*, *demand*, and *say* are negative on this component. These are not the only verbs that are negative, however. Nonspeech verbs like *remember*, *forget*, *imagine*, *pretend*, *hate*, and *love* also show up with negative scores on this component. Moving further into the next 10-17 significant principal components, this problem only worsens.

The reason for this murkiness is likely due to quirks of PCA. To fully appreciate these quirks, it is useful to first note an important property of the sorts of **Q** PCA produces: **Q** is orthogonal and thus right-invertible. Since **Q** is right-invertible, the following equality holds.

[24]This is actually a well-known issue in the factives literature. See Egré 2008; Spector and Egré 2014; Anand and Hacquard 2014 for recent discussion.

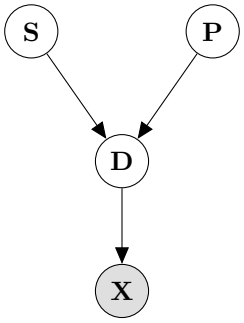


Figure 2.8: Graphical model corresponding to $S \xrightarrow{\mathrm{P}} D \rightarrow X$. (Same as Figure 2.1.)

$$\mathbf{D} = \mathbf{S}\mathbf{Q}^{-1} = \mathbf{S}\mathbf{Q}^{\top}$$

Thus, PCA can be viewed as performing factor analysis: given verb-by-frame matrix $\mathbf{X}$ find a verb-by-(latent) distributional feature matrix $\mathbf{S}$ as well as a (latent) distributional feature-by-frame matrix $\mathbf{P}$. Note that this way of viewing the problem now maps directly onto the framework discussed in Section 2.1. Replacing $\mathbf{Q}^{-1}$ with $\mathbf{P}$, the problem is one in which syntactic distributions $\mathbf{D}$ are the product of projection principles $\mathbf{P}$ "acting on" the features of $\mathbf{S}$. In this case, "acting on" is instantiated as mapping from one vector space—one in which the distributional features for words w_i fall—to another—one in which the distribution for words w_i fall. This can be represented in a graphical model as Figure 2.8.[25]

There is some sleight of hand going on here, though. How valid is it to replace $\mathbf{Q}^{-1}$ with $\mathbf{P}$? The answer is: not valid in the least. This replacement is potentially problematic in that it implies that $\mathbf{P}$ is invertibility. But to say that the projection principles $\mathbf{P}$ can be inverted to produce a syntax-to-semantics mapping $\mathbf{Q}$ is in fact

[25]This figure—identical to Figure 2.1 but repeated here for convenience—suppresses the prior parameters and (nuisance) parameters involved in generating $\mathbf{X}$ from $\mathbf{D}$.

to state a very strong thesis regarding the nature of projection. If (i) a given $\mathbf{P}$ were invertible and (ii) a noiseless representation of $\mathbf{D}$ were accessible, this would imply that $\mathbf{S}$ would be perfectly reconstructable.

I have no evidence for or against this invertibility assumption, though I think that there are strong methodological reasons for not making it. The reason is that, if $\mathbf{P}$ is invertible and that assumption is inherent to whatever procedure one carries out to find $\mathbf{S}$ and $\mathbf{P}$, then there is no harm done (assuming $\mathbf{D}$ is measured without noise, or at least well-estimated, by $\mathbf{X}$, which are themselves strong assumptions). But if $\mathbf{P}$ is not invertible and the invertibility assumption is inherent to whatever procedure is carried out to find $\mathbf{S}$ and $\mathbf{P}$, one may end up with poor results, since the correct $\mathbf{P}$ is, by definition, not in the procedure's codomain. As a methodological strategy, then, it seems that, if it is possible not to make the invertibility assumption, one should not. I point this out here for the following reason: if one attempts to induce some projection principle $\mathbf{P}$ using a factor analysis procedure like PCA, the true projection principles $\mathbf{P}$ (if they exist) could only be discovered if they are invertible. This suggests that one should seek a form of factor analysis that does not require the invertibility assumption.

Though it clarifies the relationship to projection, this does not yet explain why PCA produces the murkiness seen above. Lee and Seung (1999) note two possibly relevant properties. First, PCA learns holistic representations; each component (latent feature) in $\mathbf{Q}^{-1}$ has consequences for each observed feature in $\mathbf{D}$ (and thus $\mathbf{X}$)—i.e. most cells of $\mathbf{Q}^{-1}$ are not (close to) zero. This can be seen, for example, in Figure 2.5, where few of the frames have near-zero values on either component.

Second, PCA is greedy in the sense that it loads as much information as possible into as few of the initial principal components as possible. This is useful for compression, since only the most informative features can be saved with little loss of information, but it is not necessarily useful for analysis, since it may distribute many intuitively distinct features over a few components.

These properties of PCA solutions arise from two sources: (i) PCA allows both positive and negative values in $\mathbf{S}$ and $\mathbf{Q}^{-1}$ and (ii) PCA puts no constraints on the sparseness of its features—i.e. relative prevalence of (near) zero values in $\mathbf{S}$ and $\mathbf{Q}^{-1}$. This results in holistic representations because one latent feature's upweighting an observed feature can be directly counteracted by another latent feature's downweighting that same observed feature. It also results in uneven feature informativity (greediness) because latent features can make unfettered use of the real line as opposed to preferring (near) zero values, and thus the PCA computation loads as much information as possible into the initial features.

To ameliorate these problematic properties, I turn to a family of methods known as non-negative matrix factorization (NMF).[26] Like PCA, NMF attempts to factor $\mathbf{D}$ into $\mathbf{S}$ and $\mathbf{P}$. Unlike PCA, however, NMF puts slightly stronger constraints on $\mathbf{S}$ and $\mathbf{P}$; specifically, as the name implies, NMF requires both $\mathbf{S}$ and $\mathbf{P}$ to be non-negative. This results in a "parts-based respresentation" as opposed to a holistic representation, as in PCA (Lee and Seung, 1999). In such a parts-based representation, features target specific aspects of the observable distribution as opposed

[26]NMF methods have been used for some time in the document classification literature (cf. Xu et al., 2003) and are becoming popular for semantic representation (cf. Murphy et al., 2012; Fyshe et al., 2014, 2015).

to making slight alterations to every aspect. This comes about because latent features cannot specify which aspects of the observed distribution they disprefer, only the ones they prefer. In this case, they can only upweight—never downweight—the association between a verb and a frame.

One nice side effect of moving to NMF is that, though the constraints on the forms of **S** and **P** are tightened by disallowing negative values, the requirement that **P** be invertible is loosened. (This is not to say that NMF forces us into noninvertibility, since I could define a procedure that forced constrained **P** to be invertible, but this is not necessary. The procedure I use for conducting NMF will not enforce such a constraint for the methodological reason I discuss above.)

One choice that must be made in moving to NMF is what form **S** and **P** take beyond being non-negative. Typically, NMF is used to find non-negative real-valued **S** and **P**. One thing I have ignored up until this point that is relevant to the current choice point is the fact that positivity and negativity along a component are not the only things that matter when interpreting the results of factor analysis. Extent along that component must also be taken into account. For instance, *hate* and *love* are associated with the third component almost four times more strongly than *remember* is. This in turn has consequences for how strongly these verbs are associated with each frame: verbs that score more negatively on the third component—e.g. *hate* and *love*—are positively associated more strongly (as far as this component is concerned) with frames that load negatively—e.g. *NP Ved WH S*—than verbs that score less negatively—e.g. *remember*. But if these components—i.e. latent distributional features—plausibly correspond to latent semantic features—e.g. factivity—what

does it mean for the feature to be unbounded in this way? Moving to NMF does not help this, since in the typical case, NMF features are also unbounded.

To remedy this interpretational issue, I further constrain $\mathbf{S}$ beyond simple non-negativity (leaving the cells of $\mathbf{P}$ unbounded and non-negative). Since unboundedness in the verb representations is the problem, one route that could be taken is to assume the cells of $\mathbf{S}$ lie on some closed interval containing 0. A natural choice for such an interval is $[0, 1]$. Another is to put an even stronger constraint on the cells of $\mathbf{S}$ by forcing them to lie in $\{0, 1\}$, thus making $\mathbf{S}$ a binary mask. One benefit of this second route is that it allows us to introduce sparsity into the representation at a fundamental level: cells must be either 0 or 1.[27]

Except for the non-negativity constraint on the loading matrix $\mathbf{P}$, this brings the model quite close in form to those proposed by Griffiths and Ghahramani (2006) in a nonparametric context (see also the review in Griffiths and Ghahramani 2011 and references therein).[28] As noted by Navarro and Griffiths (2008), it is also closely related to Shepard and Arabie's (1979) ADCLUS model of similarity judgments, which assumes that the similarity between two objects is approximated by the weighted sum of the features they share (cf. Tversky, 1977). The importance of this second relationship arises in Section 2.4, where I construct mappings from $\mathbf{S}$ to similarity judgments $\mathbf{Y}$.

[27] As noted in the next section, I also induce sparsity in a less fundamental way: by introducing the equivalent of an L1 regularizer on $\mathbf{P}$ into the fitting procedure.

[28] Indeed, the parametric version of this model is quite similar to Smolensky's (1986) Harmonium, which has been recently revived under the name Restricted Boltzmann Machine (Hinton and Salakhutdinov, 2006; Salakhutdinov et al., 2007; Larochelle and Bengio, 2008; Hinton and Salakhutdinov, 2009; Coates et al., 2011) after the fact that they correspond to a constrained form of Boltzmann machine (Ackley et al., 1985).

2.2.6 Non-negative projection model

As before, specifying the forms of **S** and **P** does not tell us how to find them. In this case, I take a Markov Chain Monte Carlo (MCMC)-based sampling approach. But before moving onto the analysis of these fits, I first need to specify two further aspects of the model: a noise model and a way of selecting the correct number of features—i.e. a stopping criterion.

2.2.6.1 Noise model

One difficulty with NMF—and indeed, with most matrix factorization methods besides PCA—is the need to specify a noise model. Such a noise model is often necessary because an algorithm for discovering a specific sort of factorization may be intractable or non-existent without one.

Luckily, there is natural such model in this case. Note that, up until now, I have worked under the assumption that **X** was well-approximated by averaging acceptability judgments. This was useful for expediting a qualitative analysis of patterns in the judgment data. But as I noted in footnote 18, raw means are not technically appropriate for this kind of data due to random variation in scale use. This can lead to poor estimates of the true acceptability of a verb-frame pair. Two routes are often used to remedy this: (i) transformation of the data prior to analysis or (ii) modeling the relationship between the (unobserved) acceptability and the rating explicitly. This second route is the natural choice for the noise model.

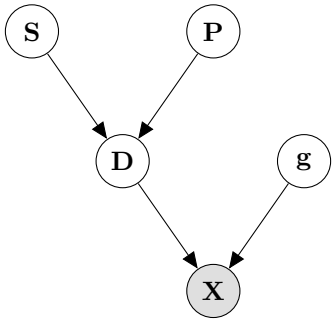


Figure 2.9: Graphical model corresponding to $S \xrightarrow{\mathrm{P}} D \rightarrow X$ with the addition of the ordinal logit mixed model parameters $\mathbf{g}$.

The specifics of this model will not concern us here,[29] besides to say that it requires some nuisance parameters $\mathbf{g}$, on which $\mathbf{X}$ is assumed to be dependent. Figure 2.9 shows the graphical model with the addition of these parameters. As with the parameters of interest, these (nuisance) parameters are sampled using MCMC.

2.2.6.2 Stopping criterion

Another obstacle for implementing the projection model is deciding on a reasonable number n of features to define the columns of $\mathbf{S}$ and the rows of $\mathbf{P}$. This is not a problem when using PCA, since the number of principal components will always be equal to the number of syntactic contexts in the original data. A common approach in these cases is to fit the model with many different values of n then assess the point at which adding features no longer improves the model's explanation of the data.[30] This is similar to the idea behind the Kaiser-Guttman criterion and

[29]The model is an ordinal logit mixed model with strictly nonnegative cutpoints and random effects for participants. The interested reader can find a formal specification in Appendix A and a computational implementation on my github.

[30]Another option would be to use a nonparametric prior over binary matrices, such as the Indian Buffet Process (Griffiths and Ghahramani, 2006, 2011). I have implemented such a version, but found convergence issue related to the discreteness of $\mathbf{S}$. Benjamin van Durme (p.c.) notes that this is a motivation for using unit-valued matrices, which remedy the "stickiness" of binary matrices

others described above for PCA, which tend to rely on amount of variance explained or some measure related to the variance. (In this sense, many of these criteria are comparable to the common adjusted R^2.) Since the model's noise component does not attempt to minimize a variance-based measure, however, a more general measure based on likelihood must be used.

The measure I use is known as the Watanabe-Akaike—or Widely Applicable—Information Criterion (WAIC), which Gelman et al. (2013) argue is preferable to other common information criteria used for comparing hierarchical models—e.g. the Deviance Information Criterion (DIC)—especially in cases where "the posterior distribution is not well summarized by its mean" (ibid, p. 182).[31] The current model is such a case since the posterior over $\mathbf{S}$ is multimodal; swapping columns i and j in $\mathbf{S}$ and rows i and j in $\mathbf{P}$ will result in a solution that is equally good in terms of both the posterior density and likelihood. WAIC attempts to approximate the results of Leave One Out Cross Validation (LOO-CV)—i.e. jackknifing—and has two components: (i) the log posterior predictive density (LPPD), which measures the model fit, and (ii) the WAIC estimate of the effective number of parameters (p_{WAIC}[32]), which measures the number of parameters that are doing explanatory work. The first term serves to measure the models fit to the data, and the second serves to penalize models for fitting the data too closely.

to some extent.

[31]Watanabe (2010) gives the first specification of this measure. See also Watanabe 2013 for discussion of the related Widely applicable Bayesian Information Criterion (WBIC).

[32]I use Gelman et al.'s second method of estimating p_{WAIC}, which they recommend because "its series expansion has a closer resemblance to the series expansion for LOO-CV and also in practice seems to give results closer to LOO-CV" (ibid, p. 174).

2.2.6.3 Model fitting

The MCMC sampler was implemented in Python using the version 2.3 of the `pymc` package (Patil et al., 2010). For each number of features n between 1 and 15, the sampler was randomly initialized and run for 11 million iterations with a burn-in of 1 million and a thinning interval of 10000. After each run, the traces of the loading matrix $\mathbf{P}$, response model parameters, and the deviance were analyzed for autocorrelation. In most cases, at least the deviance trace showed worrying amounts of autocorrelation. In those cases, the sample with the lowest deviance was extracted and the sampler was initialized with that sample and rerun using the same sampling parameters. This was repeated until only low lag (or no) autocorrelation was found in the traces. This took two repeats of this procedure for most values of n.[33]

To induce sparsity in the mapping matrix $\mathbf{P}$, exponential priors ($\lambda = 1$) were placed on each cell. This is analogous to L1 regularization, which can be derived in a Bayesian context by placing Laplace priors on the parameters. And if $x \sim \text{Laplace}(\lambda)$, then $|x| \sim \text{Exponential}(\lambda)$.[34]

2.2.6.4 Results

Figure 2.10 shows both the LPPD and the WAIC of the model (both scaled by -2) with each setting of n. (The p_{WAIC} for each model is proportional to the gap

[33] Ideally, the sampler would be rerun with sampler parameters an order of magnitude larger than the ones used. However, this is infeasible, given that even under GPU acceleration, the current procedure took 12 days.

[34] With this sort of regularization, NMF is sometimes referred to as Nonnegative Sparse Coding (NNSC Hoyer, 2002, 2004) or the related Nonnegative Sparse Embedding (NNSE Murphy et al., 2012).

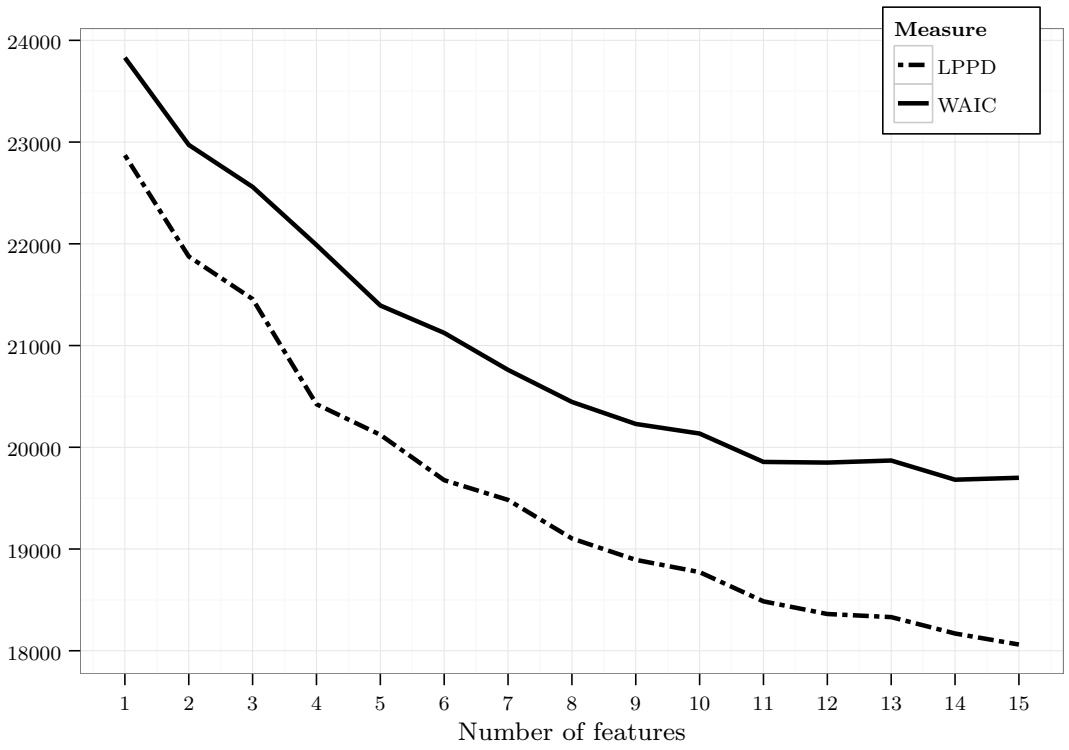


Figure 2.10: Log Pointwise Predictive Density (LPPD) and Watanabe Akaike (Widely Applicable) Information Criterion—both scaled by -2—for models with different numbers of features. The gap between the LPPD and WAIC lines is proportional to the effective number of parameters as computed by Gelman et al.'s (2013) second method (p_{WAIC2}).

between the two lines.) I see that WAIC bottoms (tops) out at 14 features, despite the fact that LPPD continues to decrease (increase). This continual decrease is expected, since adding features will never worsen the model fit. (A model with $n-1$ is a special case of the model with n features, where one feature in the model with n features is set to 0 for all verbs.) All further analyses therefore focus on the model with 14 features.

To analyze the output of the fit model, I first extract the sample with the minimum deviance (maximum likelihood) across all samples. Here, I focus only on

the parameters of interest, $\mathbf{S}$ and $\mathbf{P}$; a graph depicting the noise model parameters (ordinal logit cutpoints) can be found in Appendix A. Figure 2.11 shows the feature matrix $\mathbf{S}$ (analogous to the PCA score matrix) for this sample and Figure 2.12 shows the projection matrix $\mathbf{P}^\top$ (analogous to the PCA loading matrix).

Figure 2.11 shows that the non-negative projection models finds roughly four quite general features (1-4) with the remainder (5-14) being somewhat fine-grained. Further, the general features tend to project many syntactic contexts weakly, while the more specific features tend to project a few syntactic contexts very strongly. This is interesting in the sense that it looks like the general features have something like a baselining function: they serve to situate the verbs that have them in roughly the right part of distribution space, while the more specific features refine this placement.

In some sense, this is similar to the behavior of PCA, which also finds major features (principal components that explain a lot of variance) and then makes small refinements using later components. The important difference between PCA and the current model is that the small refinements appear to target specific classes with less noise.

This results in specific features like feature 14, which seems to target verbs of content of speech (*tell* and *promise*); feature 13, which seems to target the (experiencer object) emotive factives (*worry*, *amaze*, *bother*); feature 12, which seems to target implicative verbs (*remember*, *forget*, *bother*); feature 10, which (excluding *feel*) seems to target verbs involving permission/obligation (*demand*, *allow*, *forbid*,

deny[35]); and feature 8, which targets preferentials (*want*, *need*, *demand*, *expect*, *hate*, *love*).

Features 13 (*worry*, *amaze*, *bother*) and 14 (*tell* and *promise*) are particularly interesting in comparing the non-negative model to PCA. I noted above that, despite the fact that *tell* and *promise* appear to be good in the expletive subject contexts (*It Ved...*)—likely due to the fact that a nonreferential reading of the subject is not necessary—PCA discovers a component that seems to correspond to content of speech. That is, it separates *tell* and *promise* from other verbs. One problem with the PCA solution was that verbs like *imagine* and *remember* show up with about the same score as *tell* on the same component. This problem is remedied in the current case in that the content of speech verbs have the feature and the non-content of speech verbs don't; there is not equivocation with the binary representation (at least at each sample).

One place where PCA appears to do somewhat better is in the more general features. The general features discovered by PCA—i.e. principal components 1 and 2—corresponded quite well to previously described classes of verbs: representationals and preferentials. Indeed, this analysis even appeared to discover that some verbs fall into both classes. In contrast, the general features discovered by the non-negative model are somewhat muddled. For instance, feature 1 appears to be quite random; feature 2 corresponds roughly to representationality, but with the addition of *worry* and *demand* and to the exclusion of *hope* and *pretend*; feature 3 includes

[35] *Deny* may show up in this class because of its lack of permission use in the *NP V NP NP* context.

some, but not all, of the preferentials with the addition of many representationals; and feature 4 corresponds to a subset of representationals (again, with the addition of *worry*). In this more muddled solution, *hope* still cross-classifies with both *want* and *think* on different features, but it is unclear what to make of this, since the classes themselves are hard to interpret.

The likely reason for the more muddled solution in this case has to do with the nature of the weaker projective relationships found in **P**. Because those relationships are weaker, it is less costly for the model to posit that a verb has that feature rather than some of the more specific features, which tend to project more heavily over fewer syntactic contexts. This could result in the model being less certain about which verbs have a more general feature and which do not, thus meaning that the particular configuration found in the more general features in Figure 2.11 could be non-representative of the samples overall, which is always a problem with analyzing point estimates (and particularly discrete ones).

Unfortunately, due to the multimodality mentioned earlier, there is not much to be done about this here. Cells (and thus features) are not identifiable over the course of sampling, so it is not possible to estimate the probability of a verb having a particular feature. It is important to note, however, that, though this solution appears muddled from the point of view of the analyst, my own labellings of an algorithm's output are of no consequence in the larger scheme. Indeed, the logic I laid out in Section 2.1 explicitly eschews such labellings, at least as a matter of determining the amount of semantic information in syntactic distribution. And as I show in Section 2.4, this logic bears fruit, since at least some of the hard-to-

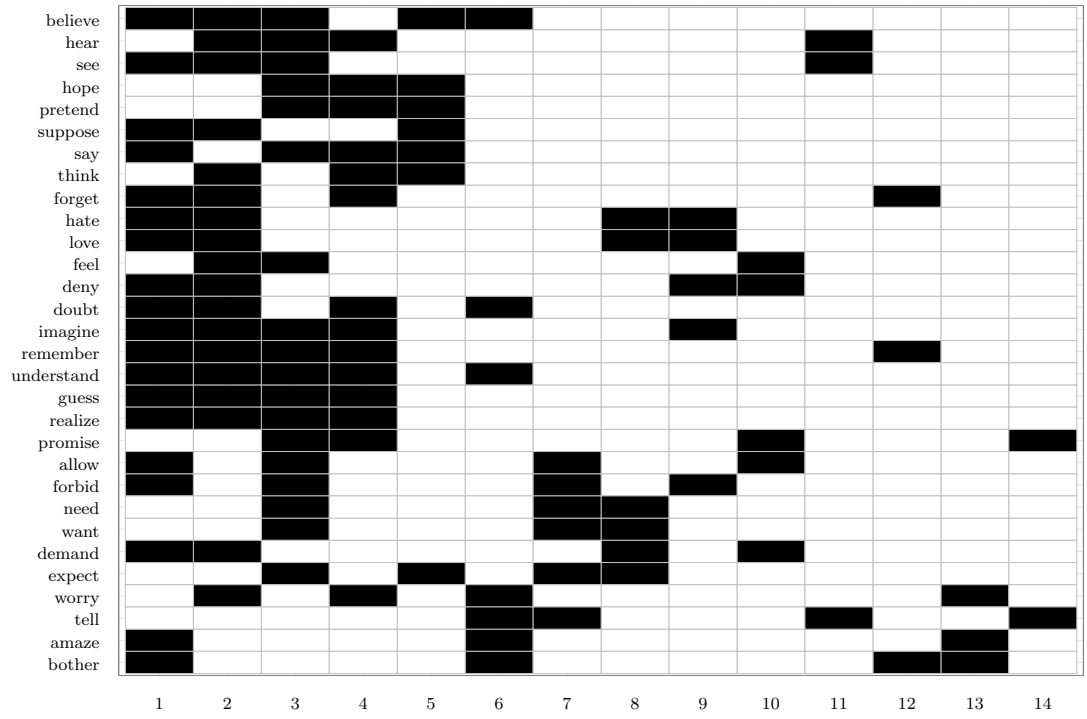


Figure 2.11: Verb features (**S**) inferred by non-negative projection model. Black cells represent 1s.

interpret features extracted using the non-negative model turn out to be predictive of participants' similarity judgments.

2.2.7 Discussion

In this section, I presented an experiment aimed at getting a measure of how acceptable a variety of propositional attitude verbs are in different syntactic contexts. My goal was two-fold. First, I assessed how closely the sorts of regularities found in these data correspond to the attitude verb distinctions discussed in Section 2.1. I carried this out by using two standard exploratory analyses—hierarchical clustering and principal component analysis—and one novel (at least in this do-

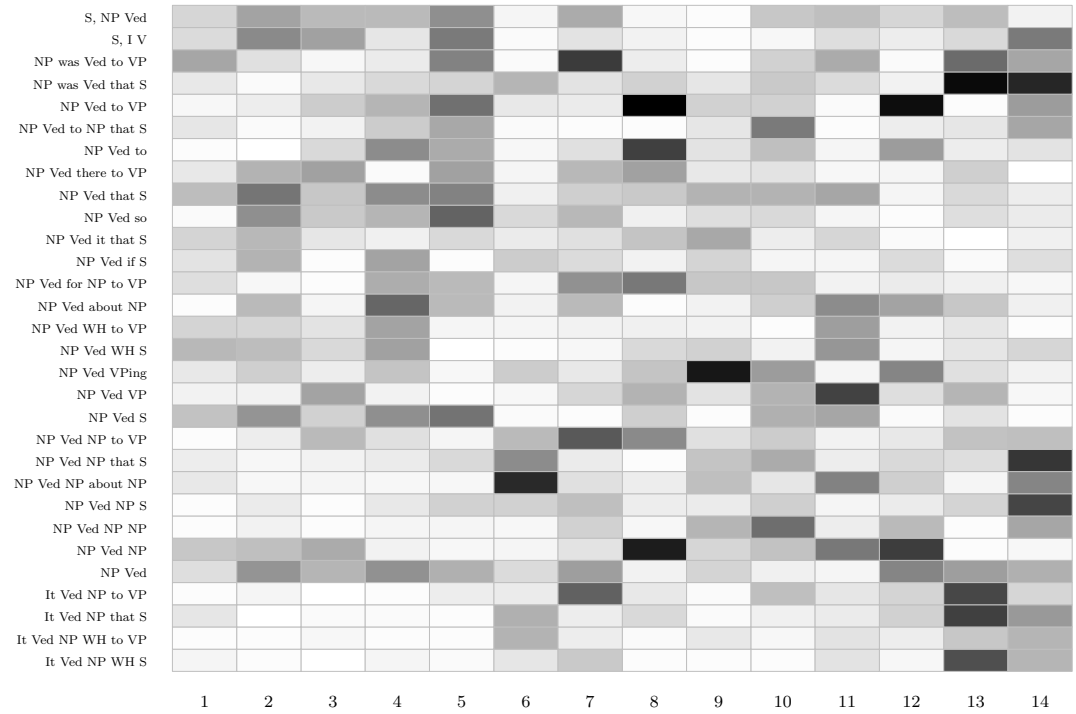


Figure 2.12: Relationship between features and syntactic frames ($\mathbf{P}^\top$) inferred by non-negative projection model. Darker cells represent larger values.

main) analysis—nonnegative matrix factorization. Second, I assessed the strengths and weaknesses of each exploratory analysis with respect to how well they satisfy various methodological considerations. This assessment was driven in part by the sorts of regularities I discovered in the data, along with theoretical considerations.

I came away with the following generalizations. First, the representationality distinction is robustly discovered by every method. I noted, however, that certain methods miss the fact that the representationals and preferentials may not be mutually exclusive. For instance, *hope* and *expect* seem to share aspects of both classes.

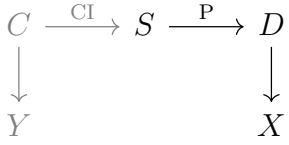
This lead us to the first methodological consideration: whatever method is used for extracting regularities should be able to represent overlapping classifications. This motivated the move to a method that produces representations that allow overlap. Principal component analysis (PCA) fit this bill. PCA was able to discover the fact that *hope* and *expect* have both representational components if the first two principal are interpreted as corresponding to representationality and preferentiality, respectively. This seemed plausible given the the quadrants in which various verbs that have been classified as such lie. An interesting distinction was also seen lying along the third principal component. This component appeared to correspond to factivity, and as has been claimed for factivity, this component involved loadings on both question and nonquestion complements.

As I moved into the later principal components, however, generalizations about the semantics that might be associated with that component became murkier. For instance, the fourth principal component appeared to correspond to content of

speech to some extent, but there were various aspects of this component that looked quite noisy. I argued that this murkiness likely arises from two properties of PCA that are undesirable for current purposes: PCA learns holistic representations and loads information greedily into initial components, thus making possibly distinct regularities indistinguishable. These properties, in turn, arise from two sources: the availability of both positive and negative values and a lack of representational sparsity. These considerations motivated us to move to a model that addresses them by constraining distributional regularities to be defined only in terms of non-negative matrices—our non-negative model of projection. Beyond remedying these problems, I noted that this model also satisfies a theoretical consideration pointed out in Section 2.1: mappings from semantics to syntax seem unlikely to be invertible.

The method by which the non-negative projection model explains the data is interesting. Much like PCA, it appears to find a few very general features that situate the verbs having those features within a general distributional space and many much more specific features that target particular areas of that space. Unlike PCA, the more specific features appear to less noisily target features like content of speech.

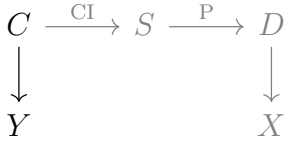
I would like to end this section by recapitulating how these results relate to the larger goals of the chapter. In Section 2.1, I gave a formal characterization of the logic I hew to throughout the chapter. I focused in this section on one small part of this logic—-that found in the lower right corner. This piece concerns how regularities might be extracted from syntactic distributions.



Much of the section consisted in exploring the consequences of making various choices for these regularity extraction procedures. This exploration was spurred on by two concerns: relating these results to previous literature in the domain of propositional attitude verb semantics and satisfying various methodological considerations. I would like to stress, however, that though these considerations are relevant to the ultimate question regarding what aspects of meaning syntactic distribution could be used to learn, I have not yet satisfied Fisher et al.'s methodological critique in that I nonetheless performed much of the exploratory analysis by attempting to label various regularities. To assist in filling out this part of the above logic, I turn in the next section to two methods for gathering a proxy for the semantics **Y**.

2.3 Experiments 2 & 3: verb similarity

In this section, I present two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. This fits into the running diagram as gathering data **Y** about the concepts **C**.



The first experiment (Experiment 2) employs a generalized semantic discrimination—or triad—task, in which participants are given lists of three words and asked to choose the one least like the others in meaning (Wexler, 1970; Fisher et al., 1991).[36] The second experiment (Experiment 3) employs an ordinal (likert) scale similarity task, in which participants are asked to rate the similarity in meaning of a word pair on a 1-7 scale.

There are two reasons I use both tasks. First, the generalized semantic discrimination task replicates the methodologies used in the Fisher et al. 1991 and Lederer et al. 1995, whose logic I employ in this chapter, and the ordinal scale task is a standard way of measuring similarities, regardless of the formal properties of the objects being compared. Second, I would like to assess to what extent the semantic properties employed in these similarity tasks are task-dependent. This may in turn suggest how close the representations $\mathbf{Y}$ used to make the judgments are to the concepts themselves.

In Sections 2.3.1 and 2.3.2, the experiments themselves are described along with some exploratory analyses that qualitatively relate their results to the results of Experiment 1. In Section 2.3.3, the two measures are themselves compared to

[36] I follow Fisher et al. (1991) here, but interestingly, a similar paradigm—using a larger set of words per trial—is put to use in Chang et al. 2009 to assess the coherence of topics in a topic model.

assess their agreement.

2.3.1 Experiment 2: generalized semantic discrimination task

2.3.1.1 Design

In this experiment, I aim to get a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. To do this, I constructed a list containing every three-combination of the 30 verbs from Experiment 1 (4060 three-combinations total). Twenty lists of 203 items each were then constructed by random sampling.

These lists were then inserted into an Ibex (version 0.3-beta15) experiment script with each three-combination displayed using an unmodified `Question` controller (Drummond, 2014). This controller displays an optional question above a list of answers. In this case, the question was omitted and the verbs making up each three combination constituted the possible answers. Participants could select an answer either by typing the number associated with each answer or clicking on the answer. All materials, including the instructions participants received, are available on my github.

2.3.1.2 Participants

Sixty participants (28 females; age: 34.5 [mean], 31 [median], 18–68 [range]) were recruited through AMT using a standard HIT template designed for externally hosted experiments and modified for the specific task. All qualification requirements

were the same as those described in Section 2.2.2. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $3.

2.3.1.3 Data validation

The data validation procedure is the same one described in Section 2.2.3 with the exception that I calculate Cohen's κ instead of Spearman's ρ.[37] The median Cohen's κ between participant responses is 0.45 (mean=0.45, IQR=0.52-0.37).[38] To find outliers, I use Tukey's method. No comparisons fall below Q1-1.5*IQR and none fall above Q3+1.5*IQR. Thus, I exclude no participants.

The median agreement here is quite a bit lower than the interrater agreement found by either Fisher et al. or Lederer et al. (Spearman's ρ=0.78).[39] This is likely driven by the fact that I am investigating a much smaller portion of the lexicon and thus am bound to find that participants have less certainty about which verbs are more semantically similar.[40]

Another possible contributor to this lower correlation is that Cohen's κ is more

[37]Both Fisher et al. and Lederer et al. compute Spearman rank correlations over count matrices of the form found in Figure 2.13. The method they use is not available to us without significant alteration since I collected data from more than two participants per list. Instead, I opt for a standard measure of interrater agreement here. This measure is preferable in any case since (i) it allows us to assess each participant's reliability at the same time as I assess overall agreement and (ii) it can be applied to the raw data instead of a statistic of the data, as in the cases of Fisher et al. and Lederer et al..

[38]An analysis of the distribution of Fleiss' κ (the multi-rater generalization of Scott's π) by list corroborates this analysis (median=0.45, mean=0.45, IQR=0.48-0.40).

[39]Fisher et al. report Spearman's ρ=0.81 (Exp. 1); 0.78 (Exp. 2); 0.76 (Exp. 3), 0.79 (Exp. 4), 0.72 (Exp. 5). Lederer et al. report Spearman's ρ=0.81.

[40]If this is indeed true, interrater agreement on this and other similarity judgment tasks could be a way of investigating the "semantic density" of a lexical neighborhood. Modeling reaction time, as a proxy for uncertainty, might also be fruitful in future research.

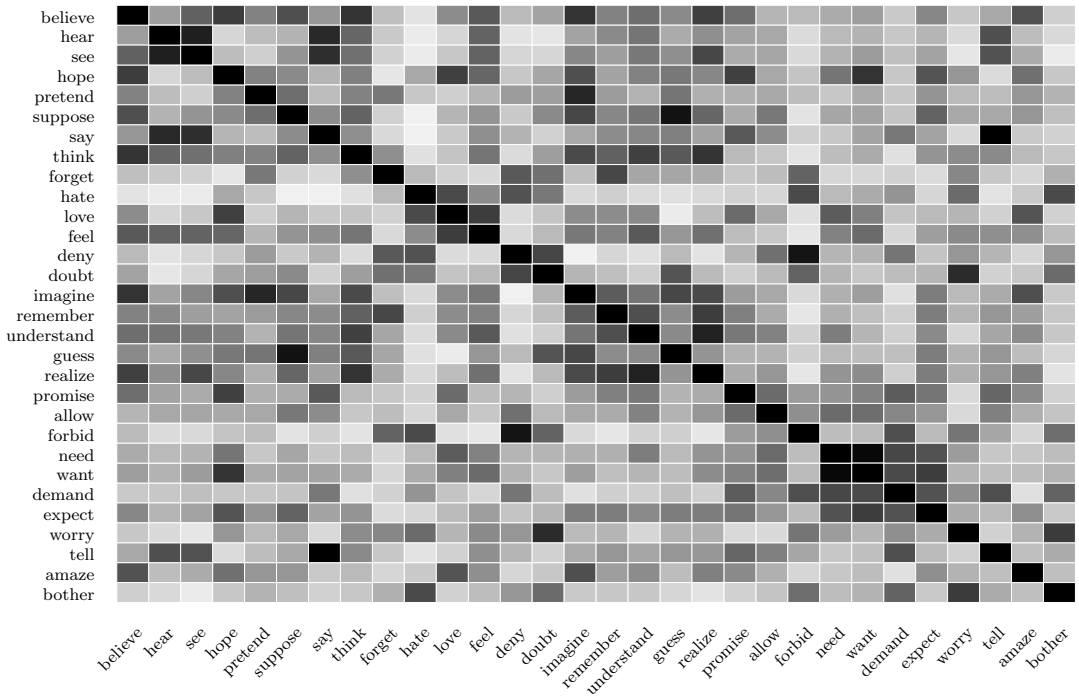


Figure 2.13: Similarity rating for each verb-verb pair from generalized semantic discrimination experiment. Darker shades represent more times chosen similar. Note that the diagonal elements are not observed and are set to the maximum over all other cells.

conservative than Spearman's ρ. As I see in Section 2.3.2.3, however, the conservativeness of Cohen's κ is not likely to be the culprit here, since even Spearman's ρ shows roughly the same amount of agreement among participants using a different measure.

2.3.1.4 Results

Figure 2.13 shows the number of times each pair of verbs occurred together and were not chosen as dissimilar. That is, if *believe*, *think*, and *want* occurred together, and *want* was chosen as the odd man out, then the similarity between

believe and *think* is incremented. Verbs are arrayed along the x- and y-axes in the same order they were on the y-axis in Figure 2.2. This arrangment allows for a visual assessment of the agreement between the syntactic clustering and the similarity judgments. Roughly, the more clearly dark blocks of cells appear in the graph, the more the syntactic clustering and the semantic judgments are in agreement.

Blocks along the diagonal suggest high agreement. There are roughly three blocks of size greater than three-by-three along the diagonal: a first group of representationals (*believe*, *hear*, *see*, *hope*, *suppose*, *say*, *think*), a second group of representationals (*imagine*, *remember*, *understand*, *guess*, *realize*), and a group of preferentials (*need*, *want*, *demand*, *expect*).

Blocks off the diagonal suggest that a larger group was split in two by a disagreement regarding some elements. The particular case of this I see in the Figure 2.13 is among the two blocks of representationals. What appears to be happening here is that participants did not rate representationals with negative affect (*forget*, *hate*, *deny*, *doubt*) as similar to the other representationals.

This effect of negative affect appears to be quite strong, as can be seen in Figure 2.14. This figure shows each verb's embedding derived from two-dimensional nonmetric multidimensional scaling (NMDS)[41] applied to the generalized semantic discrimination dissimilarity matrix (Shepard, 1962a,b; Kruskal, 1964a,b). This dissimilarity matrix is derived by counting the number of times a pair of verbs oc-

[41]Multidimensional scaling (MDS) maps from a distance matrix into an n-dimensional coordinate space such that the distances between elements in the n-dimensional space correspond as closely as possible to the distances listed in the distance matrix. The definition of "as closely as possible" determines the type of MDS carried out. NMDS results if this relationship is constrained only to be monotonic.

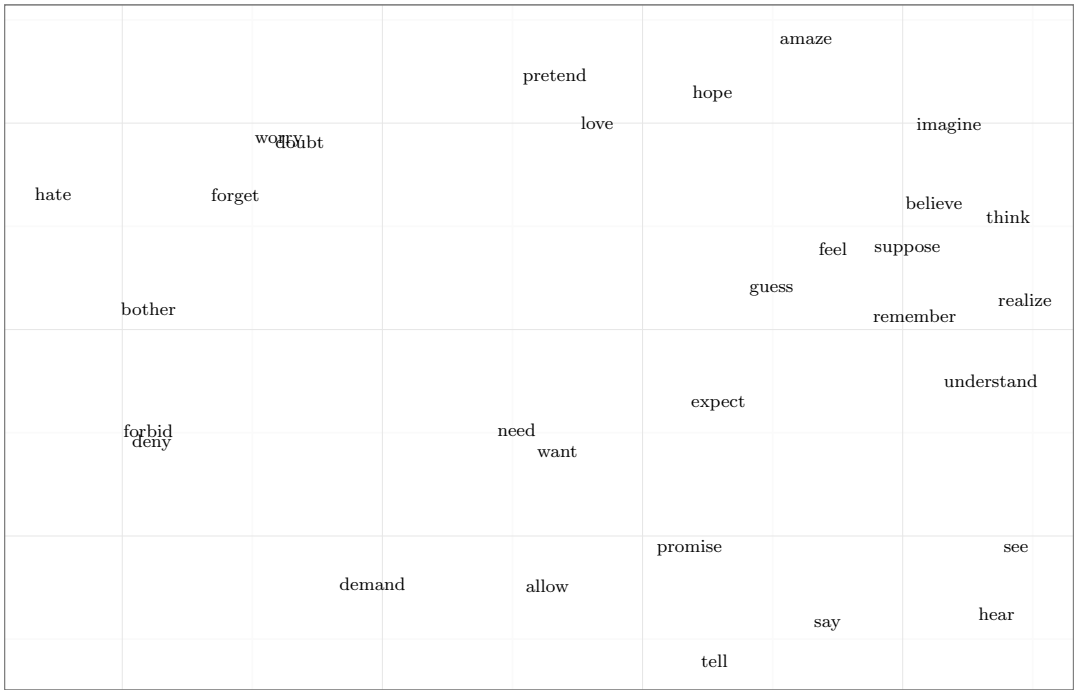


Figure 2.14: Embedding derived by two-dimensional nonmetric multidimensional scaling applied to the generalized semantic discrimination judgments represented in Figure 2.13.

curred together and one was chosen as the most dissimilar in the triad. In the *want*, *think*, *believe* example above, the dissimilarity between *want* and *think* and the dissimilarity between *want* and *believe* would both be incremented.

The effect of negative affect can be seen in the fact that verbs with such a component—*worry*, *doubt*, *forget*, *hate*, *bother*, *forbid*, *deny*—tend to cluster together (here, to the upper left). This sensitivity to affect mixes with a sensitivity to the clusters noted above and in acceptability judgments. The representational cluster (*believe*, *hear*, *see*, *hope*, *suppose*, *say*, *think*, *imagine*, *remember*, *understand*, *guess*, *realize*) can be seen to the right—with clear pockets of more fine-grained clusters, such as perception and speech (*tell*, *say*, *hear*, *see*) to the lower right and doxastic state to the upper right (*understand*, *remember*, *realize*, *suppose*, *think*, *believe*, *imagine*).

The two interesting cases discussed in Section 2.2.4—*expect* and *hope*—turn out to be interesting here as well. Note that *expect* falls midway between the representational cluster and what looks to be a preferential cluster (*need*, *want*, *demand*, *allow*). *Hope*, on the other hand, falls much further into representational territory. In fact, this layout is somewhat similar to that seen in Figure 2.4, where the embedding of *hope* on the first and second principal components of the acceptability judgments puts it closer to the representationals than *expect*.

2.3.2 Experiment 3: ordinal similarity

2.3.2.1 Design

As in Experiment 2, I aim to get a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. To do this, I constructed a list containing every pair of the 30 verbs from Experiment 1 along with the verb *know* (460 unordered pairs, 920 ordered pairs). Twenty lists of 62 ordered pairs were then constructed such that every verb was seen an equal number of times and no pair—either unordered or ordered—was seen twice.

These lists were then inserted into an Ibex (version 0.3.7) experiment script with each pair displayed using an unmodified `AcceptabilityJudgment` controller (Drummond, 2014). This controller displays the verb pair separated by a pipe character—e.g. *think* | *want*—above a discrete scale. Participants could use this scale either by typing the associated number on their keyboard or by clicking the number on the scale. A 1-to-7 scale was used with endpoints labeled *very dissimilar* (1) to *very similar* (7). To encourage them to make a symmetric similarity judgment, participants were instructed to rate "the similarity between the meanings of the two verbs" as opposed to rating how similar the first verb was to the second (or vice versa). All materials, including the instructions participants received, are available on my github.

2.3.2.2 Participants

Sixty participants were recruited through AMT. All qualification requirements were the same as those described in Section 2.2.2. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $1.

2.3.2.3 Data validation

The data validation procedure is the same one described in Section 2.2.3. The median Spearman rank correlation between participant responses is 0.40 (mean=0.41, IQR=0.52-0.32). To find outliers, I use Tukey's method. No comparisons fall below Q1-1.5*IQR and none fall above Q3+1.5*IQR. Thus, I exclude no participants.

2.3.2.4 Results

Figure 2.15 shows the mean similarity rating for each pair of verbs, collapsing over the two orderings (before or after the pipe character) in which the verbs in the pair were presented. As in Figure 2.13, verbs are arrayed along the x- and y-axes in the same order they were on the y-axis in Figure 2.2, so as for that figure, blocks along the diagonal suggest high agreement and blocks off the diagonal suggest that a larger group was split in two by a disagreement regarding some elements.

There are roughly three blocks of size greater than three-by-three along the diagonal: a first group of representational verbs (*believe*, *hear*, *see*, *hope*, *suppose*, *say*, *think*, *know*), a second group of representational verbs (*imagine*, *remember*, *under-*

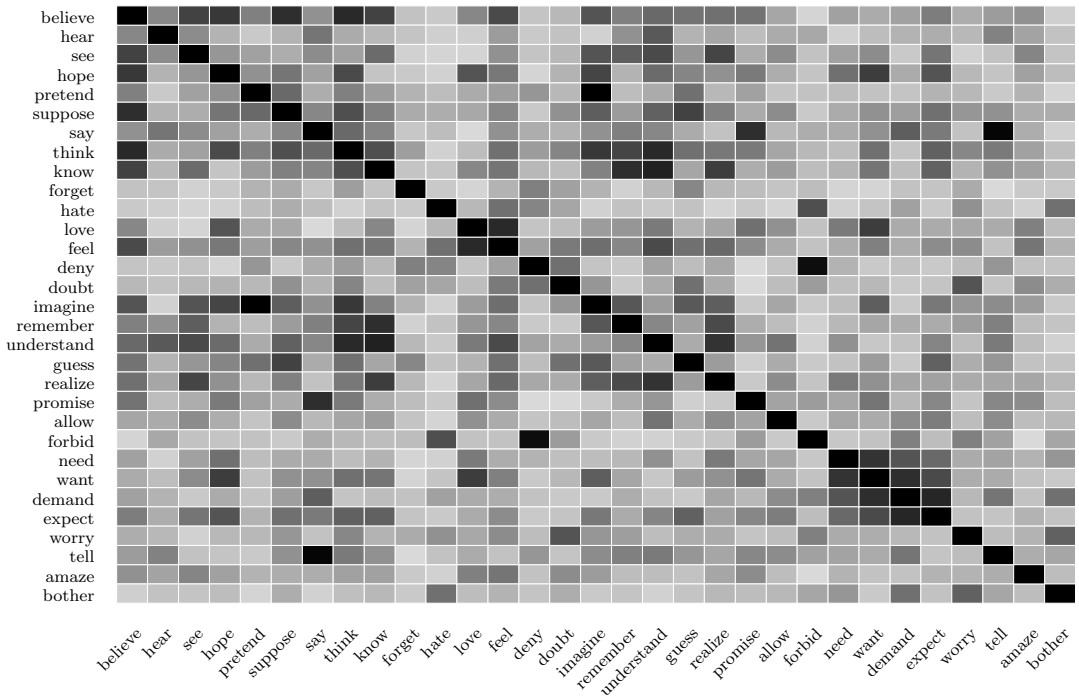


Figure 2.15: Similarity rating for each verb-verb pair from ordinal scale experiment. Darker shades represent higher mean ratings. Note that the diagonal elements are not observed and are set to the maximum over all other cells.

stand, *guess*, *realize*), and a group of preferential verbs (*need*, *want*, *demand*, *expect*). Each of these three corresponds to a block found in the last experiment, suggesting high agreement between this task and the generalized semantic discrimination task.

Also as before, there are two blocks of representational verbs off-diagonal, and again, what appears to be happening here is that participants did not rate representational with negative affect (*forget*, *hate*, *deny*, *doubt*) as similar to the other representational verbs. Thus it seems that negative affect is again having a large effect on participants' judgments.

This is corroborated in Figure 2.16, which shows each verb's embedding derived from two-dimensional NMDS applied to the likert distance matrix. This distance matrix is derived by subtracting each cell in Figure 2.15 from 7, thus essentially inverting the likert scale.

The effect of negative affect can be seen in the fact that verbs with such a component—*worry*, *doubt*, *forget*, *hate*, *bother*, *forbid*, *deny*—tend to cluster together (here, encompasses the entire left side of the diagram). This sensitivity to affect mixes with a sensitivity to the clusters noted above and in acceptability judgments. Some of the representational verbs (*believe*, *see*, *suppose*, *think*, *imagine*, *remember*, *understand*, *guess*, *realize*) can be seen to the upper right. Some of the clear pockets of more fine-grained clusters, like perception (*hear*, *see*, *feel*) are broken up, while others, like speech (*tell*, *say*), remain coherent.

Finally, the two interesting cases discussed in Section 2.2.4—*expect* and *hope*—turn out to be interesting here as well. They fall next to each in the lower right of the diagram, directly between the the representational verbs and the preferential

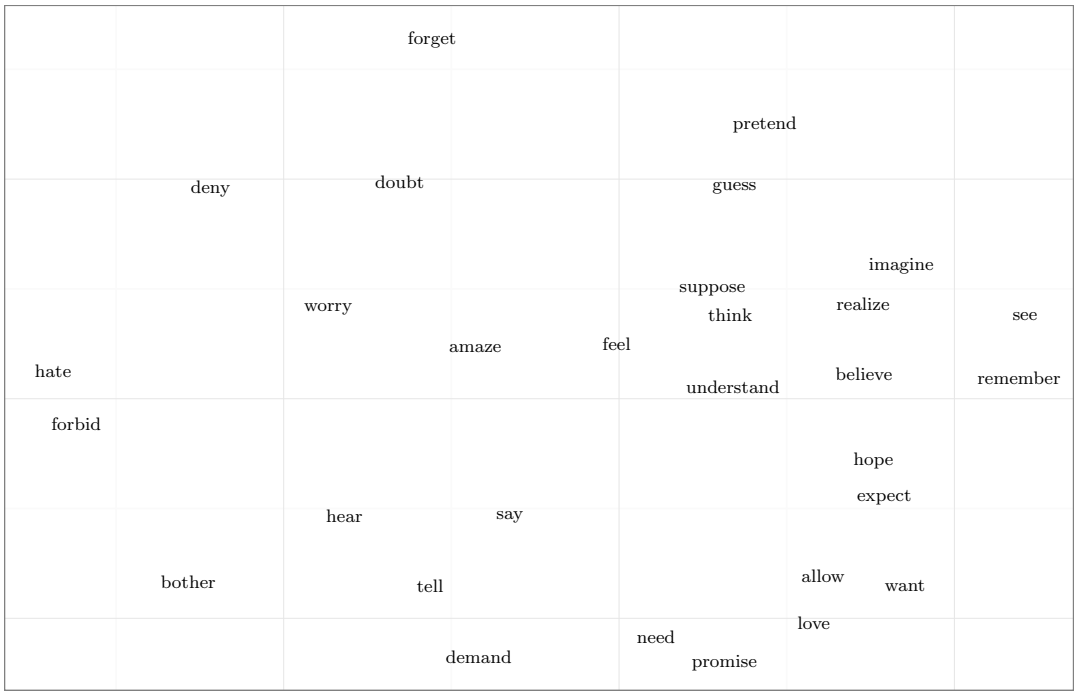


Figure 2.16: Embedding derived by two-dimensional nonmetric multidimensional scaling applied to the ordinal scale judgments represented in Figure 2.15.

verbs.

2.3.3 Comparison of (dis)similarity datasets

In the previous two subsections, I noted quite a few points of agreement between the two semantic similarity measures. And indeed, overall, the correlation between responses on the generalized semantic discrimination task and those on the likert scale task are quite high (Pearson's r=0.76, $p < 0.001$).[42] This suggests that these two task are tapping the similar aspects of participants' semantic knowledge. But though these measures tend to tap similar semantic knowledge, they diverge in

[42]If the likert scale judgments are z-scored or ridit scored prior to averaging, the correlation goes up slightly (Pearson's r=0.79 for both transformations).

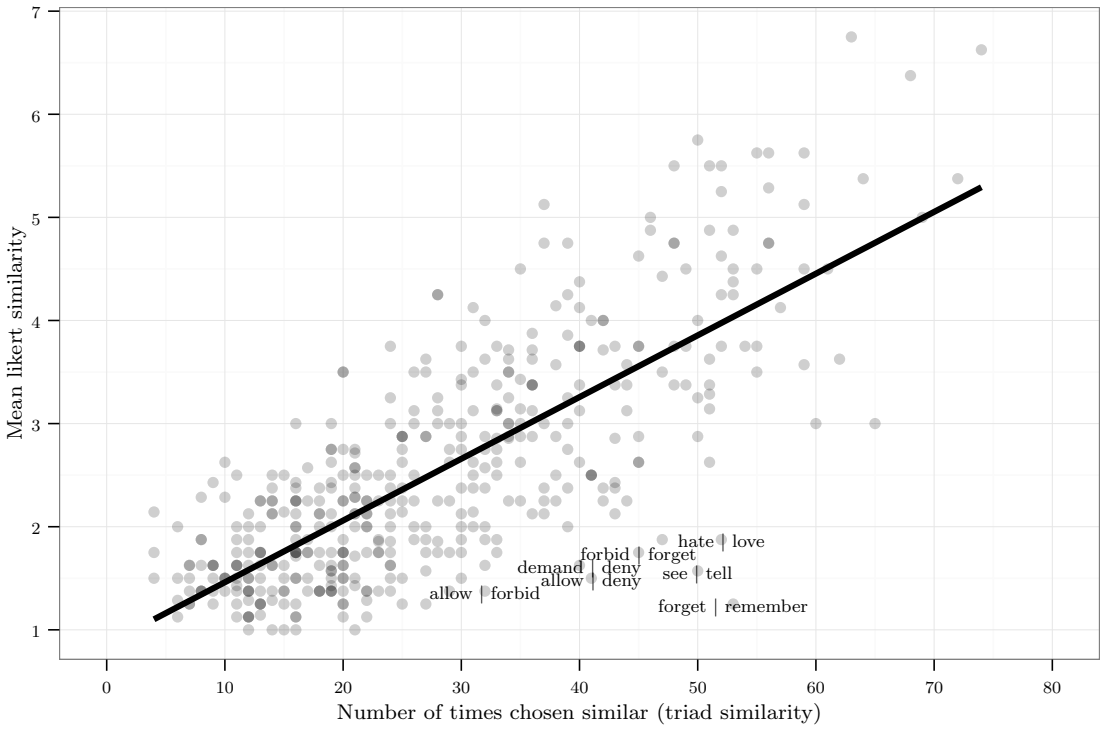


Figure 2.17: Relationship between generalized semantic discrimination similarity responses and ordinal scale similarity responses. Only low outlier pairs are labeled. (See Table A.1 in Appendix A for high outlier pairs.)

some respects. This can be seen in Figure 2.17, which plots the number of times each verb pair was rated similar in the generalized semantic discrimination task against the mean likert scale rating for that pair. The line superimposed on this graph gives a robust regression fit to these data. If a pair is above the line, it was rated higher than expected in the likert scale task given its rating in the generalized semantic discrimination task. If a pair is below the line, it was rated higher than expected in the generalized semantic discrimination task given its rating in the likert scale task.

Though much of the variability around the line is likely noise in participant responses, I also see what may be regularities. To investigate these regularities,

I compute each pair's standardized residual with respect to the robust regression fit. The variance in the data is likely heteroscedastic with respect to the generalized semantic discrimination ratings (studentized Breusch-Pagan = 135.54, $p < 0.001$), and thus standardizing by the residual standard error inferred by the model is not warranted. Instead, I infer a scedastic function conditioned on the generalized semantic discrimination ratings by fitting a generalized linear model with inverse-gamma link to the absolute value of the residuals. I then standardize each residual with respect to the robust regression fit by the inverse-gamma model's prediction. Table 2.1 shows all pairs whose residual is 2.5 standard deviations below the mean according to this method—i.e. rated more similar in the generalized semantic discrimination task than expected given the likert ratings. Table A.1 in Appendix A shows all pairs whose residual is 2.5 standard deviations above the mean. I focus on the outliers in Table 2.1, as the pattern there is clearer.

Many of these pairs appear to be antonymous along some dimension of their meaning. For instance, the antonymous pairs *remember* | *forget*, *hate* | *love*, *allow* | *forbid*, and *allow* | *deny*[43] are rated quite highly in the generalized semantic discrimination task but quite low in the likert scale task. Pairs like *demand* | *deny* and *see* | *tell* don't seem to immediately fit this generalization.

It is less clear what is happening in the case of *see* | *tell*, but one possibility for *demand* | *deny* is that participants are simultaneously contacting a distinction in the SOURCE and GOAL roles and negation on the modal quantifier. If x demands

[43]Note that the sense of *deny* that participants are likely getting in this case is the one that comes out in the double object frame and not the clausal complement frame. That is, denying someone something entails not allowing that person that thing.

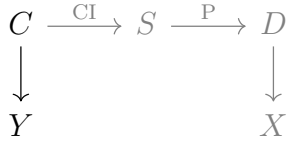
verb 1	verb 2	Standardized residual
forget	remember	-3.23
see	tell	-2.86
allow	deny	-2.79
demand	deny	-2.55
forbid	forget	-2.55
allow	forbid	-2.54
hate	love	-2.50

Table 2.1: Pairs rated more highly in the generalized semantic discrimination task than in the likert scale task.

y from z, z is the SOURCE and x is the GOAL with respect to x's demands, a strong deontic modality; in contrast, if x denies z y, x is the SOURCE and z is the GOAL with respect to x's denials, the negation of a strong deontic modal.

2.3.4 Discussion

In this section, I presented two experiments aimed at getting a measure of how similar in meaning naïve speakers take the propositional attitude verbs from Experiment 1 to be. Within the logic laid out in Section 2.1, this corresponds to getting two different proxies $\mathbf{Y}$ for the semantics.



The first experiment (Experiment 2) employed a generalized semantic discrimination task, in which participants are given lists of three words and asked to choose the one least like the others in meaning. The second experiment (Experiment 3) employed an ordinal (likert) scale similarity task, in which participants are asked to

rate the similarity in meaning of a word pair on a 1-7 scale.

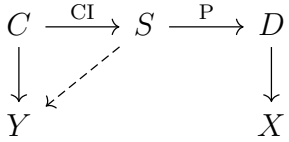
In the data from both tasks, I showed that, qualitatively, participants appear to be sensitive to the representationality distinction in their similarity judgments, though the representational cluster appears to be split in both cases by participants' sensitivity to negative affect. By comparing the results of the two experiments, I suggested that some aspect of meaning involving negation also appears to be differentially accessed in each task. In particular, participants in the ordinal scale task were much more sensitive to antonymy than participants in the generalized semantic discrimination task. In the next section, I present multiple ways of quantifying the relationship between the results presenting in this section and those presented in Section 2.2.

2.4 Quantifying the syntax-semantic connection

Having now presented qualitative aspects of all three datasets—albeit, using formal means—I explore in this section various ways of quantifying the relationship between the acceptability judgments presented in Section 2.2 and the semantic similarity judgements presented in Section 2.3. I do this in two ways. First, following Fisher et al., I assess the overall correlation between the acceptability judgments and the semantic similarity judgments.

$$\begin{array}{ccccc} C & \xrightarrow{\text{CI}} & S & \xrightarrow{\text{P}} & D \\ \downarrow & & & & \downarrow \\ Y & \longleftarrow\!\!\cdots\cdots\cdots\!\!\longrightarrow & & & X \end{array}$$

Second, in order to discover more fine-grained relationships, I map from the regularities **S**, extracted from the acceptability judgments, to the semantic similarity judgments **Y**.



I begin with the basic correlational analyses. This provides an overall measure of the extent to which syntactic distribution correlates with the semantics (making the relevant assumption laid out in Section 2.1). I then move into the more sophisticated analyses involving mapping from **S** to **Y**. This second makes it possible to delve into the relationship between the distributional regularities extracted using the non-negative projection model and the similarity judgments directly.

I present two such analyses. The first assesses both the nature of participants perception of semantic similarity relative to the distributional features and the salience of those features. Salience in this case is quantitatively instantiated as feature weights in the mapping model(s). This in turn gives us a way of assessing to what extent the distributional features **S** might be related to the true linguistically relevant semantic features, which in the formal sketch presented in Section 2.1 are a (homomorphic) function of the conceptual space **C**. The second uses the resulting models to assess how useful different syntactic contexts could plausibly be for a learner in discriminating the semantically relevant distinctions.

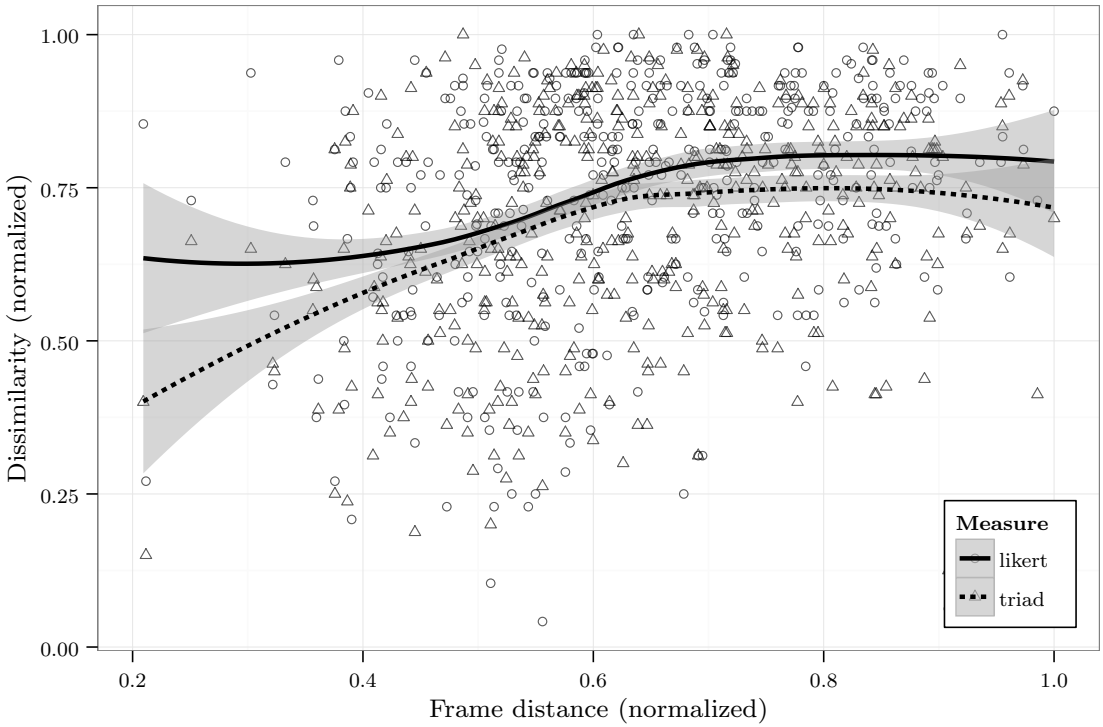


Figure 2.18: Relationship between frame distances based on acceptability judgment data present in Section 2.2 and dissimilarities based on the similarity judgment data presented in Section 2.3. Lines show local regression fits.

2.4.1 Basic correlational analysis

To get a measure of the overall correlation between the acceptability and similarity judgments, it is necessary to define a measure of distance over the acceptability judgments. In Section 2.2 I used Euclidean distance, viewing each verb as a point in 30-dimensional space, with each dimensions corresponding to a frame (see footnote 19). This choice of distance measure is arbitrary but suitable for a first pass.

Figure 2.18 plots these acceptability-based distances against the corresponding dissimilarity for each pair. Both axes are normalized by dividing each distance/dissimilarity by the maximum for that measure, which is necessary in order

to compare the generalized semantic discrimination and ordinal datasets on the same axis. Superimposed on these points are local regression fits, regressing the generalized semantic discrimination and ordinal scale dissimilarities on the acceptability-based distances.[44]

The correlation between the acceptability-based distances and both the generalized semantic discrimination dissimilarities (Spearman ρ=0.27; Mantel(iter=10000), $p < 0.001$) and the ordinal dissimilarities (Spearman ρ=0.28; Mantel(iter=10000), $p < 0.001$) are significant and comparable. Much of this positive correlation appears to be driven by agreement on verbs that are close together on both measures, with less agreement regarding those that are further apart. This suggests that syntactic distribution may be useful in signaling to learners which verbs are similar but not which ones are dissimilar. One possible implication of this is that a learning mechanism is only licensed in making conclusion about how well two verbs features match, not how much they mismatch.

A similar analysis can be run over the binary features extracted using the non-negative factor analysis model seen in Figure 2.11. Here again, a distance measure must be defined. Perhaps the simplest in this case is Hamming (Manhattan) distance, which corresponds to simply counting up the number of features that two verbs mismatch on.[45] The correlation between these feature-based distances and both the generalized semantic discrimination dissimilarities (Spearman ρ=0.27;

[44]The default parameters for the `loess()` function were used.

[45]Another way of conceptualizing this distance, which will be useful for understanding the analysis in the next section, is geometric. Suppose that each verb lies on the vertex of a unit hypercube, where $\mathbf{s}_i$ (row i of Figure 2.11) identifies which vertex the verb lies on. Then, the Hamming distance between verb m and verb n in feature space corresponds to counting the number of edges of the hypercube that one would need to traverse to get from $\mathbf{s}_m$ to $\mathbf{s}_n$.

Mantel(iter=10000), $p < 0.001$) and the likert dissimilarities (Spearman ρ=0.25; Mantel(iter=10000), $p < 0.001$) are again significant and comparable. Further, they hew quite closely to the distances computed from the raw data itself. This suggests that very little, if any, of the semantic similarity information in the original data was lost in abstracting the 14 binary features from the judgments over the 30 frames.

2.4.2 Distributional features and similarity

Looking at overall correlations does not yet tell us which distributional features appear to be active in the similarity judgments (or at least correlated with such an active feature). To carry this analysis out, I run two different types of regressions: a multinomial regression, in which the generalized semantic discrimination judgments are regressed on (similarities derived from) the features of the verbs included in each triad, and an ordinal regression, in which the ordinal scale judgments are regressed on (similarities derived from) the features of the verbs included in each pair.

Each of these analyses requires that I define some way of deriving similarities from the binary features extracted in Section 2.2. There are a few natural ways of doing this. The first is to take the raw additive inverse of the Hamming (Manhattan) distance in N (the number of features). This maps straightforwardly onto the basic correlational analysis in the last section in that the correlation between this measure and the similarity judgments must be the same as the correlation between Hamming distance and distances derived from the similarity judgments.

$$\begin{aligned}\text{HammingSim}_N(\mathbf{s}_i, \mathbf{s}_j) &= N - \text{Hamming}(\mathbf{s}_i, \mathbf{s}_j) \\ &= N - \sum_{k=1}^{N} |s_{ik} - s_{jk}| \\ &= N - \|\mathbf{s}_i - \mathbf{s}_j\|\end{aligned}$$

Note that this measure ensures that, if $\mathbf{s}_{mk} = \mathbf{s}_{nk}$ for all k, the N-Hamming similarity is at its maximum of N. This implies that, if $m = n$, $\text{Hamming}(\mathbf{s}_m, \mathbf{s}_n) = \max_{i,j} \text{Hamming}(\mathbf{s}_i, \mathbf{s}_j) = N$, and if $\mathbf{s}_{mk} \neq \mathbf{s}_{nk}$ for all k, then $\text{Hamming}(\mathbf{s}_i, \mathbf{s}_j) = 0 = \min_{i,j} \text{Hamming}(\mathbf{s}_i, \mathbf{s}_j)$. This seems reasonable.

Such a measure may be problematic unmodified, however, since as I note in Section 2.3, similarity judgments seem to be sensitive to different aspects of verbs' meanings to different degrees. To remedy this, I might include strictly positive weights that represent the importance of each feature in that particular task. To retain the above properties, the N-Hamming similarity would need to be generalized in the following way.[46]

$$\begin{aligned}\text{WeightedHammingSim}_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_j) &= \sum_{k=1}^{N} w_k - w_k|\mathbf{s}_{ik} - \mathbf{s}_{jk}| \\ &= \|\mathbf{w}\|_1 - \|\mathbf{w}(\mathbf{s}_i - \mathbf{s}_j)\|_1\end{aligned}$$

[46] $\text{WeightedHammingSim}_{\mathbf{w}}$ is known more commonly as weighted Minkowski (1-norm) distance. And like weighted Minkowski (1-norm), $\text{WeightedHammingSim}_{\mathbf{w}}$ is equivalent to HammingSim_N—unweighted Minkowski (1-norm)—iff $\mathbf{w} = \mathbf{1}_N$, where $\mathbf{1}_N$ is a vector of 1s of length N.

Under the assumption that $\mathbf{w}$ is strictly positive, this second function—and *a forteriori*, the first—defines a positive-definite kernel function.[47]. As such, I henceforth abbreviate it to

$$K_{\text{Hamm}_{\mathbf{w}}}(i,j) = \text{WeightedHammingSim}_{\mathbf{w}}(\mathbf{s}_i, \mathbf{s}_j)$$

In fact, this kernel is equivalent to the one employed in Shepard and Arabie's (1979) ADCLUS model but for one major difference: where ADCLUS only counts matches between features that are both 1,[48] $K_{\text{Hamm}_{\mathbf{w}}}$ counts matches between features that are both 1 and 0. (I return to why I use $K_{\text{Hamm}_{\mathbf{w}}}$ instead of something like the ADCLUS kernel shortly.)

This comparison makes clear the relationship between the current question and more general questions in mathematical cognitive science regarding generalization—e.g. Shepard's (1987) exponential law of generalization. The relationship defined by $K_{\text{Hamm}_{\mathbf{w}}}$ is linear in the the distributional features encoded in $\mathbf{S}$. This raises a possible worry, in that many generalization and discrimination phenomena are best fit by models that involve exponential decay (see Tenenbaum and Griffiths 2001 for discussion). And indeed, returning to Figure 2.18, it seems this nonlinearity may be appearing here in the relationship between frame distance and dissimilarity ratings; it is easier to predict which words are similar if they are nearby in frame space than if they are far away, and further, the drop-off in predictability appears

[47]More generally, WeightedHammingSim defines a positive-definite kernel if $\|\mathbf{w}\|_1 > 0$ and a positive-semidefinite kernel if $\|\mathbf{w}\|_1 \geq 0$.

[48]In this way, the ADCLUS kernel is like a feature-weighted version of Tversky's (1977) Contrast Model, though there are some other differences in that the Contrast Model is a valid positive (semi)definite kernel only under specific combinations of weights.

to be logarithmic. This in turn suggests that there is a need to scale distributional distances nonlinearly—perhaps, exponentially. The problem is that, unlike the sorts of cases that Shepard and others consider, the current feature space is discrete, meaning that most current models of these nonlinearities are inappropriate.

To remedy this, I consider Kondor and Lafferty's (2002) diffusion kernels on graphs. As Kondor and Lafferty note, diffusion kernels "can be regarded as the discretisation of the familiar Gaussian kernel of Euclidean space" and that $K_{\text{Hamm}_{\mathbf{w}}}$ (in its equiweighted variant) underperforms diffusion kernels in categorical prediction tasks. This latter suggests that these sorts of kernels may be useful in this case as well, which corresponds to Kondor and Lafferty's diffusion kernel for the hypercube (see footnote 45). This is a special case of the diffusion kernel for arbitrary strings over alphabet $\mathcal{A}$ with number of symbols $|\mathcal{A}|$, where $\text{WeightedHammingSim}_{\mathbf{w}}$. (I abbreviate $\text{WeightedHammingSim}_{\mathbf{w}}$ as $\text{WHS}_{\mathbf{w}}$ below for readability.)

$$K_{\text{Diff}}(i,j) = \left(\frac{1-\exp[-|\mathcal{A}|\beta]}{1+(|\mathcal{A}|-1)\exp[-|\mathcal{A}|\beta]}\right)^{\text{WHS}_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_j)}$$

This gives the elegant characterization for binary alphabets $\mathcal{A} = \{0,1\}$. (See Kondor and Lafferty 2002 for a derivation of both of these kernels.)

$$\begin{aligned} K_{\text{Diff}_{\mathbf{w}}}(i,j) &= \left(\frac{1-\exp[-2\beta]}{1+\exp[-2\beta]}\right)^{\text{WHS}_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_j)} \\ &= (\tanh\beta)^{\text{WHS}_{\mathbf{w}}(\mathbf{s}_i,\mathbf{s}_j)} \end{aligned}$$

In the following, I consider the four natural choices that this discussion delimits: the equiweighted linear model ($K_{\text{Hamm}_\mathbf{1}}$), the weighted linear model $K_{\text{Hamm}_\mathbf{w}}$), the equiweighted diffusion model ($K_{\text{Diff}_\mathbf{1}}$), and the weighted diffusion model ($K_{\text{Diff}_\mathbf{w}}$). For each type of regression, corresponding to each similarity dataset, I fit all four models and then compare their performance using two metrics that are similar to WAIC, the metric used to determine a stopping criterion for the non-negative regression model. In the case of the weighted models, $\mathbf{w}$ is learned. In the case of the diffusion models, a second parameter β (akin to a inverse decay parameter in the general case of nonparametric density estimation) must be set; I also learn this parameter from the data.

I further place exponential priors—equivalent to L1 regularization—on both $\mathbf{w}$ and β for the relevant models. This simultaneously serves as to bias against the more complex models, since their MLE estimates might not be the same as their MAP estimates under this regularization, and it also instantiates a variable selection procedure.

2.4.3 Multinomial logit mixed model

I use a standard multinomial logit mixed model with a softmax link and subject random effects, which account for each participant's implicit bias toward a particular kind of response. (See Appendix A for details on these components.) The model was implemented in python using the `pymc` package and was fit using the Powell optimization implemented in the `scipy optimize` module (Jones et al., 2001), which

attempts to find the the model's Maximum A Posteriori (MAP) estimate given the data. For each of the four similarity models, this optimization was repeated 100 times with random initialization and the MAP estimate selected.

2.4.3.1 Model comparison

In Section 2.2, I use WAIC to perform model comparison for the non-negative projection model. This method is not available to us here because the parameters were not derived via sampling. I still need a way of trading model fit with model complexity, however. I thus fall back to model comparison measures that can be computed using only point estimates. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are two such common measures.[49] Table 2.2 shows these model comparison measures, along with the deviance (the log-likelihood scaled by -2) for each of the four models.

Two generalizations are clear from this table. First, both diffusion kernel models outperform both linear kernel models on both AIC and BIC. This suggests that, as in many other areas of cognition, similarity in distributional features decays exponentially with the distance in feature space. Second, both weighting models outperform their unweighted counterpart. This suggests that certain distributional features are more salient than others—a suggestion that is corroborated by the earlier observation that different judgment tasks may tap different aspects of the semantics. Given these clear cut results, I focus on only the weighted diffusion

[49] These measures are less desirable, since they require assumptions about the posterior—namely, multivariate normality—that may or may not hold. Violation of these assumptions may not be too problematic, however, given the clear separation between the models' performance in the current case.

Kernel	Feature weighting	Deviance	AIC	BIC
Linear	None	26103	26469	27825
Linear	Weighted	26032	26426	27886
Diffusion	None	26069	26437	27800
Diffusion	Weighted	25744	**26140**	**27606**

Table 2.2: Model comparison measures for multinomial logit mixed model. The minimum values for AIC and BIC are bolded.

model for the remainder of this subsection.

2.4.3.2 Feature weights

Figure 2.19 shows the weights **w** learned for the multinomial logit mixed model with weighted diffusion kernel. (The MAP estimate for β was 0.37.) I see that the multinomial logit mixed model utilizes three of the very general features from the non-negative projection model (features 2, 3, and 4), and four of the more specific features (features 5, 6, 7, and 13). The highest weighted feature (feature 2) is the one that I pointed out comes closest to the representationality distinction. Interestingly, the rest of the features—besides feature 13, which corresponds to emotive factivity but has a small weight—are ones that I pointed out as harder to interpret. This is interesting in the sense that it is an apparent justification of Fisher et al.'s critique of beginning with a labelling of verb classes.

2.4.3.3 Frame informativity

Knowing the relative importance of each distributional feature is interesting, but it tells us little about how a learner might go about accessing those features. That is, it does not tell us how much discriminative power each frame has for abduc-

ing the features strongly associated with that frame. To assess this, it is necessary to find some way of measuring frames' informativity relative to the weighting on distributional features presented in the last subsection and the projective relationship between that distributional feature and that frame, as encoded in **P** (Figure 2.12 in Section 2.2). With regard to the projective relationships, two things are important: (i) strength of the relationship and (ii) relative uniqueness of that relationship. The first is required for obvious reasons: for a frame to be important it should be important for an important feature. The second is required because the frame should not simultaneously be important for two different important features: if a frame is similarly important for two similarly important features, a verb's showing up in that frame helps little in being able to tell which of those two features the verb has.

I thus use a measure that can take into account both strength in particular projective mappings and inequality[50] across projective relationships. The Gini coefficient, which measures the unevenness of a distribution, is useful in this case. I compute two Gini coefficients for each frame: unweighted Gini, computed directly from the columns of **P** (the transpose of Figure 2.12); and weighted Gini, computed by applying the weights **w** (Figure 2.19) to the columns of **P** and then performing the same computation on the rows of the resulting matrix. Unweighted Gini, then, provides a measure of discriminative power relative just to the distributional characteristics of the data, and weighted Gini provides a measure of discriminative power relative to the salience of those characteristics as they relate to verbs' meanings

[50]Another common term for inequality in this sense is sparsity. See Hurley and Rickard 2009 for reasons to use Gini, over other common measures, as a measure of sparsity.

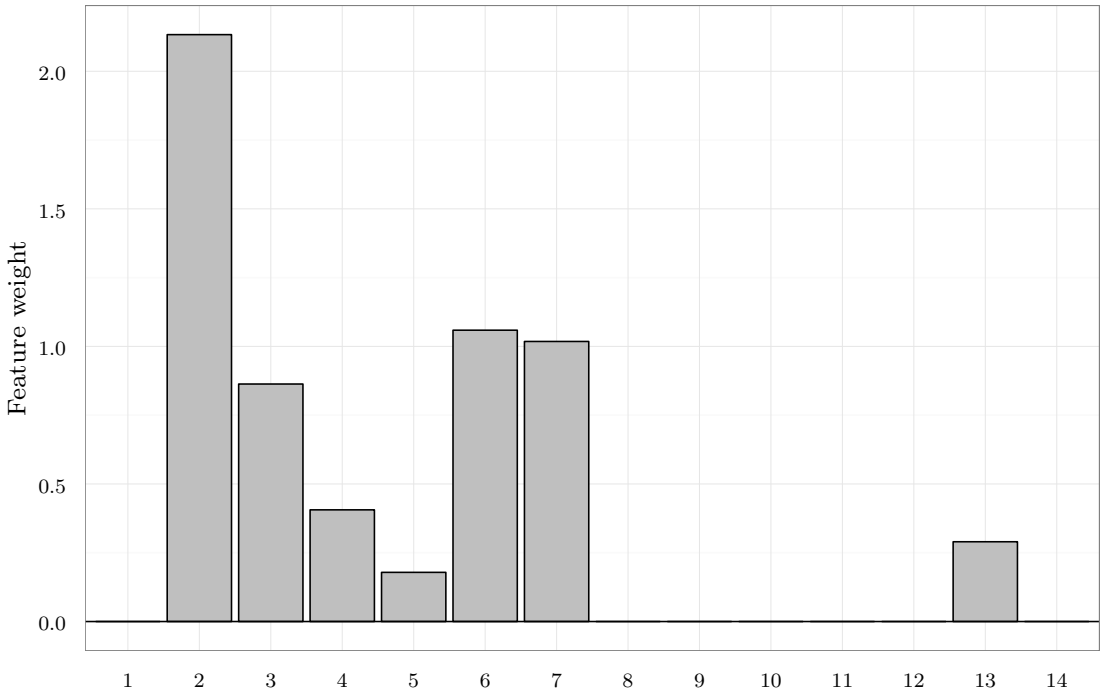


Figure 2.19: Feature weights for multinomial logit mixed model with weighted diffusion kernel. Features correspond to those in Figure 2.11.

(presuming that the similarity judgments are a good proxy for this).

Figure 2.20 shows the results of both computations. The black bars give the unweighted Gini for a frame and the black+grey bars give the weighted Gini. (The grey bar thus gives the difference between weighted Gini and unweighted Gini.) The frames are ordered along the y-axis by their weighted Gini, and so the graph can be interpreted as follows: if one were trying to learn the semantics of a novel word and only got to choose a single syntactic context, they should prefer the syntactic contexts toward the top. For instance, the first six syntactic contexts involve both NP objects—either on the surface or implicitly through passivization—and sentential complement, so knowing whether a verb takes an NP object and a

sentential complement would be quite useful in determining its semantically relevant distributional features.

One thing this graph does not tell us is how useful combinations of syntactic contexts are. Note that this is not derivable from the discriminative power of the syntactic contexts alone, as seen in Figure 2.20, since two different frames may be projected from very different sets of features. This means that seeing a verb in two different frames could either make it easier to discriminate the semantically relevant distributional feature(s) that that verb has—to the extent that the syntactic contexts intersect on features they project strongly from—or make it harder—to the extent that the syntactic contexts don't intersect on features they project strongly from. To assess the discriminative power of a combination, over and above the syntactic contexts that constitute that combination, I then must define a way of assessing the discriminative power of the intersection of features that those syntactic contexts project from. But how does one measure the intersection when the latent features and syntactic contexts are associated via continuous values?

To see how to do this, it is useful to first consider how one might do this if the projective relationships were binary (as the relationship between verbs and features is). If this were the case, the discriminative power of a single syntactic context might be defined as the inverse of the number of feature it is projected from, which is maximized when a syntactic context is projected from only one feature. Intersection for two syntactic contexts could be defined as the latent features both project from—bitwise AND applied to the columns corresponding to the two syntactic contexts—and the discriminative power of the combination could be computed

from the resulting vector in the same way as it is for single frames. Then, a combination is useful over and above its constituent syntactic contexts, if the discriminative power is increased relative to both frames.

In the continuous case, bitwise AND is not available, but its natural generalization, the Hadamard (pointwise) product, is. The Hadamard product is intuitively correct for what I aim to do. If two different syntactic contexts project strongly from some feature, the product of that projection will be large; if only one syntactic context in a combination projects strongly from some feature and the other, weakly, then the product will be middling; and if neither syntactic context in a combination strongly projects, the product will be small. This means that, (i) to the extent that two syntactic contexts agree on only a subset of each of their projections and (ii) to the extent that the disagreements are large, the combination adds discriminative power. To analyze the additional discriminative power that a combination of syntactic contexts adds, I can then compute the Hadamard product of the constituent contexts projection relationships.

Our aim is to ascertain how predictable this value is from the discriminative power of the syntactic contexts and how much is due to their interaction. This can be done by constructing a regression to relate the discriminative power of the constituent contexts to that of the combination. I then residualize the combination's value by that regression's prediction to get a measure of the combination's gain in discriminative power. Because the value of discriminative power is bounded by 0 and 1, I need a regression that assumes a dependent variable on a bounded interval. Beta regression does the trick here. I regress the combination discriminative power

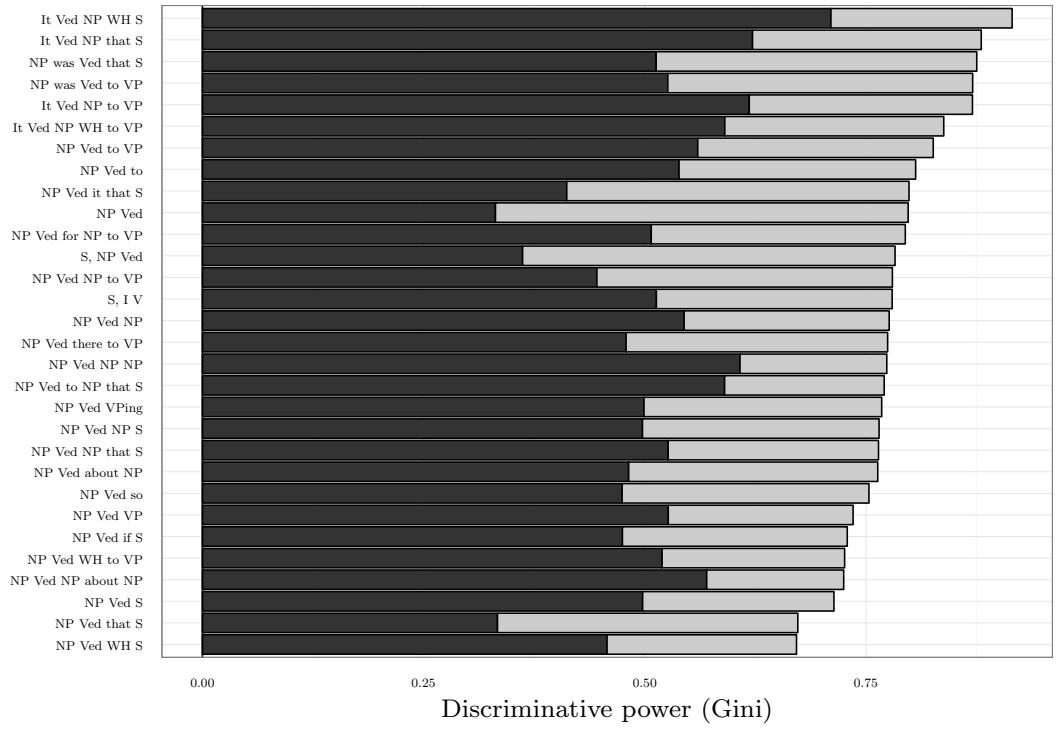


Figure 2.20: Unweighted (black) and weighted (black+grey) Gini computed using feature weights from multinomial logit mixed model with weighted diffusion kernel (Figure 2.19) and projection principles inferred from non-negative projection model (Figure 2.12).

(weighted Gini), computed as above, on both of its constituent's discriminative powers (after a logistic transformation) and their interaction. I then residualize the combination discriminative power by the regression predictions. Figure 2.21 shows these residualized values.

I see a few interesting patterns in these data. First, on the whole most frames enter into at least some useful interactions, suggesting that all frames have at least some extra discriminative power in combination with others. Second, some frames show more gains in discriminative power in combination with nearly all frames, while others are very selective about the frames they interact with. This tends to

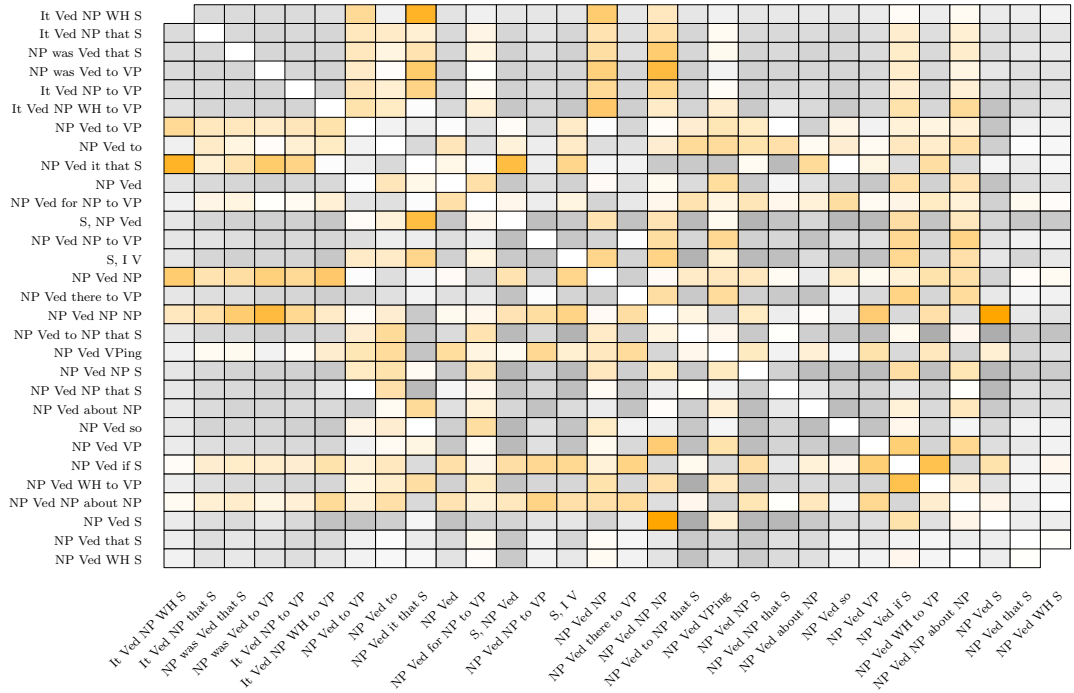


Figure 2.21: Weighted frame combination Gini residualized by beta regression on weighted Gini of each frame in combination (full bars in Figure 2.20). Grey represents a positive residual and orange, a negative.

pattern with initial discriminative power. For instance, both *NP Ved that S* and *NP Ved WH S* show only non-negative increases in discriminative power in combination with other frames, though they also have lower discriminative power to begin with. And frames with higher discriminative power to begin with do not tend to interact with fewer frames. For instance, *NP Ved to VP* tends not to show much if any increase in interaction. (Though interestingly, it does interact positively with both *NP Ved that S* and *NP Ved WH S*.)

To some extent this correlation between initial discriminative power and interactiveness is unsurprising, since if a frame is already highly discriminative, there is less seeing a verb in another frame could do to reduce uncertainty about the features of that verb. On the other hand, this was the point of looking at the residualized scores, and indeed, some high informativity frames show large numbers of frames they interact with. For instance, the top four frames look very much like the bottom two but for their negative interaction with various low interactivity frames.

2.4.4 Ordinal logit mixed model

I now turn to the ordinal logit mixed model analysis of the ordinal scale data. I use a standard ordinal logit mixed model with strictly positive cutpoints and subject random effects, which account for each participant's implicit bias toward particular parts of the scale. (The subject random effects model is the same as the response model described in Section 2.2; see Appendix A for details on this components.) The model was implemented in python using the `pymc` package and was fit using

the Powell optimization implemented in the `scipy optimization` module, which attempts to find the model's MAP estimate given the data. For each of the four similarity models, this optimization was repeated 100 times with random initialization and the minimum deviance model selected.

2.4.4.1 Model comparison

As for the multinomial logit mixed model reported above, I use AIC and BIC for the initial model comparison. Table 2.3 shows the model comparison measures for each of the four models.[51] Similar generalizations hold as in the last section, though the results are less clear cut. Both the diffusion kernel and weighting improve the fit in comparison to the unweighted linear model. This improvement is not additive, however, as can be seen in the fact that the weighted diffusion model only does slightly better than either the weighted linear model or the unweighted diffusion model in terms of AIC, and the unweighted diffusion model does far better than either weighting model in terms of BIC. This contrasts with the finding from last section that the weighted diffusion model bested both the weighted linear model and the unweighted diffusion model by a wide margin on both measures.

The reason for this difference could rest on the nature of the task. Likert similarity scale tasks require participants to map whatever representation of similarity they have for a given domain into a discrete scale. Regardless of how this mapping is done, if similarities decay exponentially with distance and objects are fairly uni-

[51]Note that the values in Table 2.3 should not be compared to those in Table 2.2 since the datasets are not the same. The reason the numbers are lower in the current table is that fewer total datapoints were collected for the likert scale task.

Kernel	Feature weighting	Deviance	AIC	BIC
Linear	None	9811	10521	12699
Linear	Weighted	9682	10420	12685
Diffusion	None	9714	10420	**12611**
Diffusion	Weighted	9678	**10418**	12689

Table 2.3: Model comparison measures for ordinal logit mixed model. The minimum values for AIC and BIC are bolded.

formly distributed through a space, one would expect most verb-verb pairings to yield low ratings on the scale. And this is indeed what I find.

But there is a catch. Because participants tend to use the scale differently, one needs to take into account differences among participants mappings in the analysis. I did this using an ordinal logit model with participant random effects. This means that, in fitting the model, the inference algorithm has to simultaneously decide whether the many low scores seen in the data arose as a consequence of participants mappings, which could map larger intervals of the latent similarity measure to low points on the scale than high points, or as a consequence of the similarities themselves. This would yield a result wherein the linear model looks better than it should because the exponential decay in similarity is being explained in the mappings from similarity to likert scale as opposed to the similarities themselves. That is, the linear model has a way of mimicking the diffusion model by pushing the explanation into the response model.

This does not appear to be the case, however. If the linear model were mimicking the diffusion model, one would predict the lower ratings to have exponentially larger interval sizes than the smaller, but in fact, their size is comparable. Across participants, the interval size for a 1 rating on the likert scale in both the unweighted

diffusion (median=0.56) and weighted diffusion models (median=0.58) is about the same as—or even a bit smaller than—the interval sizes in the unweighted linear (median=0.24) and weighted linear models (median=0.36). And this pattern holds for the remainder of the likert scale points as well (see Figure A.3 in Appendix A).

In fact, rather than the linear models mimicking the diffusion models, it appears that the diffusion models are mimicking the linear models. One of the main differences between the linear and diffusion models is in the distribution of similarities they produce: the diffusion models tend to yield more uneven distributions of similarities, with many low similarities and few high similarities. In contrast the linear models tend to spread similarities more evenly. In degenerate cases, however, the diffusion models can act like the linear models if their inverse decay parameter is relatively large or if the distances that get exponentiated are very small (below 1). This appears to be what is happening here.

In the multinomial logit mixed models, both diffusion models showed much more uneven similarities (Gini=0.30 [unweighted diffusion], 0.46 [weighted diffusion]) than their linear counterparts (Gini=0.11 [unweighted linear], 0.17 [weighted linear]). In contrast, in the current models, both diffusion models show similar uneveness in their similarities (Gini=0.01 [unweighted diffusion], 0.14 [weighted diffusion]) than their linear counterparts (Gini=0.11 [unweighted linear], 0.15 [weighted linear]). This appears to be driven by a large inverse decay value in the unweighted diffusion model ($\beta = 2.81$), where distances cannot be below 1, and small distances in the unweighted diffusion model, which arise as a consequence of small feature weights (max=0.02, mean=0.003). Further, the feature weights with non-negligible

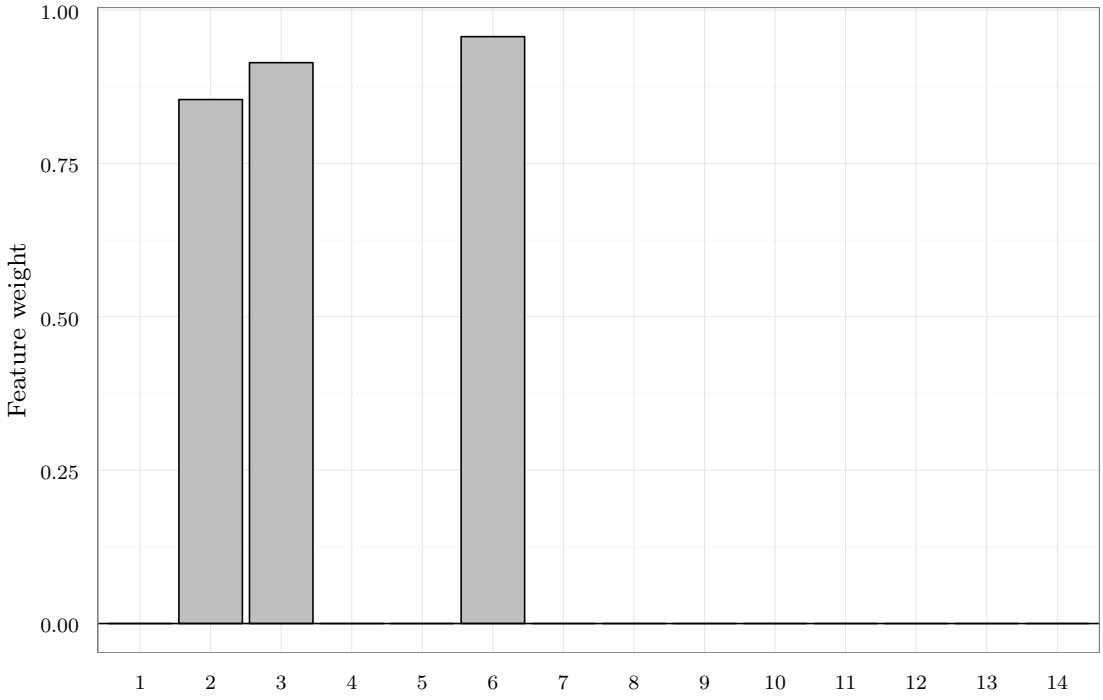


Figure 2.22: Feature weights for ordinal logit mixed model with weighted linear kernel.

measure in the diffusion model match exactly those in the linear model, suggesting that these two models are doing essentially the same thing. For the remainder of this section, then, I analyze the weighted linear model.

2.4.4.2 Feature weights

Figure 2.22 shows the weights **w** learned for the ordinal logit mixed model with weighted diffusion kernel. The features with non-negligible weights in the ordinal logit mixed model are a subset of those that have non-negligible weights for the multinomial mixed model. This again suggests that the generalized semantic discrimination task and the ordinal scale task pick up on similar things aspects of

the meaning. It also suggests that, whatever each is picking up on, the multinomial logit model picks up on more distributionally relevant aspects of it.

2.4.4.3 Frame informativity

As with the multinomial logit model, I measure frame informativity in terms of that frames unweighted and weighted Gini index relative to **w** (Figure 2.22) and **P** (Figure 2.12 in Section 2.2). Figure 2.23 shows the results of these computations. I again focus on weighted Gini and change in rank between unweighted and weighted Gini.

On the whole, the most discriminative frames have less discriminative power than in the multinomial logit mixed model, though the worst have no less. This may be due to the fact that the feature weights are much more even in this case than the last. With regard to specific frames, here again, I find many of the NP object frames have high discriminative power. (This may also be why the expletive object frame *NP Ved it that S* is so discriminative.) Interestingly, the *NP V S* and *S, I V* frames show up higher than in the multinomial logit mixed model, though this may have to do with the fact that all frames are closer in discriminative power overall, and thus small changes can change rank.

Figure 2.24 shows data for the ordinal logit mixed model analogous to that found in Figure 2.24 for the multinomial logit mixed model. In this case, there is a much different overall pattern of results, where overall the increases in discriminative power are low, and they are spread out across frames. This contrasts with the

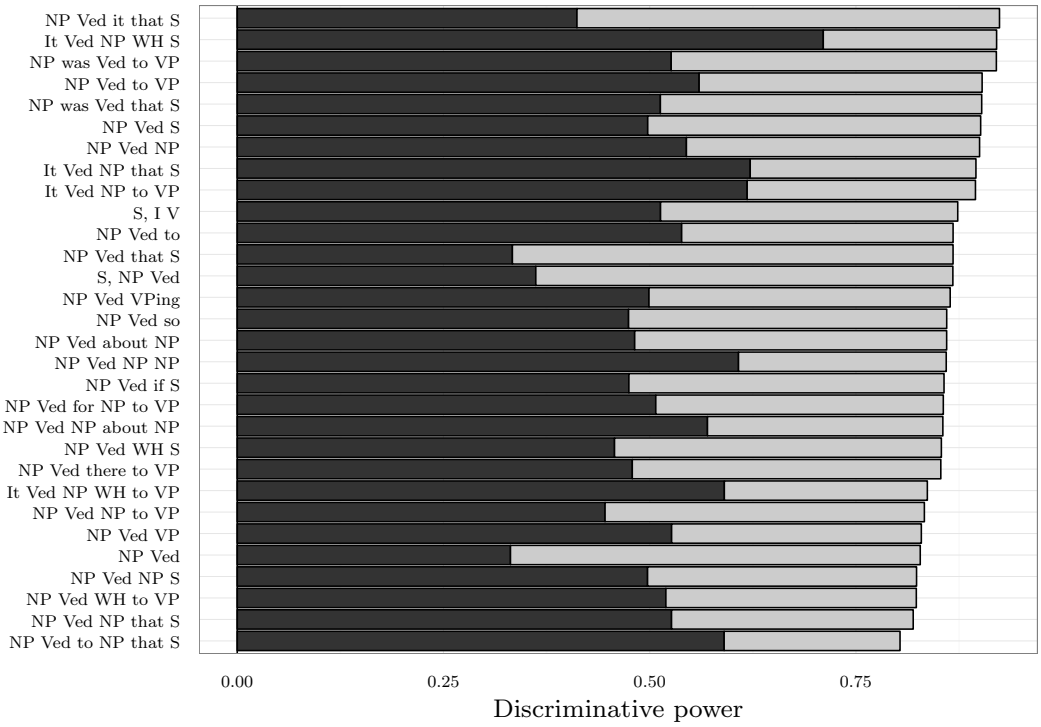


Figure 2.23: Unweighted (black) and weighted (black+grey) Gini computed using feature weights from ordinal logit mixed model with weighted linear kernel (Figure 2.22) and projection principles inferred from non-negative projection model (Figure 2.12).

previous situation, where certain frames had, in a certain sense, maxed out on their discriminative power in such a way that combining them with others would not result in better discrimination. This situation mirrors that seen in Figure 2.23 in that no frames really outperform any others to a great extent.

2.4.5 Discussion

In this section, I explored various ways of quantifying the relationship between the acceptability judgments presented in Section 2.2 and the semantic similarity judgements presented in Section 2.3. First, following Fisher et al., I assessed the

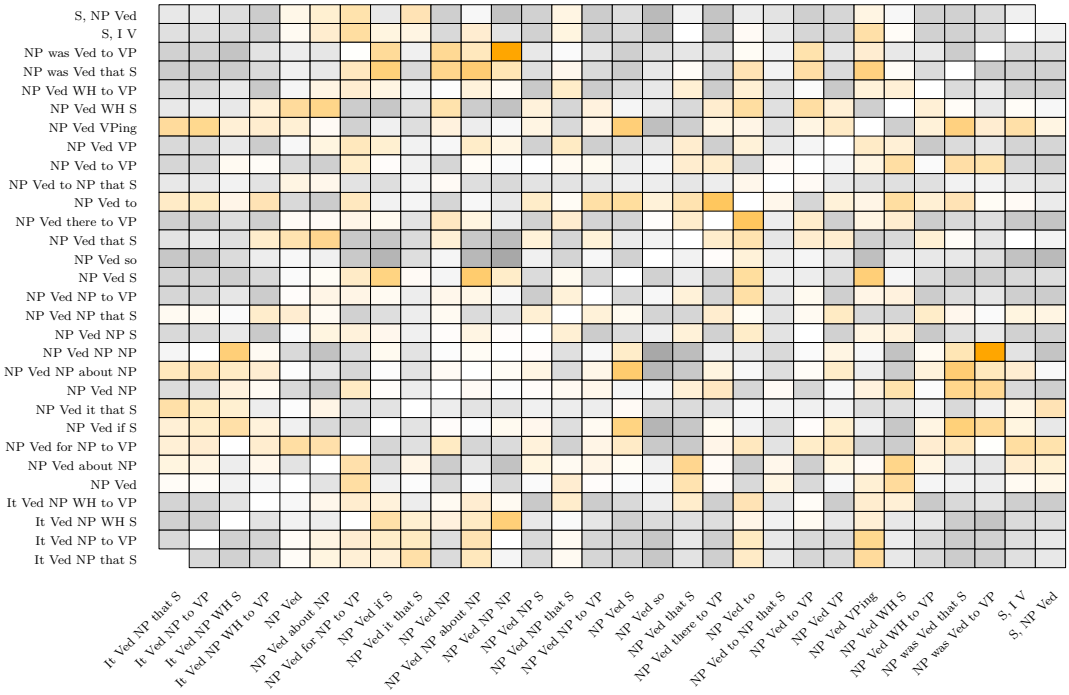
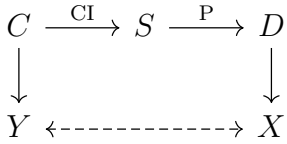
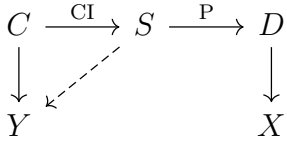


Figure 2.24: Weighted frame combination Gini residualized by beta regression on weighted Gini of each frame in combination (full bars in Figure 2.20). Grey represents a positive residual and orange, a negative.

overall correlation between the acceptability judgments and the semantic similarity judgments.



Second, in order to discover more fine-grained relationships, I map from the regularities **S**, extracted from the acceptability judgments, to the semantic similarity judgments **Y**.



I began with the basic correlational analyses. These analyses were carried out by defining a distance measure between verbs based on **X** (the "raw" data) as well as one on **S**. In both cases, I found significant correlations with distances derived from both of the similarity datasets **Y**. In the case of both the distance on **X** and the distances on **S**, these correlations were of roughly the same size.

I then moved into more sophisticated analyses that allow us to delve into the relationship between the distributional regularities extracted using the non-negative projection model and the similarity judgments directly. I presented two such analyses. The first assessed how best to model participants' perception of semantic similarity relative to the distributional features and the salience of those features. For the generalized semantic discrimination task, a model that assumes

exponential decay in similarity (the diffusion kernel) outperforms one that assumes linear decay. For the ordinal scale task, the linear decay model wins out. I suggested that this may be related to the way participants make semantic discrimination judgments v. direct semantic similarity judgments.

The second analysis used the model fits that resulted from the first analysis to assess how useful different syntactic contexts could plausibly be for a learner in discriminating the semantically relevant distinctions. I found that two of the general features (features 2 and 3) and one specific feature (feature 6) showed up as important in both analyses. These were also the only features that were assigned non-negligible weights in the ordinal logit mixed model, while in contrast the multinomial logit mixed model utlized these features plus another general one (feature 4) and three other specific ones (features 5, 7, and 13). This may suggest that the generalized semantic discrimination responses are derived more directly from $\mathbf{C}$ than the ordinal responses.

2.5 General discussion

In this chapter, I presented experiments aimed at quantitatively assessing how much information about propositional attitude verbs' meaning lies in their competence distribution. I showed that the measure of syntactic distribution and the measure of semantics extracted from these experiments are significantly correlated. In analyzing the data from these experiments I developed a computational model, based on the linguist's notion of projection, which extracts features from the compe-

tences distribution proxy collected in this experiment. I then analyzed these features both qualitatively and quantitatively to assess their relationship to previous proposals about the relationship betwen syntax and semantics in the attitude verb domain, finding that in large part those previous proposals are corroborated. In the next chapter, I turn to the question of whether one finds the same amount of semantic information present in performance distribution as well.

Chapter 3: A computational model of syntactic bootstrapping

In Chapter 2, I show that syntactic distribution, as proxied by acceptability judgments, goes quite far in predicting the fine-grained properties of a word's semantics, as proxied by semantic similarity judgments. As discussed in the last chapter, acceptability judgments are one of the most direct measures available of what I have been calling the *competence distribution* of a word. This direct measure, however, is not necessarily representative of the sort of data learners have access to in verb-learning; rather, they have access to what I have been calling *performance distributions*—e.g. cooccurrence counts between verbs and subcategorization frames. With this in mind, the first goal of this chapter is to show that similar levels of semantic information lie in the kinds of *performance distributions* that learners plausibly have access to at the same time as presenting a model that takes advantage of this information.

To assess the semantic information that lies in these performance distributions, I adapt the nonnegative projection model proposed in Chapter 2, which was applied to acceptability judgment data, to verb-subcategorization frame counts from a corpus. This can be done quite straightforwardly due to the modular nature of this model: at a high level, the response model used for the acceptability judgments

need merely be replaced with a count model. The resulting model can be conceived of on two levels: on the first level, it generates competence distributions (the sorts of syntactic structures a verb is good with) by projecting semantic features into a distribution space using the projection rules. Then, from these competence distributions, it generates performance distributions (the sorts of syntactic structures a verb occurs with) by sampling verb-frame pairs according to the competence distributions.

I propose that this model abstractly characterizes the syntactic bootstrapping process, and on analogy with the nonnegative projection model proposed in the last section, I call this model the nonnegative syntactic bootstrapping model. This model is, in many ways, the midway point between two types of models of learning semantics from syntactic distribution. On the one hand are category-based models of semantic representation, which tend to be founded in probabilistic approaches to semantic representation like Latent Dirichlet Allocation (LDA Blei et al., 2003); and on the other hand are vector space models of semantic representation, which tend to be founded broadly in matrix factorization techniques such as Singular Value Decomposition (SVD) like Latent Semantic Analysis (LSA Deerwester et al., 1990) and, more recently, neural embedding models like skip-gram with negative sampling model (Mikolov et al., 2013). I refer to these latter sorts of models as feature-based models for reasons that become clear.[1]

The chapter begins in Section 3.1 with a review of previous models, both prob-

[1]This is, by necessity, a rough characterization, since at a higher level, the goal of both general approaches is to factor the observed data into some representations involving latent objects (categories or features), and thus they have similarities. But the sorts of representations each traffics in is distinct enough to warrant discussion.

abilistic and vector space, that learn word-meaning representations from syntactic distribution. The upshot of this discussion is to motivate the need for a midway point between the representational freedom of vector space models, which can result in hard-to-interpret features—cf. the discussion of PCA in Chapter 2—and the sorts of representations that common category-based probabilistic models produce, which tend to be easier to interpret but which I show have other undesirable properties from the point of view of semantic representation.

The particular undesirable property I focus on is the fact that these representations are what I refer to as *globally normalized* and thus don't allow us to naturally represent words that may have multiple features simultaneously. In Section 3.1.1, I discuss this normalization property at length, showing that it can be converted into a sort of representation I refer to as *locally normalized*, arguing that this representation does not fall prey to the interpretability issues inherent in the vector space models. I then suggest that the conversion process itself is of interest because it shows a deep connection between the category view of word-meaning and a feature view of word-meaning. The upshot of this suggestion is that normalized representations can be fruitfully thought of as topological in nature and the unnormalized representations can be thought of as logical in nature.

In Section 3.2, I present the nonnegative syntactic bootstrapping model, which incorporates the desirable features of both the vector space models and the probabilistic models discussed in the previous two sections. This model utilizes the same nonnegative projection model employed in the previous chapter, but it replaces the response model used there with a model of counts. I show that the features that this

model induces correlate with the similarity judgments presented in the last chapter to approximately the same extent as the acceptability judgments in the last chapter. The upshot of this—the main result of this chapter—is that syntactic distributions in corpora—performance distributions—carry a significant amount of fine-grained information about attitude verb syntax.

In Section 3.4, I conclude by showing how to build an incremental learning algorithm for this model.

3.1 Computational models and syntactic distribution

Computational models of word-meaning induction from syntactic distribution tend to fall into two main classes: category-based models that attempt to induce verb classes from syntactic distributions culled from corpus counts (LDA and related models) and vector space models that attempt to induce spatial representations from these corpus counts (LSA-based models and neural embedding models). I begin with a review of work within both domains, focusing in particular on the sorts of representations they traffic in.

3.1.1 Category models

3.1.1.1 Prior approaches

Category-based models, which might involve either hard or soft-clustering approaches, tend to find their inspiration in Levin's (1993) now classic handbook as well as various resources that derive at least partially from it. One of the more

straightforward approaches within the category-based frameworks is exemplified by Schulte im Walde and Brew 2002; Schulte im Walde 2003, 2006.[2] Schulte im Walde's approach was to estimate the parameter of a multinomial distribution over frames for each verb (with additive smoothing $\lambda = 0.5$), then apply k-means clustering with various distance metrics (Manhattan distance, Euclidean distance, KL divergence, information radius, skew divergence, and cosine distance) and various values of k. Choice of optimal k was based on optimizing two indices of agreement with a gold-standard categorization. Thus, under this approach verbs are viewed as falling into distinct hard clusters.

Other approaches make similar assumptions but utilize slightly different methods. For instance, Stevenson and Merlo (1999) and Merlo and Stevenson (2001) were interested in whether grammatical features that constitute subcategorization frames—e.g. active v. passive—could be used to discover a fixed set of three predefined classes. To do this, they utilized both an unsupervised method, hierarchical clustering converted to a flat hard clustering like that produced by k-means, as well as a supervised method, decisions trees trained on the gold standard classes (see also White et al. 2014 for a similar approach focused on the fine-grained semantics of propositional attitude verbs). Stevenson and Joanis (2003) extend this methodology with a semi-supervised approach to feature selection. Simalar to Stevenson and Merlo's hierarchical clustering approach, Schulte im Walde (2000) used a relative entropy-based hierarchical clustering and well as a generative latent class model

[2]Each of these papers present slightly different analyses while keeping the general approach roughly the same. Schulte im Walde 2006 presents perhaps the most comprehensive set of analyses; this discussion is based on that paper.

proposed in Rooth 1995.

More recent approaches eschew hard clusters for soft clusters. That is, instead of representing a word as belonging to a single category, the word's representation is fundamentally associated with a discrete distribution over some number of categories (possibly an unbounded number). One common method for performing this soft clustering is to employ probabilistic methods originally designed for document classification. Perhaps the most popular current framework for carrying this out is that of Latent Dirichlet Allocation (LDA; Blei et al., 2003) and related models that employ priors that allow the number of latent categories to vary, such as the Hierarchical Dirichlet Process (HDP; Teh et al., 2006). Because it is useful for the sake of grounding discussion, I give the generative story for the Latent Dirichlet Allocation (LDA) model (Blei et al., 2003), since it forms the core of many lexical representation models. For convenience, I convert the common document speak into word speak, where V is the number of words to represent—i.e. soft cluster—and F is the number of possible contextual features (cooccurring words, cooccurring structures, etc.).

In this model, verbs are associated with discrete distributions over categories—sometimes referred to as *topics* in reference to this model's document classification origins—parameterized by a multinomial/categorical parameter θ_i of length K, where $(K + 1)$ is the total number of categories. These categories are in turn associated with a multinomial/categorical distribution over, e.g., possible contextual features (words, strings, syntactic structures, etc.) of a word. The LDA generative story is quite simple.

1: **for** word category k in $1 : K$ **do**

2: Choose a distribution over contextual features $\phi_k \sim \text{Dirichlet}(\beta \mathbf{1}_K)$

3: **end for**

4: **for** word i in $1 : V$ **do**

5: Choose a distribution over word categories $\theta_i \sim \text{Dirichlet}(\alpha \mathbf{1}_K)$

6: **for** occurrence j in $1 : n_i$ **do**

7: Choose a category $s_{ij} \sim \text{Categorical}(\theta_i)$

8: Choose a contextual feature $x_{ij} \sim \text{Categorical}(\phi_{s_{ij}})$

9: **end for**

10: **end for**

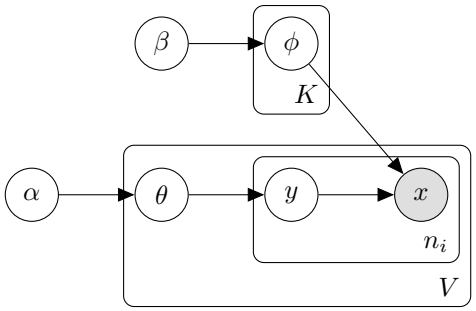


Figure 3.1: Plate diagram for Latent Dirichlet Allocation (LDA)

Griffiths et al. (2007) propose using this sort of approach for word-word cooccurrences, where a word's semantic representation is discovered by associating each of K categories with a distribution over words, and each word is associated with a distribution over the K categories.[3] Building on work in the computational literature on selectional preferences (cf. Resnik, 1996; Ritter and Etzioni, 2010; Ó Séaghdha,

[3]Griffiths et al. also briefly consider some alternative model structures, but for current purposes, I focus on the simpler model, since the more complex models represent meanings themselves in the same way.

2010), Ó Séaghdha and Korhonen (2014) propose various models that take into account both the words that a particular word occurs with as well as the syntactic relationships those cooccurring words have to the word of interest.[4]

Vlachos et al. (2008, 2009) construct a similar model with a nonparametric prior that infers verb representations from verb-subcategorization frame cooccurrences. In the same vein as Vlachos et al., Alishahi and Stevenson (2008) present a model that learns subcategorization frames as distributions over syntactic features, constructions as distributions over subcategorization frames, and verbs as distributions over constructions (see also Barak et al. 2013, 2014b,a, which utilize the same model and focus on very coarse-grained attitude verb classes). The impetus for this extra construction level in Alishahi and Stevenson's case is to capture some notion of semantic structure underlying multiple frames (Jackendoff, 1990; Goldberg, 1995).

A similar idea drives Lewis and Steedman's (2013) approach. Lewis and Steedman use the output of a Combinatorial Categorial Grammar parser to induce types using a k-nearest neighbors style clustering analysis. They then apply vanilla LDA to verb-type cooccurrence counts, representing verbs as distributions over categories that are in turn associated with distributions over types.

3.1.1.2 Representational assumptions

The representational assumption that these approaches all have in common is what I refer to as the global normalization property: the representation of a

[4]See also Gormley et al. 2012, who augment the vanilla topic model of selectional restrictions to discover a low dimensional representation of the properties that categories are constructed from.

word's semantics must sum to 1.[5] The hard clustering models trivially assume this, since they assume that the verb representation is given by a single category. This can be thought of as a one-hot vector representation, with zeros in all positions besides that corresponding to a particular category. This view seems problematic, at least for encoding propositional attitude verb semantics for reasons discussed extensively in Chapters 1 and 2: propositional attitude verbs' meanings are multi-faceted. *Want* and *hope* share a property (preferentiality), but so do *think* and *hope* (representationality).

One way to get around this problem is, of course, to multiply the number of categories so that, e.g., one category encodes nonfactive belief verbs (*think*), one encodes factive belief verbs (*know*), one encodes nonfactive nonbelief verbs (*want*), and another encodes factive nonbelief verbs (*love*). But this clearly misses a generalization that, e.g., *think* (a representational nonfactive) and *know* (a representational factive) cross-classify to the exclusion of *want* (a preferential nonfactive) and *love* (a preferential factive). Indeed, in the limit, this leaves every verb in a separate category.[6]

The soft clustering models—e.g. LDA—provide more representational freedom, which at least partially fixes this problem by assuming that a particular verb may be associated with multiple categories. The caveat to this freedom is that,

[5]The hierarchical clustering approach is something of an exception to this, at least in principle. In practice, however, the hierarchy is cut at a certain threshold to create mutually exclusive categories, resulting in the same sort of representation produced by hard clustering models like k-means. Thus, it falls under this generalization.

[6]Of course, the whole question of this dissertation is essentially where this limit is, and it seems unlikely that, as a matter of syntactic distribution, a learner would ever have enough evidence to place each verb in its own category.

on any particular occurrence of the verb, that occurrence falls into a particular category. This is problematic in a sense related to the problematic aspect of the hard clustering approach: many verbs seem to retain the same semantic features on each occurrence, and so modeling them as though each occurrence is the product of a different category seems incorrect. For instance, on each occurrence of *hope*, the speaker is presumably committed to both the desire entailment and the belief entailment.

This is not to say that one or the other component might not be foregrounded on a particular use; indeed, it seems quite likely that on any one use, either the belief or preference component is foregrounded. But regardless of this presumably pragmatic foregrounding effect, both entailments remain. For instance, the belief component seems foreground in (1b) since B presumably means to convey that she believes its possible that John went to the store—or perhaps merely that it's possible that he did.

(1) **A:** Did John go to the store?

B: I hope he did.

But even with this foregrounding, the desire component does not go away: B in (1) is still committed to all wanting it to be the case that John went to the store. For instance, she cannot follow up with (2) and not contradict herself.

(2) **B:** ...but I don't want him to have.

Thus, this is very different from a case of polysemy—where the word might shift

between (possibly systematically) related meanings, and thus be amenable to a univocal semantic description, but nonetheless has only one particular meaning, with particular entailments, on a particular use. For instance, (3) might have an aperture reading or an obstruction reading.

(3) The ghost went through the door.

And perhaps the event being described could be described by both expressions. But no one would argue that the aperture reading—with its entailment that the ghost move through the door frame—also has the obstruction entailment—that the ghost also passed through the door filling that doorframe. This can be seen in the fact that (3) can truly describe a scene in which the ghost passes through an open door frame. This suggests that *hope* is not really the same sort of beast as the standard polysemy examples. It really does require a description incorporating both components of its meaning.

The question then becomes: why should both components not determine cooccurring words, structures, etc. on each occurrence? One common tack is to ignore how the model itself views the representation—as a probability of a category on a particular occurrence—and try to treat the distribution associated with each word as something like encoding weights between particular words and its features. This gives rise to a representation that, unlike the one-hot hard clustering representation, can be viewed as encoding a word's multiple features simultaneously.

I think this view is reasonable but not in its barest form. Recall the example of *think*, *want*, *hope*. Suppose we encode these three verbs using two categories/features—

representationality and preferentiality—which we can represent as a two-element vector. The vector associated with *think* might be (0, 1), whereas the vector for *want* might be (1, 0). But what about the vector for *hope*? Ideally, it would encode an equal relationship between representationality and preferentiality. But since each verb's representation is, at base, a distribution over categories, and one category is sampled on each occurrence of a verb, the verb's category/feature relationships must sum to one. This means that the only way to encode an equal relationship between *hope* and both representationality and preferentiality is to associate *hope* with the vector (0.5, 0.5). In turn, this means that a verb's relationship to its categories/features cannot be interpreted on its own but only relative to the entire representation.

The reason why this problem arises has to do with the fact that the simplex representation encodes not only the strength of the relationship between a word and a feature but also information about how likely that relationship is to manifest itself. But these two properties seem separable. The relationship between a word and a component of its meaning seems more essential than the likelihood of observing that component's effects. How, then, does one derive the "true" relationship between a word i and feature k from θ_i? One possibility is to loosen the restriction that the representation be globally normalized.

3.1.2 Vector space models

In the last section, I noted that many category-based models, both hard clustering and soft clustering models, constrain their representations of verb semantics to be globally normalized. I then suggested that these globally normalized representations were not intuitive as featural representations, which is seemingly what one needs to represent the multi-faceted nature of propositional attitude verbs. In this section, I consider a semantic representation that does not fall prey to the problems inherent to globally normalized representations.

There are a broad range of vector space models to semantics. The general approach to semantic representation in these models is to view words as points in some space of observable features. For instance, in the case relevant to this dissertation, one might conceive of verbs as lying in some space whose dimensions correspond to subcategorization frames and whose values correspond to some relationship between the verbs and the frame—e.g. cooccurrence count, term frequency-inverse document frequency (tf-idf), pointwise mutual information (PMI), etc. Any sort of observable feature is possible as a dimension. (The same is of course true of the category-based models, as discussed in the last section.)

One touchstone case of these vector space models is that arising from the Latent Semantic Analysis/Indexing (LSA/LSI) literature (Deerwester et al., 1990). As is true of LDA (discussed briefly above), LSA's original application was to document representation, but by replacing documents with words, LSA provides a natural way of representing words (Landauer and Dumais, 1997). Assuming the spatial view of

words, the idea behind LSA—an idea which recurs throughout this section—is that learning a semantic representation consists in factoring a matrix encoding the spatial position of each word within the observable dimensions into two distinct matrices: one that represents the relationship between a verb and various latent features or components of that verb (the score matrix) and another that represents the relationship between the latent features and the observed features (the loading matrix). The particular way that LSA does this is known as Singular Value Decomposition (SVD). This method underlies the algorithms that compute Principal Component Analysis (PCA), as used in the last chapter.

As noted in Chapter 2, one nice aspect of this sort of model is that it naturally encodes the notion of projection from semantic features to syntactic distribution: semantics-to-syntax projection is projection from one vector space (the semantics) to another (the syntax).[7] One problem with this sort of approach is that, left unconstrained as in the case of LSA, it results in feature values that are hard to interpret even if the features themselves correspond to a clear semantic class.

As noted in Chapter 2, one remedy for this—the one employed in that chapter—is to enforce non-negativity and sparsity and to employ unit- or binary-valued features. The first two of these are also employed in Murphy et al. 2012; Fyshe et al. 2014, 2015, though the last is not.

This general spatial meaning approach has been extended rapidly in recent years. It has been put to particularly common use within the deep learning litera-

[7] In fact, it is a misnomer to call this projection in the second case, since the mapping in question is not an endomorphism.

ture. One model popular for inducing word representation is that given in Mikolov et al. (2013), who present their skip-gram with negative sampling model (cf. Rumelhart et al. 1986 for an early example of this sort of model; see also Bengio et al. 2006; Collobert and Weston 2008). Roughly, this model attempts to learn representations in some real-valued space that can be used to predict words on either side of that word. Like standard LSA models, these neural word embedding models appear to be performing a sort of implicit matrix factorization, which Levy and Goldberg (2014b) argue is based on a variant of a matrix containing the point-wise mutual information between objects (e.g. verbs) and observable features (e.g. subcategorization frames).

Levy and Goldberg (2014a) note that these representations, often called neural word-embeddings "are considered opaque, in the sense that it is hard to assign meanings to the dimensions of the induced representation." (p. 303) and they give a similar method that uses the same notion of predicting context, but instead of predicting string-adjacent words, their model predicts adjacent adjacent words in a dependency parse. They note that, whereas the string adjacent version produces "broad topical similarities...the dependency-based contexts yield more functional similarities of a cohyponym nature."

It is useful, however, to separate interpretability of a dimension itself and interpetability of a value along that dimension. In the last chapter's section on PCA, I showed several cases in which the component of the meaning that a particular dimensions was sensitive to was at least somewhat clear. For instance, it appeared that PCA discovered meaning features like representationality, preferentiality, factivity,

and communicativity. But the relative position of words along these dimensions was uninterpretable.

This arises because the feature values can fall anywhere in $(-\infty, \infty)$ on the reals. This problematic aspect was noted in the last chapter in the discussion of PCA. For instance, what does it mean to be negative on a feature? Should that be interpreted as not having the feature? And what does it mean for a particular verb to a value, e.g., four times greater than another verb, but in the same direction? It's always possible to pass individual real values through a normalizing (or "squashing") function—for example, the standard logistic function—but unless that value itself is passed through that function as a part of whatever objective was used to fit the model, it is unclear what the interpretation of that operation should be.[8]

For this reason, this sort of value is very different from the globally normalized representations discussed in the last section, in that the value of one feature does not affect the values of the others. In this sense, these representations are unnormalized, since nor constraint applies to their sum. But for this same reason, this sort of representation is more powerful, since a verb is free to be related to a particular feature to an extent independent of its relationship to other featues.[9] The problem is that it leaves the feature values themselves uninterpretable without further constraint. Interpretability of both the feature itself and its value are important, though. If a learner's job is to discover features of a word's meaning such as representationality and preferentiality, which themselves are seemingly symbolic,

[8]Note that it is the dot product of two vectors and not their values that is passed through a logistic in a model Mikolov et al.'s model.

[9]I do not mean to say that this representation is necessarily more expressive, however.

then it is a prerequisite to have a value that itself can be interpreted symbolically.

This is one of the main benefits of the category-based approach; the categories that verbs are associated with can be interpreted symbolically yet the representation itself encodes uncertainty about which aspect of the meaning is relevant on any one occurrence of the word. The problem, noted in Chapter 2, is that some words seem to have multiple components of their meaning on each occurrence. This is intimately related to the normalization property discussed briefly in the last section. In the next section, I present a model that can incorporates the benefits of the category-based model while not requiring such constrained representations.

3.2 The model

The current model is the same as the model in Chapter 2 in terms of the priors on $\mathbf{S}$ and $\mathbf{P}$. In that chapter, I utilize a prior over the distributional regularities $\mathbf{S}$ with a finite number of features K, fitting the model with various K and then perform model comparison.[10]

$$\pi_k \mid \alpha, \beta \sim \text{Beta}(\alpha, \beta)$$

$$s_{ik} \mid \pi_k \sim \text{Bernoulli}(\pi_k)$$

I further retain the previous chapter's exponential prior on $\mathbf{P}$. This prior

[10]I have also implemented a nonparametric version using the Indian Buffet Process prior (Griffiths and Ghahramani, 2006, 2011), but as in Chapter 2, only the parametric version is tested here.

induces sparsity in the same way as an L1 regularizer.[11]

$$b_{kj} \mid \lambda \sim \text{Exponential}(\lambda)$$

Where the current model diverges from the previous chapter is in the form of **D** (the representation of the competence distribution) and **X** (the acceptability judgment data). In that chapter, **D** is a deterministic product of **S** and **P**. This **D** is then "passed through" an ordinal response model, whereby larger d_{ij} get mapped to distributions with more mass over higher scale points (denoting higher acceptability of a verb with a particular syntactic context). Because **P** (the projection rules) is non-negative real-valued, **D** is also non-negative real-valued. Indeed, this would also be possible in the current setup, since counts are also non-negative real-valued and thus, **X** (now, the counts) could be thought of as a product of adding some sort of noise to **D**.

There are two problems with this approach. First, it produces a representation of the competence distribution that is too sensitive to the relative counts found in the corpus unless some sort of frequency damping model is employed (cf. Goldwater et al. 2011). This seems right for representing performance distributions, but not competence distributions, which I argued in the last section, should involve something more like a unit interval representation.

[11]Indeed, an Exponential(λ) prior is equivalent to an L1 regularizer with weight λ. In the more general case of real-valued parameters—i.e. not non-negative parameters as in this case—the equivalent of L1 regularization in a Bayesian context is a Laplace prior. But if $x \sim \text{Laplace}(\lambda)$, then $|x| \sim \text{Exponential}(\lambda)$.

This suggests that a model that conditions its learning of competence distributions on performance (count) distributions must have some way of representing the competence distributions that factors out the relative counts. But if **D** were represented as in the last chapter—as non-negative real-valued—**D** would retain some residue of the empirical distributions and thus would not be a good candidate for a competence distribution representation.

To remedy this, I pass **SP** through a Beta distribution to produce a distribution over **D** with support only on the unit interval. As I will see, this boundedness forces the model to explain the count aspects of the empirical distributions in some other way.

$$d_{ij} \mid \mathbf{S}, \mathbf{P} \sim \text{Beta}([\mathbf{SP}]_{ij}, 1)$$

In the current model, the way that the model is forced to explain the count data is via an auxiliary (nuisance) variable **g**. This variable can be thought of as encoding, in g_i, the relative prevalence of verb i. The count of verb i with syntactic context j, x_{ij}, is then modeled as a draw from a Poisson distribution with parameter $g_i d_{ij}$.

$$g_i \mid \gamma, \delta \sim \text{Gamma}(\gamma, \delta)$$

$$x_{ij} \mid g_i, d_{ij} \sim \text{Poisson}(g_i d_{ij})$$

One metaphor for thinking about this component is that each word i is associated with a "potential energy" g_i and each syntactic context j associated with that word "gets" up to that much energy. The proportion that each context actually uses of the energy it could use is encoded in d_{ij}. If one knew both g_i and d_{ij} for a word i and syntactic context j, they would then expect to see those two together on average $g_i d_{ij}$ times.

The energy in this metaphor corresponds quite directly to a word's overall count. Since poisson distributions can be compounded, $g_i \sum_j d_{ij}$ gives the expected number of occurrences of word i. This means that, though the overall distribution of proportions d_{ij} matter for the counts, they themselves are not constrained by having to explain counts in the same way as, e.g., a globally normalized simplex representation (multinomial parameter), since each can take on a value in $(0, 1)$.

As I show shortly, this layout has two benefits beyond the theoretical one just discussed: first, $\mathbf{g}$ can be completely collapsed in inference due to a conditional conjugacy; second, because of this collapse, x_{ij} is distributed as a three parameter version of the two parameter Negative Binomial (the poisson-gamma mixture distribution), which is known to describe empirical (count) distributions in language well (Church and Gale, 1995).

Figure 3.2 gives the plate diagram for the above model. The generative story is given by:

1: **for** feature k in $1 : K$ **do**

2: Choose a feature probability $\pi_k \sim \text{Beta}(\alpha, \beta)$

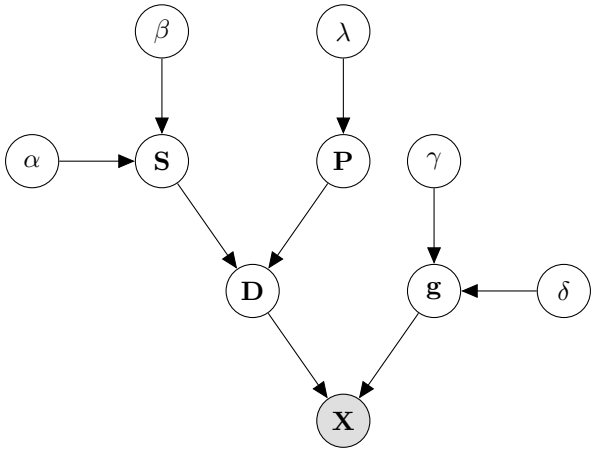


Figure 3.2: Plate diagram for non-negative projection model.

3: **for** verb i in $1 : V$ **do**

4: Choose a feature value $s_{ik} \sim \text{Bernoulli}(\pi_k)$

5: **end for**

6: **for** syntactic context j in $1 : F$ **do**

7: Choose a projection strength $b_{kj} \sim \text{Exponential}(\lambda)$

8: **end for**

9: **end for**

10: **for** verb i in $1 : V$ **do**

11: Choose a verb prevalence $g_i \sim \text{Gamma}(\gamma, \delta)$

12: **for** syntactic context j in $1 : F$ **do**

13: Choose a competence distribution strength $d_{ij} \sim \text{Beta}([\mathbf{SP}]_{ij}, 1)$

14: Choose a cooccurrence count $x_{ij} \sim \text{Poisson}(g_i d_{ij})$

15: **end for**

16: **end for**

3.2.1 Batch learner

In this section, I focus on two pieces of the inference equations for a learner that computes the posterior over latent verb features $\mathbf{S}$, projection rules $\mathbf{P}$, and competence distributions $\mathbf{D}$ given counts $\mathbf{X}$—that is, a learner that computes $\mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D} \mid \mathbf{X}; \Psi)$, where $\Psi = \{\alpha, \beta, \lambda, \gamma, \delta\}$, by "reversing" the above generative story. The particular pieces I focus on are the likelihood $\mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)$ and the posterior on $\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})$. I provide the full equations necessary for constructing a Gibbs sampler as well as the gradients necessary for conducting Maximum A Posteriori (MAP) estimation over the continuous matrices in Appendix B.

Note that $\mathbf{g}$ does not occur in the above probability functions. Since the goal is a model of word-learning, $\mathbf{g}$ is not particularly interesting in that it encodes count information that one needs to control for but which is only partially related to the representations of interest $\mathbf{S}$ and $\mathbf{P}$. In the next section, I show how it is possible to compute the above posterior without explicitly finding a probability distribution over $\mathbf{g}$. This is possible due to a useful conditional conjugacy.

3.2.1.1 A useful conditional conjugacy

It is well-known that the gamma distribution is a conjugate prior of the poisson distribution. A less well-known, but related, conjugacy arises from the product distribution constructed from a gamma random variable and another random variable independent of the gamma. This product distribution is conditionally conjugate to the poisson with respect to the gamma distribution. This can be taken advantage

of to analytically integrate out $\mathbf{g}$.

$$\begin{aligned}
\mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D} \mid \mathbf{X}; \Psi) &\propto \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D} \mid \mathbf{X}; \Psi) \\
&= \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g} \;\; \mathbb{P}(\mathbf{g}, \mathbf{S}, \mathbf{P}, \mathbf{D}, \mathbf{X}; \Psi) \\
&= \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi) \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g} \;\; \mathbb{P}(\mathbf{X} \mid \mathbf{g}, \mathbf{D}) \mathbb{P}(\mathbf{g}; \gamma, \delta)
\end{aligned}$$

Let us focus for the moment on the integral, since this is where the conditional conjugacy becomes important. First, note that, if d is an arbitrary positive random variable with parameters α, β, $g \sim \text{Gamma}(\gamma, \delta)$, $x \sim \text{Poisson}(gd)$, and $g \perp\!\!\!\perp d$ then

$$\begin{aligned}
\mathbb{P}(d \mid x; a, b, \gamma, \delta) &\propto \mathbb{P}(d, x; a, b, \gamma, \delta) \\
&= \int_{\mathbb{R}_+} \mathrm{d}g \;\; \mathbb{P}(g, d, x; a, b, \gamma, \delta) \\
&= \mathbb{P}(d; a, b) \int_{\mathbb{R}_+} \mathrm{d}g \;\; \mathbb{P}(x \mid g, x) \mathbb{P}(g; \gamma, \delta)
\end{aligned}$$

This yields the same form as our model, whereby the posterior is the product of the prior over d and the integral over g. Since this will come up later, note also that this integral is equivalent to the likelihood of d, $\mathbb{P}(x \mid d; \gamma, \delta)$. This integral is quite easily solved analytically using the standard conjugacy technique.

$$\begin{aligned}
\int_{\mathbb{R}_+} \mathrm{d}g \;\; \mathbb{P}(x \mid g, d)\mathbb{P}(g; \gamma, \delta) &= \int_{\mathbb{R}_+} \mathrm{d}g \;\; \frac{(gd)^x}{x!} \exp\left[-(gd)\right] \frac{\delta^\gamma}{\Gamma(\gamma)} g^{\gamma-1} \exp\left[-\delta g\right] \\
&= \frac{d^x \delta^\gamma}{x!\Gamma(\gamma)} \int_{\mathbb{R}_+} \mathrm{d}g \;\; g^{x+\gamma-1} \exp\left[-g(d+\delta)\right] \\
&= \frac{d^x \delta^\gamma}{x!\Gamma(\gamma)} \frac{\Gamma(x+\gamma)}{(d+\delta)^{x+\gamma}} \\
&= \frac{\Gamma(x+\gamma)}{\Gamma(x+1)\Gamma(\gamma)} \frac{d^x \delta^\gamma}{(d+\delta)^{x+\gamma}}
\end{aligned}$$

Note that the normalizing constant in the above equation is just that of a negative binomial distribution. Indeed, the previous equation just the PMF of a negative binomial distribution with the parameterization NegativeBinomial(γ, $\frac{d}{d+\delta}$). This distribution gives the distribution over number of successes before γ failures in a series of bernoulli trials with probability $\frac{d}{d+\delta}$ of success. One can think of a success here as seeing a particular syntactic context.

For current purposes, let us assume that $d \in (0, 1)$, since in the model it is beta distributed and thus has support only on $(0, 1)$. Figure 3.3 shows the PDF for this likelihood at various values of x and δ and γ set to 1—that is, in which

What can be seen here is that, for small δ, one needs a quite large x to conclude that d is near 1. But as δ gets larger, the probability of d being near one rises quickly for any count. This has to do with the fact that, as δ goes to 0, the quantity $\frac{d}{d+\delta}$, the probability of success in the negative binomial, is close to 1 regardless of the value of d, and so we expect many successes—i.e. a large x—before a failure. On the other hand, as δ gets larger, $\frac{d}{d+\delta}$ gets smaller and smaller regardless of the value of d, since it is bounded at 1, and so the larger the count, the more the likelihood will prefer d near one.

This is interesting, but the existence of this conjugacy alone does not imply anything about the more complex situation present in the above model, in which there are multiple d_{ij} for any one verb. All x_{ij} are dependent on a single unobserved g_i, which one would like to integrate out. But this dependency could cause problems that do not arise in the simple case. It turns out, however, that due to certain properties of the Poisson PMF, we can take advantage of the conjugacy above.

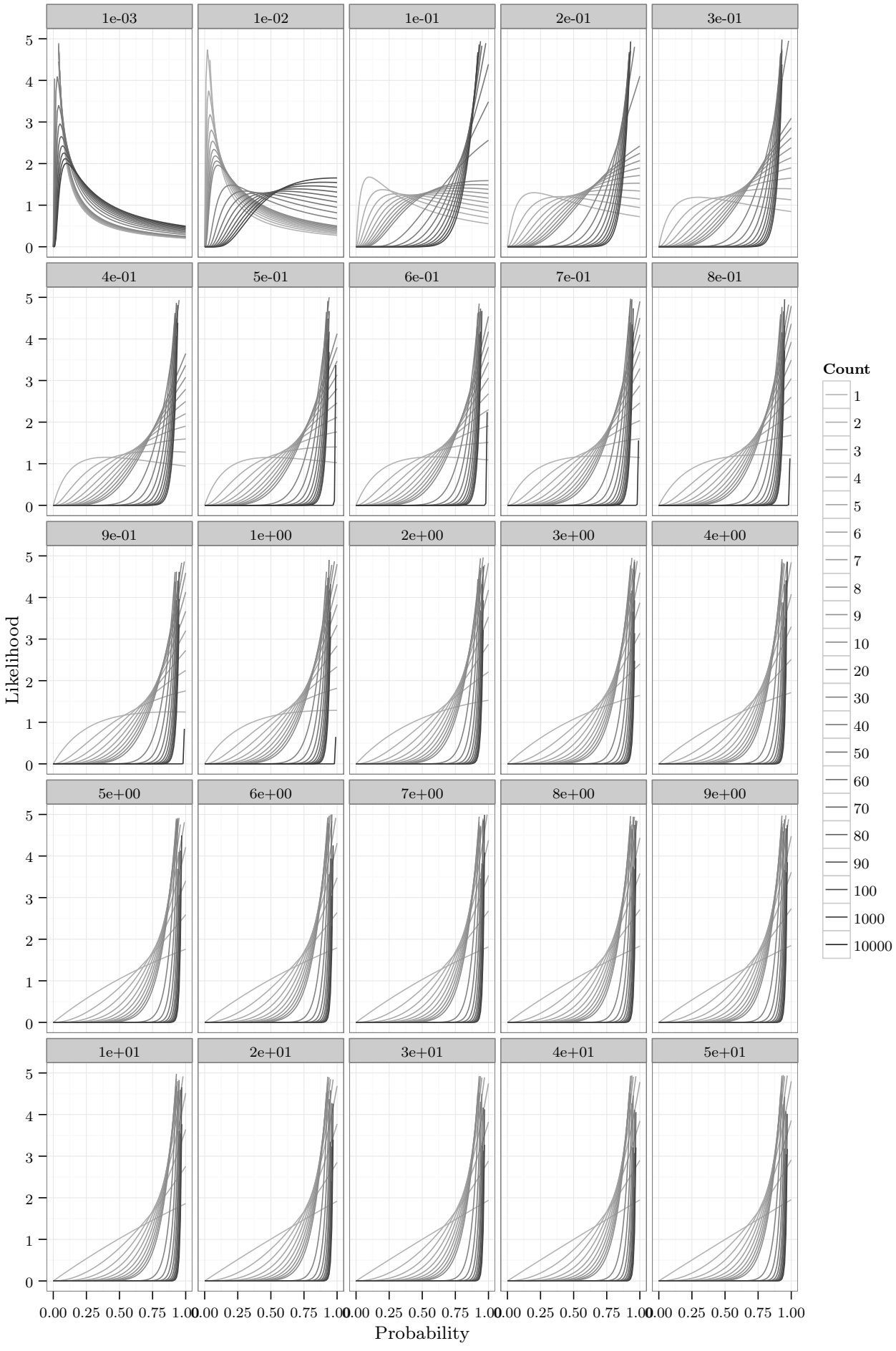


Figure 3.3: Likelihood of d given x and δ. The value of δ is given as the title of each facet.

$$\int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \mathbb{P}(\mathbf{X} \mid \mathbf{g}, \mathbf{D})\mathbb{P}(\mathbf{g}; \gamma, \delta) = \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \prod_{i=1}^{V} \mathbb{P}(g_i; \gamma, \delta) \prod_{j=1}^{F} \mathbb{P}(x_{ij} \mid g_i, d_{ij})$$

$$= \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \prod_{i=1}^{V} \frac{\delta^\gamma}{\Gamma(\gamma)} g_i^{\gamma-1} \exp\left[-\delta g_i\right] \prod_{j=1}^{F} \frac{(g_i, d_{ij})^{x_{ij}}}{x_{ij}!} \exp\left[-(g_i, d_{ij})\right]$$

$$= \left[\prod_{i=1}^{V} \frac{\delta^\gamma}{\Gamma(\gamma)} \prod_{j=1}^{F} \frac{d_{ij}^{x_{ij}}}{x_{ij}!}\right] \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \prod_{i=1}^{V} g_i^{\gamma-1} \exp[-g_i\delta] \prod_{j=1}^{F} g_i^{x_{ij}} \exp[-g_i d_{ij}]$$

$$= \frac{\delta^{\gamma V}}{\Gamma(\gamma)^V} \left[\prod_{i=1}^{V} \prod_{j=1}^{F} \frac{d_{ij}^{x_{ij}}}{x_{ij}!}\right] \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \prod_{i=1}^{V} g_i^{\gamma-1} \exp[-g_i\delta] g_i^{\sum_{j=1}^{F} x_{ij}} \exp\left[-g_i \sum_{j=1}^{F} d_{ij}\right]$$

$$= \frac{\delta^{\gamma V}}{\Gamma(\gamma)^V} \left[\prod_{i=1}^{V} \prod_{j=1}^{F} \frac{d_{ij}^{x_{ij}}}{x_{ij}!}\right] \int_{\mathbb{R}_+^V} \mathrm{d}\mathbf{g}\ \ \prod_{i=1}^{V} g_i^{\gamma-1+\sum_{j=1}^{F} x_{ij}} \exp\left[-g_i \left(\delta + \sum_{j=1}^{F} d_{ij}\right)\right]$$

$$= \frac{\delta^{\gamma V}}{\Gamma(\gamma)^V} \prod_{i=1}^{V} \left[\prod_{j=1}^{F} \frac{d_{ij}^{x_{ij}}}{x_{ij}!}\right] \int_{\mathbb{R}_+} \mathrm{d}g_i\ \ g_i^{\gamma-1+\sum_{j=1}^{F} x_{ij}} \exp\left[-g_i \left(\delta + \sum_{j=1}^{F} d_{ij}\right)\right]$$

$$= \frac{\delta^{\gamma V}}{\Gamma(\gamma)^V} \prod_{i=1}^{V} \left[\prod_{j=1}^{F} \frac{d_{ij}^{x_{ij}}}{x_{ij}!}\right] \frac{\Gamma\left(\gamma + \sum_{j=1}^{F} x_{ij}\right)}{\left(\delta + \sum_{j=1}^{F} d_{ij}\right)^{\gamma+\sum_{j=1}^{F} x_{ij}}}$$

This looks a good deal more complicated than the previous, but in the simple case where $\mathbb{P}(d_{mn} \mid \mathbf{X}, \mathbf{D}_{-(mn)}; \gamma, \delta)$ is desired (as in Gibbs sampling), it simplifies somewhat to.

$$\mathbb{P}(d_{mn} \mid \mathbf{x}_m, \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}; \gamma, \delta) \propto \frac{d_{mn}^{x_{mn}}}{\left(\delta + \sum_{j=1}^{F} d_{ij}\right)^{\gamma + \sum_{j=1}^{F} x_{ij}}} \mathbb{P}(d_{mn} \mid \mathbf{S}, \mathbf{P}; \gamma, \delta)$$

where $\mathbb{P}(d_{mn} \mid \mathbf{S}, \mathbf{P}; \gamma, \delta)$ is the prior on d_{mn}.

The major difference between this equation and the simpler one I began the section with is that, instead of being divided by only $(d_{mn} + \delta)^{x_{mn}+\gamma}$, $d_{mn}^{x_{mn}}$ is divided by the sum of all $\mathbf{d}_m$. One interesting thing to note about this is that it penalizes d_{mn} that are too low relative to the size of x_{mn}, but it does nothing to penalize d_{mn} that are high with respect to the relative size of x_{mn}. This is to say that, as far as the likelihood is concerned, it is worse to predict that an object does not have a feature when it does than to predict that it does, when it actually doesn't. This is not to say that there is no pressure to push the probability of unseen object-feature pairs down; just that that pressure is spread out across the matrix—or more specifically, the row corresponding to a particular verb.

How this pressure gets distributed is to some extent controlled by the prior, as I show in the next section. Thus, in a certain sense, the model is only using positive evidence—rewarding high probability instances and not punishing low probability instances. This allows the model to learn a verb's distribution—particularly a low frequency one—by making inductive hypotheses based on higher frequency verbs. This in turn results in high frequency verbs "sucking" low frequency verbs closer to

their distribution.

This has a second benefit relative to the results reported in the last chapter. There, I noted that being close in distribution was predictive of being close semantically (according to the semantic similarity judgments), but that being far away was not. This model implements this in terms of competence distributions, since if need be, it can allow one verb's distribution to assimilate to another's without too much penalty, especially of the assimilating verb is low frequency. In the next section, I show how the beta prior I place on the d_{ij} can further take advantage of this for the purpose of feature induction.

3.2.2 Factor analysis-based smoothing

The prior on $\mathbf{D}$ has the following form.

$$\begin{aligned}
\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P}) &= \prod_{i=1}^{V} \prod_{j=1}^{F} \mathbb{P}(d_{ij} \mid \mathbf{S}, \mathbf{P}) \\
&= \prod_{i=1}^{V} \prod_{j=1}^{F} \frac{\Gamma([\mathbf{SP}]_{ij} + 1)}{\Gamma([\mathbf{SP}]_{ij})\Gamma(1)} d_{ij}^{[\mathbf{SP}]_{ij}-1} (1 - d_{ij})^{1-1} \\
&= \prod_{i=1}^{V} \prod_{j=1}^{F} \frac{\Gamma([\mathbf{SP}]_{ij} + 1)}{\Gamma([\mathbf{SP}]_{ij})} d_{ij}^{[\mathbf{SP}]_{ij}-1} \\
&= \prod_{i=1}^{V} \prod_{j=1}^{F} [\mathbf{SP}]_{ij} d_{ij}^{[\mathbf{SP}]_{ij}-1}
\end{aligned}$$

For a particular d_{mn}, this simplifies to the following.

$$\mathbb{P}(d_{mn} \mid \mathbf{S}, \mathbf{P}) = [\mathbf{SP}]_{ij} d_{ij}^{[\mathbf{SP}]_{ij}-1}$$

When combined with the likelihood term of the posterior $\mathbb{P}(d_{mn} \mid \mathbf{X}, \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}; \gamma, \delta)$, the following is obtained.

$$\mathbb{P}(d_{mn} \mid \mathbf{S}, \mathbf{P}) = \frac{[\mathbf{SP}]_{mn} d_{mn}^{x_{mn}+[\mathbf{SP}]_{ij}-1}}{\left(\delta + \sum_{j=1}^{F} d_{ij}\right)^{\gamma+\sum_{j=1}^{F} x_{ij}}}$$

Thus, as before, this equations penalizes d_{mn} that are too low relative to the size of x_{mn}, but it does nothing to penalize d_{mn} that are high with respect to the relative size of x_{mn}. That is, it is still worse to predict that an object does not have a feature when it does than to predict that it does, when it actually doesn't.

Nonetheless, there is still a pressure coming from the denominator not to overpredict d_{mn} near 1. The addition of the prior modulates this pressure by (i) adding what amounts to a pseudocount (as in standard LDA) and (ii) scaling the

entire likelihood by the size of the projection strength $[\mathbf{SP}]_{mn}$, which in turn is related to verb m's features via the projection rules. Thus, the influence of the features and projection rules is felt in the form of a redistribution of pressure across $\mathbf{d}_m$ not to overpredict particular d_{mn}, based on the particular features a verb has. In the next section, I report on an experiment that deploys this model of syntactic bootstrapping on actual data.

3.3 Experiment

In this section, I report on an experiment that fits the model of syntactic bootstrapping proposed in the last section to subcategorization frame distributions in a corpus—as I have been referring to them, performance distributions. I then show that the distributional regularities extracted from the corpus explain the semantic similarity data from last chapter about as well as features extracted from the acceptability judgment data in that chapter using the same core projection model.

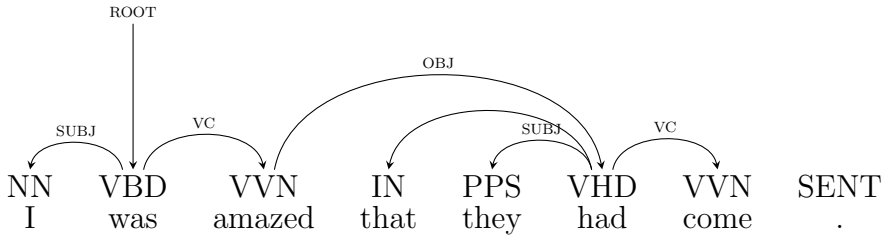
3.3.1 Data

Three subcategorization datasets containing verb-by-subcategorization frame counts were considered. Two of these datasets were previously constructed: an English corpus built by Schulte im Walde using the methods described in Schulte im Walde 2003 for German and Korhonen et al.'s (2006) VALEX lexicon, which is built using Korhonen's (2002) update of Briscoe and Carroll's (1997) set of 163 subcategorization frames—a superset of those in the well-known ANLT and COMLEX

dictionaries (Grishman et al., 1994). Both of these datasets lack some subcategorization information important to attitude verb syntax—e.g. information about the voice of the clause—and so a further subcategorization frame dataset was constructed. All datasets were submitted to the basic correlational analysis described below, but only this latter dataset, described further below, was submitted to the full modeling procedure.

Data constituting this third dataset were extracted from the Parsed uk Web as Corpus (PukWaC) dataset (Baroni et al., 2009). PukWaC is the part-of-speech (POS) and dependency parsed version of ukWaC, which is an approximately two billion word web scrape of the uk domain. To create PukWaC, ukWaC was lemmatized and POS tagged using TreeTagger (Schmid, 1994) and dependency parsed using MaltParser (Nivre et al., 2007).

To extract subcategorization frames associated with particular verbs the following post-processing was conducted. For each item tagged as a verb in a particular sentence, the parents and dependents of that verb were collected. For instance, in 3.3.1, the parent of *amazed* is *was* and the dependents of *amazed* is *had*.



For verb dependents, the tense/aspectual marker—e.g. *past* or *gerund* (*-ing*)—of the dependent was recorded by analyzing the part-of-speech tag. If the dependent was an auxiliary verb (*have* or *be*), the tense of that verb was recorded.

The tense/aspect of the dependents were mapped down into four values: *past* and *present* to TENSED, *gerund* to GERUND, *past participle* to PASTPART, and *bare* to BARE. Modal auxiliaries were mapped to TENSED.

The dependents of this verb were then checked for the presence of a subject (marked by the SUBJ dependency relation) and a complementizer (*that*, *if*, *like*, etc.) or WH word (*who*, *what*, etc.). Subjects were mapped to three values: unambiguously nominative pronouns (*I*, *he*, *she*, etc.) to NOM, unambiguously accusative pronouns to ACC, and all others to CASEUNKNOWN. This was done to get a rough sense of whether the verb occurred in a tensed clause (4a), ECM (4b), or small clause (4c) construction, since the tense information on the verb is sometimes otherwise ambiguous if the verb is in its bare form. (If the dependent verb occurred in its bare form, the dependents of that verb were also checked for the present of an infinitival marker *to*.)

(4) a. Bo thinks that he went to the store.

b. Bo wants him to go to the store.

c. Bo saw him go to the store.

The complementizers were mapped to four values: *that* and *like* (as in *seems like*) to FINITE, *for* to NONFINITE, *if* and *whether* to POLARQ, and any WH word (*what*, *who*, etc.) to *whq*.

In cases with no auxiliaries, the subject of the matrix verb—in the above example *amaze*—was recorded. If the parent of the matrix verb was an auxiliary, as in the case above, the auxiliary chains were followed until a subjects (if any) was

found.[12]. These subjects were mapped into four values: *it* to IT, *there* to THERE, everything else to REFERENTIAL, and no subject to NONE. The idea here is to get an approximation to whether the verb occurred with an expletive subject (and if so, what kind) or not.

Dependents were also checked for whether there were noun phrases or prepositional phrases marked with the relation OBJ. These were marked in three boolean features: main object 1 (true if one or two NP dependents were found), main object 2 (if two NP dependents were found), and prepositional object (if at least one prepositional phrase was found).

The final feature that was extract was whether the matrix verb was passivized. This is important for distinguishing object experiencer verbs from other verbs, which was found ot be important in the last chapter. To assess this, the presence of a form of the auxiliary *be* as the immediate parent along with past participle marking on the matrix verb was recorded. If both were, that verb was recorded as passivized.

For each observation of each verb, these feature values were concatenated to produce a subcategorization frame; 427 such combinations were observed at least once. The number of times a particular verb was found with a particular frame was then recorded to make up the performance distribution matrix $\mathbf{X}$.

[12]If the verb is embedded, it may not have a subject if the subordinate clause it is found in does not have a subject

3.3.1.1 Filtering

For this particular experiment, I focus only on verbs that plausibly occur with some sort of embedding syntax broadly construed. For instance, the sentences in (5) count as embeddings under this broad construal.

(5) a. Bo thinks that Jo went to the store.

b. Bo wants Jo to go to the store.

c. Bo saw Jo go to the store.

d. Bo loves going to the store.

Of course, the criterion for embedderhood cannot be that the verb only occurs in embedded clauses, since many embedders also allow nonembedding structures.

(6) a. Bo thinks about Jo.

b. Bo wants Jo.

c. Bo saw Jo.

d. Bo loves Jo.

The natural criterion would then seem to be that the verb at least sometimes occurs in embedded clauses. The problem here is that, due to noise in the parsing, many verbs that are not actually embedders appear as though they have embedded syntax. Two fairly prevalent instances of this are cases where free relatives are parsed as though they are embedded question clauses (7a) and purposes clauses as though they are infinitival clauses (7b).

(7) a. I'll kiss whoever you tell me to.

b. I drank to celebrate my birthday.

These cases are likely somewhat rarer than true embedding cases relative to the frequency of the verb, however. That is, *kiss* will likely not take free relatives nearly as often as clear NPs and *drink* will likely not take purpose clause nearly as often as it takes an NP object or no object at all. One strategy for filtering out nonembedders, then, is to somehow assess the frequency of a verb-embedded clause pair relative to the frequency of the verb with any complement and an embedded clause of any type (finite, infintival, question, etc.) with any verb. Pointwise Mutual Information (PMI) was used for this purpose.

Prior to calculating PMI, verbs with frequency less than 1000 were filtered. As can be seen in Figure 3.4, this removes all but the top 1500 verbs. From this filtered set, PMI was taken between a categorical variable *verb* and a binary variable *embedded-clause*.[13] This second variable was set to true in the case that a verb on a particular datapoint had a verb dependent and false otherwise. PMI was taken according to the standard formula (Church and Hanks, 1990).

$$\text{PMI}(\text{verb}, \text{embedded-clause}) = \log \frac{\mathbb{P}(\text{verb}, \text{embedded-clause})}{\mathbb{P}(\text{verb})\mathbb{P}(\text{embedded-clause})}$$

Figure 3.5 shows the distribution of the normalized version of the PMI measure: $\text{NPMI}(\text{verb}, \text{embedded-clause}) = \frac{\text{PMI}(\text{verb}, \text{embedded-clause})}{-\log \mathbb{P}(\text{verb}, \text{embedded-clause})}$ Verbs were then fil-

[13] Dunning's (1993) G could also have been used here.

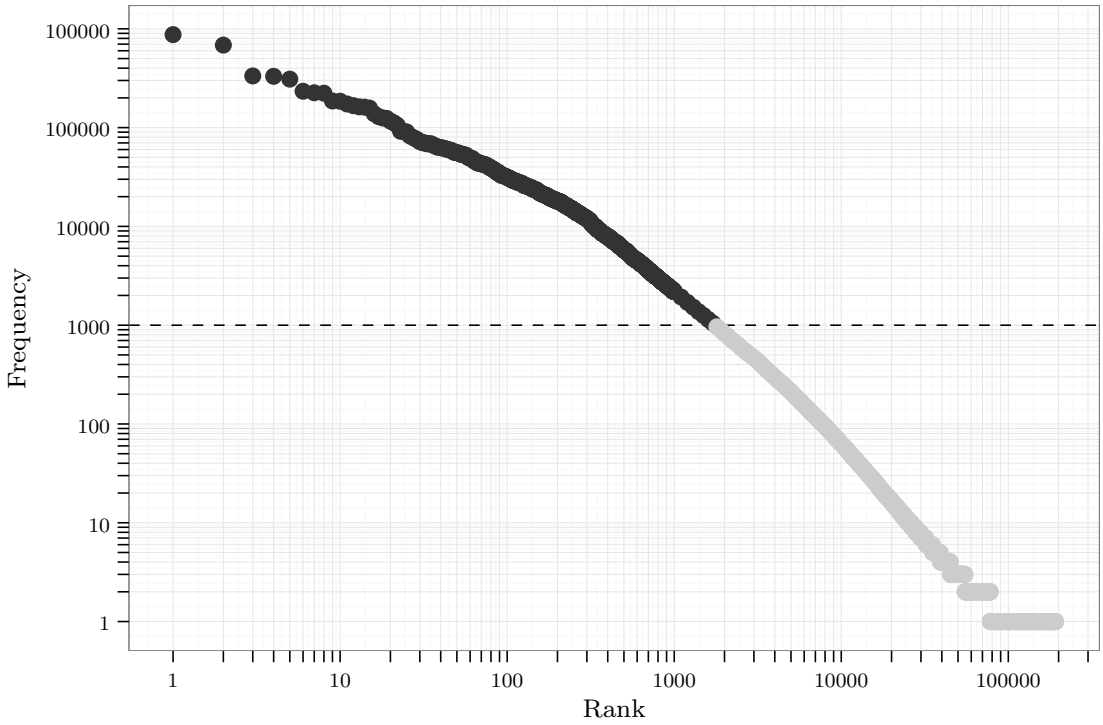


Figure 3.4: Count v. rank of particular verbs. Black points show verbs that were kept.

tered for whether they had PMI greater than 0 or were in verb set from previous chapter. All verbs included satisfy both the frequency and PMI conditions except *worry* from the original set, which fails to satisfy the PMI condition. This yields a total of 232 verbs.

3.3.2 Hybrid sampler/optimizer design

The sampler designed for this experiment implements Gibbs sampling for the posterior on **D**, **P**, and **S**—both parametric and nonparametric (IBP)—and Maximum A Posteriori (MAP) estimators (gradient descent with dynamic step-size to enforce support boundaries) for **D** and **P**. The MAP estimators can be used in

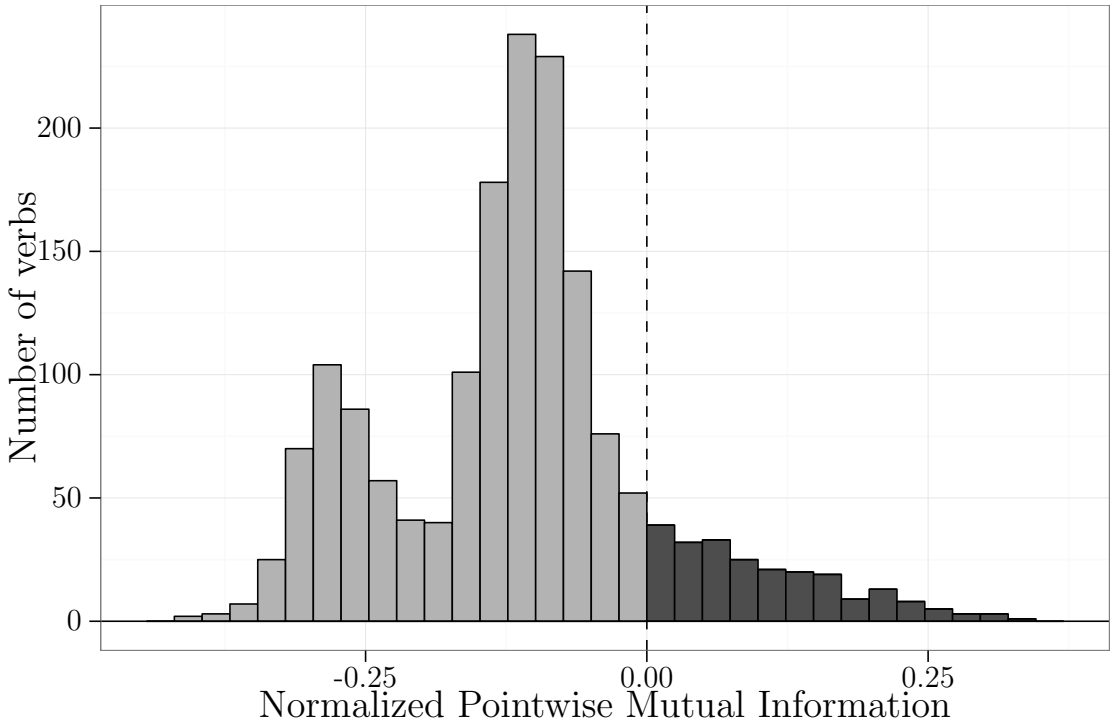


Figure 3.5: Distribution of normalized pointwise mutual information of verb and embedded clause across verbs. Black bars show verbs that were kept at embedders.

place of the Gibbs samplers when the dimensions of $\mathbf{S}$ are fixed—i.e. when $\mathbf{S}$ has a parametric prior.

For the purposes of this chapter, I investigate only the parametric model and use MAP estimators in place of samplers for $\mathbf{D}$ and $\mathbf{P}$, meaning that estimates of the posterior variance will be poor. This was done mostly for convenience, since the inference algorithms converge much more quickly when using the optimizers.

Fitting was separated into three separate stages: (i) a maximum likelihood (MLE) pre-training stage for $\mathbf{D}$; (ii) a MAP pre-training stage for $\mathbf{S}$ and $\mathbf{P}$; and (iii) a hybrid sampler/optimizer training stage for $\mathbf{D}$, $\mathbf{P}$, and $\mathbf{S}$. I describe each stage in detail below.

3.3.2.1 MLE pre-training ($\mathbf{D}$ only)

$\mathbf{D}$ was pre-trained by optimizing the log-likelihood $\log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma)$ (as noted above, δ is set to 1). This results in a MLE estimate, since the implicit prior on d_{mn} is uniform on $(0, 1)$. Optimization was carried out using gradient descent with dynamic step size to enforce the $(0, 1)$ bounds on d_{mn}. Figure 3.3 shows that likelihood of particular d_{mn} given values of x_{mn} and $gamma$. The step size is given by:

$$\text{step-size}(d_{mn}) = r \min\left(d_{mn}, 1 - d_{mn}\right)$$

where r is a learning rate parameter (set to 0.01 for all simulations). The additive updates are given by

$$\text{update}(\mathbf{D}, m, n) = \text{step-size}(d_{mn}) \tanh\left(\frac{\partial}{\partial d_{mn}} \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)\right)$$

where tanh is the hyperbolic tangent. The application of hyperbolic tangent is necessary to ensure that, when the gradient is steep, d_{mn} is not pushed outside of $(0, 1)$. Since the hyperbolic tangent has infimum -1 and supremum 1, and since the step size will never be great than the distance from the current d_{mn} and a bound, updates will never push d_{mn} outside its bound.

Figure 3.7 shows the sort of $\mathbf{D}$ this procedure produces using the PukWaC-derived dataset and Figure 3.6 shows the counts that $\mathbf{D}$ is derived from. Figure 3.7 As can be seen from the similarity of these two graphs, the MLE procedure fits $\mathbf{D}$ tightly. This is expected without any prior information to raise the probability of verb-subcategorization frame pairs with fewer (or even no) occurrences. Another way of thinking about the MLE-derived $\mathbf{D}$ is that it represents a relatively unsmoothed representation of the competence distributions. (I say "relatively" here because there is of course some smoothing coming from γ and δ.)

3.3.2.2 MAP pre-training ($\mathbf{S}$ and $\mathbf{P}$)

Subsequent to the MLE pre-training described above, $\mathbf{S}$ and $\mathbf{P}$ were pre-trained by iteratively optimizing the log-posterior of $\mathbf{S}$, $\log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \log \mathbb{P}(\mathbf{S} \mid \alpha, \beta)$ and $\mathbf{P}$, $\log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \log \mathbb{P}(\mathbf{P} \mid \lambda)$. Note that because $\mathbf{S}$ is discrete, it is not possible to use a method like gradient descent to optimize it. A proxy of this $\mathbf{S}$ of the

Figure 3.6: Log of verb-subcategorization frame counts plus 1. White represents 0 and grey is scaled with the log count.

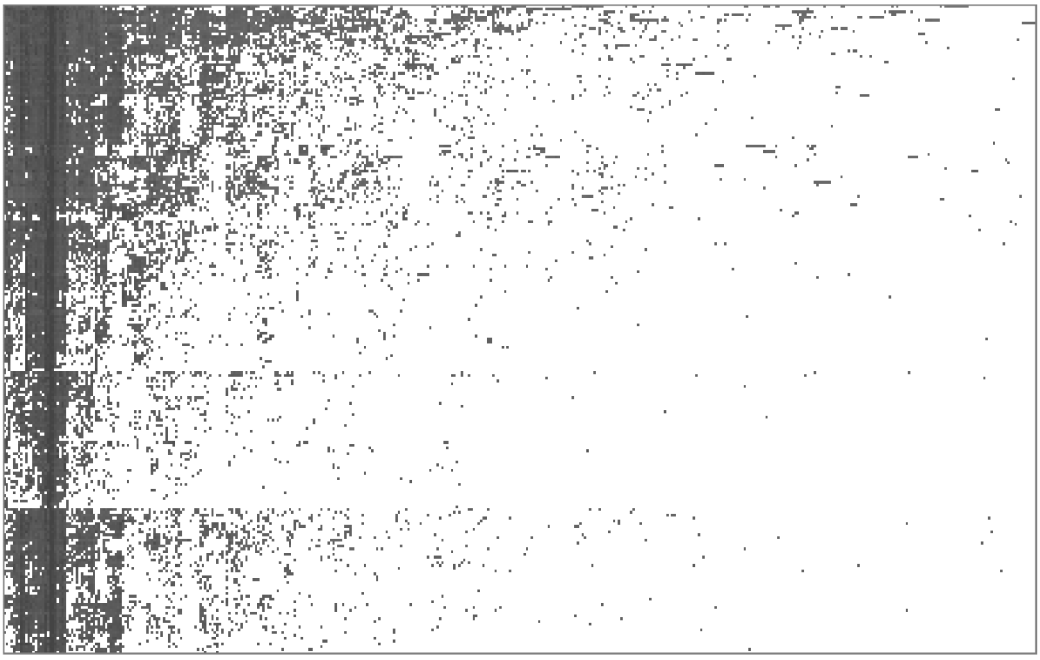


Figure 3.7: Log of $\mathbf{D}$. White represents values closer to $-\infty$ and darkest grey represents least negative values.

same form and with the same prior as $\mathbf{P}$ was used for this purpose. That is, for the purposes of this pre-training, $\mathbf{S}$ was treated as though its cells are distributed Exponential(λ). Thus, the pre-training reduces to a standard non-negative matrix factorization (NMF) but for the fact that the cells of $\mathbf{D}$ are assumed to be unobserved and beta-distributed, where standard NMF assumes the factorized matrix is observed with gaussian-distributed cells.

The step size for d_{mn} is the same as given above. The step size for both s_{mk} and b_{kn} is given by[14]

$$\text{step-size}(s_{mk}) = r \max\left[0, \min\left(s_{mk}, |\log s_{mk}| + 1\right)\right]$$

$$\text{step-size}(b_{kn}) = r \max\left[0, \min\left(b_{kn}, |\log b_{kn}| + 1\right)\right]$$

where r is a learning rate parameter (set to 0.01 for all simulations). Analogous to those in the previous section, the additive updates are given by

$$\text{update}(\mathbf{S}, m, k) = \text{step-size}(s_{mk}) \tanh\left(\frac{\partial}{\partial s_{mk}}\left[\log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})\right) + \log \mathbb{P}(\mathbf{S} \mid \lambda)\right]\right)$$

$$\text{update}(\mathbf{P}, k, n) = \text{step-size}(b_{kn}) \tanh\left(\frac{\partial}{\partial b_{kn}}\left[\log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})\right) + \log \mathbb{P}(\mathbf{P} \mid \lambda)\right]\right)$$

Because the $\mathbf{S}$ inferred by this procedure is not of the correct form, the cells of the $\mathbf{S}$ resulting from the above procedure were thresholded by the median of the column in which they lie: cells were set to 1 if they fell above the column median

[14]This step size function is a rectifier from $(-\infty, 1)$ and ln from $(1, \infty)$.

and 0 otherwise, thus converting $\mathbf{S}$ into a bit matrix for sampling in the next stage.[15]

Without any further processing, this procedure is likely to significantly decrease the log-posterior of the true model, since in most cases, the maximum value in a particular column of $\mathbf{S}$ prior to the above thresholding procedure is far below 1. This results in cells of $\mathbf{SP}$ that are far too large. To remedy this, the maximum along each column of $\mathbf{S}$ was calculated prior to the above thresholding and used to scale each row of $\mathbf{P}$. This returns $\mathbf{SP}$ to a scale somewhat similar to the one resulting from the NMF pre-training, though its cells will still be somewhat too large.

(Before moving on, it is worth noting that a MAP pre-training procedure that treat $\mathbf{S}$ as continuous but bounded on $(0, 1)$ in the same way as $\mathbf{D}$ was also tried, but the results of this procedure were poor. This approach may still be feasible with some tweaking of various hyperparameters, but in various trials runs, I could not find such an appropriate set of parameters.)

3.3.2.3 Training ($\mathbf{D}$, $\mathbf{P}$, and $\mathbf{S}$)

Subsequent to the MLE pre-training described above, $\mathbf{D}$, $\mathbf{S}$ and $\mathbf{P}$ were pre-trained by iteratively optimizing their respective log-posteriors. On each iteration, $\mathbf{S}$ was sampled using the Gibbs sampling equations described above, and $\mathbf{D}$ and $\mathbf{P}$ were incremented once using the update equations described in the previous two sections.

[15] This particular thresholding procedure has a secondary benefit in that exactly half of the cells in a particular column will be 1 and thus it is easy to tell how far $\mathbf{S}$ has moved from its initial state over the course of sampling.

Dataset	Generalized Discrimination	Ordinal Scale
PukWaC	0.129 ($p = 0.028$)	0.1348 ($p = 0.059$)
Schulte im Walde	0.189 ($p = 0.001$)	0.142 ($p = 0.023$)
VALEX	0.123 ($p = 0.0399$)	0.064 ($p = 0.218$)

Table 3.1: Spearman rank correlation between Jensen-Shannon divergence derived from three different datasets and similarity judgments. P-values derived from Mantel (permutation) test with 10000 iterations.

3.3.3 Results

3.3.3.1 Basic correlational analysis

As in the previous chapter, I begin with a basic correlational analysis to assess the relationship between distances defined on the syntactic distribution themselves and the two similarity judgment tasks presented in the last chapter. This analysis was carried out on all three datasets referenced in Section 3.3.1. For each dataset, the conditional probability of each subcategorization frame given each verb was estimated by taking the conditional relative frequency with additive smoothing (following Schulte im Walde 2006, $\lambda = 0.5$). Jensen-Shannon divergence was used as the distance measure, and so the additive smoothing is necessary here, since that measure does not tolerate zeros.

Table 3.1 shows the correlation between the distances as derived above and the two similarity judgment tasks. The correlations here are much lower than the ones seen in the last chapter between the acceptability judgment-based distance and the similarity judgment-based distance, which were 0.28 for the generalized discrimination similarities and 0.27 for the ordinal similarities. This interesting in the sense that, in that section, both distances defined directly on the acceptability judg-

ments and on the features extracted from those acceptability judgments using the nonnegative projection model showed similar levels of correlation with the distances derived from the similarity judgments. One difference here is that, whereas the subcategorization frames for the acceptability judgment task were selected specifically for to discriminate well among attitude verbs, many of the subcategorization frames extracted from the corpus may be irrelevant to distinguishing among particular attitude verbs.

If this is the case, there are two possibilities: (i) these irrelevant subcategorization may make it difficult for the model to extract semantically relevant features; or (ii) the model may be able to cut through these irrelevant subcategorization frames to discover semantic features relevant to participants similarity judgments. I show in the next section that the latter appears to be the case.

3.3.3.2 Model analysis

The model was trained with number of latent feature values ranging from 2 to 30 with 10 chains per number of latent features. The chain that converges on the lowest mean likelihood across samples was kept. Ideally, a stopping criterion similar to the one employed in the last chapter—WAIC—would be used here to select the optimal number of features. The problem is that, while the likelihood falls with higher numbers of latent features—the likelihood at 2 features is -9099346.0 and the likelihood at 30 is -9091140.0—these likelihood values do not always decrease over two adjacent pairs. This suggests sensitivity to the initialization conditions,

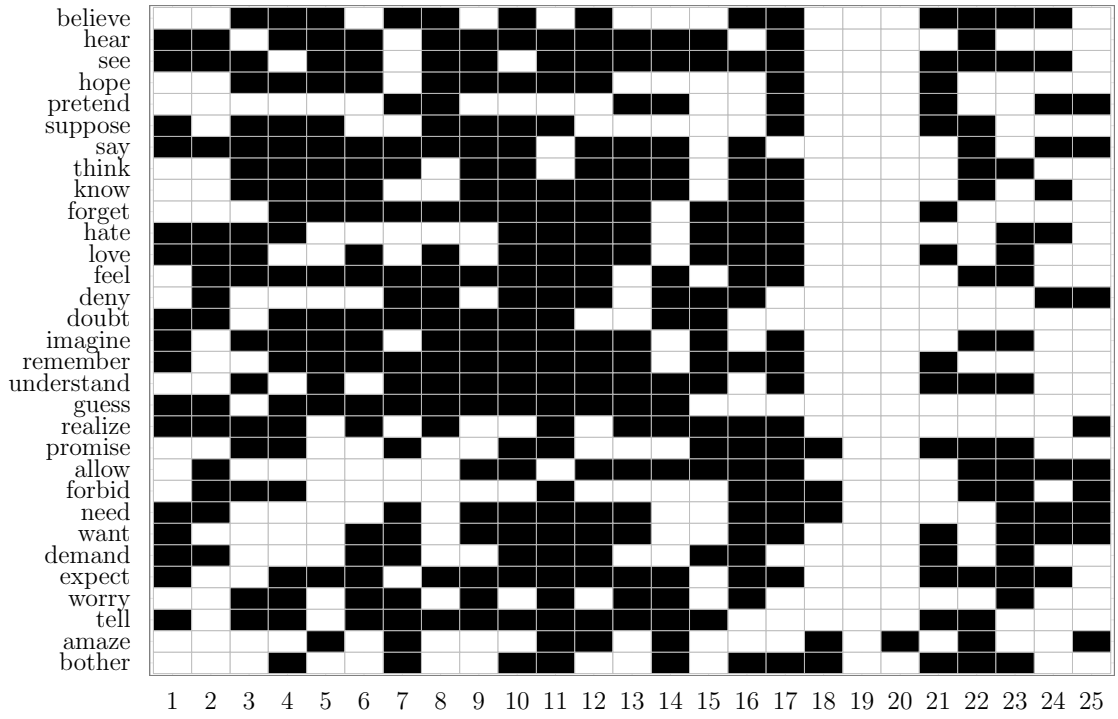


Figure 3.8: Features extracted from PukWaC dataset using nonnegative projection model of syntactic bootstrapping.

which was the impetus for running multiple chains; in spite of this, it appears that for some numbers of features the fitting algorithm was not sampling from the true posterior.[16]

To remedy this, I take the chain with the lowest mean likelihood, which was one run for 25 latent features. The final $\mathbf{S}$ was taken from this chain and all subsequent analyses are based on this $\mathbf{S}$, which can be seen in Figure 3.8.

As in the last chapter, to measure the information shared between the inferred features and the similarity judgments, I take the correlations between Manhattan distance defined on this $\mathbf{S}$ and the two similarity judgment-based distances. Both the

[16]Indeed, this sampler will not necessarily sample from the true posterior, since optimizers were used for $\mathbf{D}$ and $\mathbf{P}$.

correlation between the generalized discrimination judgments (Spearman $\rho = 0.238$, Mantel[iter=10000] $p < .001$) and the ordinal judgments (Spearman $\rho = 0.245$, Mantel[iter=10000] $p < .001$) are comparable to the correlations found in the last chapter and are also substantially greater than those reported in the last section. This is interesting in light of the fact that the correlations defined directly on the observed syntactic distributions was much lower, possibly suggesting that the model is filtering out irrelevant aspects of the distributions.

3.4 Discussion

In this chapter, I reviewed two classes of models of word representation: category-based and vector space models. I argued that, on the one hand, the category-based models has a representation of words that is constrained in an inappropriate way due to what I called the global normalization property; on the other hand, vector space models have representations that are not constrained enough making their feature values uninterpretable. I then gave a model that I argued was a natural midpoint between these models that builds on each's strengths without inheriting their problems. I showed that this model extracts features from performance (count) distributions that compare in their correlations with the similarity judgments to those extracted from the acceptability judgments presented in Chapter 2, despite the fact that the raw correlations between the performance distributions and the semantic similarity judgments were quite a bit lower.

Chapter 4: Incrementality in syntactic bootstrapping

In the previous two chapters, I showed that there is quite fine-grained semantic information present in propositional attitude verbs' syntactic distributions—both their *competence distributions* (Chapter 2) and their *performance distributions* (Chapter 3). In the course of doing this, I developed a computational model of projection that could take advantage of this information. I showed that when trained on either acceptability judgments or corpus data, using appropriate models of the data generating process, this projection model can be used to predict participants semantic similarity judgments well. The conclusion I draw from these results is that, if a learner had access to (propositional attitude) verbs' syntactic distributions in their entirety they could learn quite fine-grained aspects of those words' meanings.

One question that arises here is whether it is reasonable to assume that learners have such access to verbs' entire syntactic distribution when making inferences about those verbs' meanings. The answer to this is almost surely no. Learners receive data incrementally, and it seems likely that the inferences that underlie word-learning also operate incrementally. The question that naturally arises, then, is whether learners can take advantage of propositional attitude verbs' syntactic distributions in an incremental setting.

In this chapter, I present three experimental studies aimed at answering this question. The chapter begins with a review of the Human Simulation Paradigm (HSP; Gillette et al. 1999) and related tasks that serve as the inspiration for the current studies. This paradigm is discussed extensively below, but quickly: the idea behind HSP is to provide adult participants with a context—a scene, some cooccurring lexical items, etc.—in which a particular word was uttered but from which the word itself has been removed. Participants are then asked the free choice question: what word occurred in that context? By manipulating the particular kinds of contextual information presented, one can then investigate the amount of information provided by particular kinds of contexts about particular kinds of words (Gillette et al., 1999; Snedeker and Trueswell, 2004; Papafragou et al., 2007).

After this review of HSP, I then move onto the first two experiments, which simultaneously serve as norming studies for the third experiment but which are also of interest in their own right. The first of these norming studies is a special case of HSP akin to that used by Medina et al. (2011) and Trueswell et al. (2013) to measure the informativity of particular contexts themselves (as opposed to comparing types of contexts against each other). In this variant, participants are not learning a word given a set of contexts but rather treating each instance as a separate word. One can then measure various properties of the response distributions to particular items—e.g. how often the true word that occurred in the context was actually recovered—to assess their informativity about the true word's meaning.

In general, accurate recovery of the true word is the dependent measure analyzed in HSP—whether used in the traditional way or as a measure of item informa-

tivity. As has long been noted in this literature, this is not an ideal state of affairs, since it doesn't account for the fact that participants may give a response whose meaning is quite close to the true word's. I show this qualitatively in analysis of the first norming task, motivating a second norming task that aims to measure this closeness quantitatively.

Related to this worry is a second worry: when one asks participants to recover the word found in a particular context, the participant is explicitly asked to discretize their hypothesis around a single word meaning. The question of how discrete this hypothesis is has been a topic of recent debate—in particular, do participants keep track of uncertainty about a word's meaning? In the attitude verb domain, this question is particularly poignant, since possibly unlike the well-studied noun-learning domain, the verb domain seems to involve quite a bit more complexity that may not make it amenable to using standard methods that ask participants to choose discrete responses. To remedy this, I develop a novel extension of HSP that I refer to as Spatial HSP (SHSP). The idea behind using spatial SHSP is to avoid the possibly problematic forced discretization that comes with using a standard HSP task. The task is spatial in the sense that, instead of giving categorical responses regarding their hypothesis about a word meaning, they give similarity judgments between the word they are learning and the words they already know, thus hopefully enabling the tapping of uncertainty about the word's meaning.

4.1 The human simulation paradigm

4.1.1 The classic paradigm

The human simulation paradigm (HSP), introduced in Gillette et al. 1999, is a standard instrument used in word-learning research.[1] The task in this paradigm is simple: adult participants are given some information about the context a word was uttered in—scenes, surrounding words, structures, etc.—and they are asked which word occurred in that context. The idea behind using adult learners here is that, while there is a question of what level of conceptual development child learners have attained at a particular age, adult learners presumably have some level of conceptual development that allows them to grasp the meanings of the sorts of words children learn in the first few years of life. This in turn allows one to ask questions about the informational properties of various purported learning cues, controlling for conceptual development.

In Gillette et al. 1999, six different contextual conditions are tested, which form the basis for later work in HSP. Their first experiment investigates the usefulness of solely nonlinguistic scene information plus lexical category information. For this experiment, they chose the 24 most frequent nouns and the 24 most frequent verbs from a transcript of video-recorded play sessions. They then collected 6 video clips for each verb and played participants each of these clips in sequence, with a beep occurring when the word occurred. Participants were asked to guess at each

[1] See also Snedeker et al. (1999); Snedeker (2000) for work in this paradigm from around the same time.

beep which word occurred in that position, having been told (i) that the beeps for a particular set of clips all involved the same word and (ii) whether the word was a noun or verb. Participants did much better with nouns than with verbs for this experiment. They further showed that this could be wholly predicted by the imageability rating for particular nouns or verbs, where verbs like *run* are rated much more imageable than propositional attitude verbs like *think*.

Gillette et al.'s second task, which is the one important for current purposes, manipulated the kind of information participants had access to for making inferences about the word meaning: (i) scene information only—the original task—(ii) items from lexical categories (noun, verb, adjective) that surround the word; (iii) scene information plus lexical information; (iv) syntactic frames with nonce words replacing words from lexical categories; (v) the full sentence the word occurred in (syntactic frame + word from lexical categories); and (vi) all of the previous kinds of information. The upshot is roughly that more information is better.

Focusing in on the verbs, Snedeker and Gleitman (2004) replicate this more-information-is-better result (see also discussion in Gleitman et al., 2005).[2] They argue that verbs fall into three groups with respect to the sorts of cues used to learn them (and further that these three groups correspond to well-defined development stages): "relatively concrete verbs that describe specific actions in and on the observable world (*fall*, *stand*, *turn*, *play*, *wait*, *hammer*, *push*, *throw*, *pop*), the more abstract mental-content verbs (*know*, *like*, *see*, *say*, *think*, *love*, *look*, *want*),

[2]Indeed, they argue that it is actually a partial ordering, but this is not relevant for current purposes.

and a third set, of what have been called light verbs (*come*, *do*, *get*, *go*, *have*, *make*, *put*)." Relevant for current purposes, perhaps unsurpringly, the mental-content verbs (propositional attitude verbs) have low accuracies in the contexts (i-iii) from above but show a spike in accuracy when moving to contexts (iv-vi)—i.e. when one has syntactic information, possibly in concert with lexical information.

Papafragou et al. (2007) replicate this strong effect of syntactic information on ability to infer propositional attitude verb meanings, but they also show that scene information is not totally irrelevant. If participants are given scenes involving something that highlights a character's beliefs, participants are more willing to conjecture propositional attitude verb meanings. This is further reinforced by combining these scenes with linguistic—specifically, syntactic—information.

4.1.2 Norming HSP with HSP

A question that has arisen recently as a topic of debate in the word-learning is how memory constrains learners' ability to utilize cues to a word's meaning (Medina et al., 2011; Trueswell et al., 2013). In many accounts of noun-learning in particular, the assumption is often that learners have access to the full history of their experience with a word, or at least some nonnegligible chunk (see Yu and Smith 2012 for recent discussion).

One way memory constraints have been investigated is to ask how sensitive to the information carried by a particular piece of contextual information a learner's hypotheses are with the idea that, to the extent that learners decisions are based on

particular learning instances—the less smooth their learning trajectory—the more constrained their memory for previous instances is likely to be. Testing this requires some way of measuring the information carried by particular learning instances. It is this, rather than the particulars of the learning mechanisms proposed in this subliterature that I am interested in.

Following Gillette et al. (1999), Medina et al. (2011) selected the 24 most frequent nouns and 24 most frequent verbs from a video corpus developed by one of the authors. They then selected 288 from this corpus and presented each of 37 participants with 96 scene-only vignettes (2 per word). Unlike the standard HSP, in which participants would see 6 clips in a row for the same word, each of the clips in the Medina et al. study was completely disconnected from the others. The idea here is that, by disconnecting the vignettes, the informativity of each can be measured on its own terms. They take as their measure of informativity, the accuracy with which participants can recover the actual word that occurred in the vignette (as is standard, at the point of a beep).

Medina et al. then threshold items into two sets High Informativity (HI) and Low Informativity (LI) by their accuracy, choosing 33%, as this gave them a $1:5$ ratio of HI:LI vignettes.[3] They then use this partitioning to construct sequences of five vignettes with one HI instances and four LI instances, thus using the experiment with disconnected vignettes as a norming experiment for the second, a standard HSP task. They find that manipulating the placement of the HI instances significantly

[3]No verb or abstract noun vignettes showed up as HI vignettes, so Medina et al. drop verbs and abstract nouns from consideration for the rest of the experiments.

affects participants' ability to eventually recover the correct word. In particular, the earlier a high informativity instance the better.

4.1.3 Spatial HSP

In the current series of experiments, I adapt this strategy of first measuring the informativity of particular items, then using that to construct training sets, to HSP with syntactic frames instead of scene information, as in Medina et al. (2011). Before moving onto the actual experiments, however, I would like to first make a note on some crucial methodological alterations to the HSP tasks described above.

In Medina et al. 2011, accuracy is used as the measure of item informativity. But while accuracy is useful as a rough measure of informativity, going back to Gillette et al. 1999, accuracy has had known issues as such a measure, since it doesn't take into account responses that may be semantically close to the true word but which are nonetheless inaccurate. As I show evidence of in the next section, this makes this measure susceptible to various issues. Of particular note, no distinction is made in an accuracy measure between really bad guesses, such as ones that don't even fall into the correct syntactic category, and better guesses, which might involve words that are semantically close to the true word. This second problem is exacerbated when an item might be very informative about the semantics of the word that occurred in it, but when a semantically related high frequency word—plausibly, one that comes to mind more quickly—also fits well within the context. I show evidence that such frequency effects happen in the current experiments.

I remedy this state of affairs by implementing a, to my knowledge, novel extension of the Medina et al. norming task that takes into account not only participants' ability to recover the exact true word, but also semantically related words. To do this, I employ an ordinal scale semantic similarity judgment task of the sort used in Chapters 2 and 3 to measure the relationship between syntax in competence distribution (Chapter 2) and performance distribution (Chapter 3) to measure the relationship between participants' responses and the true word. To build this task, I extract all (lemmatized) responses from the previous task and pair them with the true word that occurred in the item they were a response to, and gather similarity judgments for those pairs. These similarity judgments are then used to get an estimate of the distribution of similarity among responses to particular items, which gives a more fine-grained view of those items' informativity.

A potential problem for understanding verb semantics related to standard HSP's use of accuracy as a measure of informativity is that HSP gathers free choice word responses. This is a problem in that, as I have discussed in earlier chapters, many of the verbs of interest in this dissertation seem to display multiple semantic features at once, and thus learners might plausibly be at least somewhat certain that a particular word has some features but not certain whether it has others. But in the classic paradigm, participants are forced to choose a particular semantics on each trial. On the whole this might not be problematic if the method participants use for selecting a discrete choice is guided by the uncertainty that they have about a word's features. We should then see these uncertainties arising out of aggregate behavior. But this is not an ideal state of affairs; it relies on faithful mappings

from the representation a particular participant has of a particular word's semantic features to whatever guides that participant's procedure for selecting a particular instantiation of those features to then give a response. Indeed, even if these mappings are faithful, passing the decision procedure through the participant's lexicon, may create undesired warping effects.

To remedy this, I present a novel extension of HSP that aims at more faithfully measuring participants' uncertainty about the semantic features a verb has, which I call the spatial HSP. In this task, instead of giving free choice responses, participants are tested using a similarity judgment task that asks them to compare a novel word they just learned from some training set—here, items culled from the norming tasks described above—to words they already know. Thus, instead of being forced to choose an instantiation of semantic features they may be uncertain about, participants can hopefully show their uncertainty more clearly. The use of the two norming tasks is then, following Medina et al. to understand how the distribution of informativity in the training sets affects participants uncertainty.

4.2 Norming tasks

In this section, I present two norming tasks that will be used to construct training sets for the task presented in Section 4.3. These two tasks together were aimed at measuring the informativity of each syntactic context about the meaning of the word that occurred there. The first task is a linguistic context-only HSP task with disconnected instances. That is, it is not the case that participants were told

they were learning the same word across instances. The second task is a likert scale similarity judgment task built from the responses gathered in the first.

4.2.1 Human simulation norming

4.2.1.1 Design

This norming task takes the form of a standard human simulation task with only linguistic context. In this task, participants are given a sentence with a blank somewhere in it and are asked to fill the blank with the word they think most people would respond with. All sentences were sampled from a corpus of child-directed speech as described below, and thus the blank replaces a real word. For instance, (1a) was an actual sentence used in the experiment.

(1) a. I **told** you I'm not having a new baby now.

b. I florped you I'm not having a new baby now.

This task was conducted online, and responses were collected using an HTML text box. This text box was filled with a greyed out placeholder verb *florp* that disappeared when a participant started typing, implemented using the standard HTML `input` tag `placeholder` attribute. This placeholder verb had tense/aspect morphology matching the verb that it takes the place of. For instance, (1b) shows the sentence derived from the true sentence (1a). Text boxes were autofocused to allow participants to use only the keyboard while performing the task.

The norming items fell into one of 40 conditions as formed from a full cross of the following three factors whose descriptions are given in subsequent sections: VERB (levels: 10 attitudes verbs), LEXICAL CONTENT (levels: *real*, *nonce*), and CONTEXT OF UTTERANCE (levels: *dinner*, *play*). Participants received one item from each condition for a total of 40 items.

In addition to providing typed responses to the HSP task, participants were asked a memory question about the item they just responded to in order to ensure they were paying attention. The rationale and construction of this memory task is described further in the next section.

4.2.1.2 Materials

4.2.1.2.1 Corpus sampling procedure For each of the 31 verbs investigated in the studies in Chapter 2, all sentences containing at least one of those verbs were extracted from Gleason corpus (Masur and Gleason, 1980), which is part of the CHILDES database (MacWhinney, 2014b,a).[4] The description of this corpus, taken from the CHILDES manual for North American English corpora (p. 44), is as follows.

> The participants are 24 children aged 2;1 to 5;2 who were recorded in interactions (a) with their mother, (b) with their father, and (c) at the dinner table. The 24 participants were recruited through nursery schools and similar networks, and were from middle-class families in the greater

[4]CHILDES provides a lemmatized version of the sentence. This lemmatized version was used for the search.

> Boston area. There were 12 boys and 12 girls. All families were White, and English was spoken as a first language in all families. Each child was seen three times: once in the laboratory with the mother; once in the laboratory with the father; and once at dinner with both mother and father. The laboratory sessions were videotaped and audiotaped, and the dinners were only audiotaped. Laboratory sessions included: (a) play with a toy auto, (b) reading a picture book, and (c) playing store.

This corpus was chosen because, unlike many corpora in CHILDES, it provides data from two different contexts that children commonly find themselves in—*play* and *meal* (*dinner*) contexts—thus heightening the chances that data sampled from this context are more representative of children's linguistic experience overall. It also provides data from an age range where children's verb vocabularies are rapidly developing—in particular, where much development of the attitude lexicon occurs (de Villiers, 2005).

In this first norming task, both contexts were considered as separate factor levels. The number of sentences each verb occurred in within both the play and dinner sections of the corpus (across children) were then tabulated, and the top ten most frequent verbs from the previous study found. For each of these high frequency verbs, up to 30 sentences were taken from the dinner sessions and up to 60 taken from play sessions (30 from the mothers' play sessions and 30 from the fathers').

Each of these sentences was then hand-checked for transcription errors and acceptability. Unacceptable sentences were marked for exclusion, including those that might seem unacceptable out of context. For instance, taking examples from the sentences extracted for *know*, such sentences often involve continuations, as in (2), or discourse-dependent processes, like the topicalization in (3).

(2) I think next time you go off the board, you **know**, I think you will dive in instead of jump in, okay?

(3) That, I **know**.

In other cases, it seemed likely that an acceptable sentence might be hard to parse in the context of a human simulation task. To retain as much fidelity to the true syntactic distribution—what I have been calling *performance distribution*—of the verb while reducing this complexity, sentences for which it was possible were modified from their original form without changing the syntactic structure or selectional relationships to the verb in question. For instance, the conditional antecedent and matrix verb in (4) were excluded to create new sentence with only the conditional consequent.

(4) ~~I said, if you have to really start really considering it,~~ it is impossible to make that kind of decision, you **know**?

After this acceptability checking and modification procedure was complete, 20 sentences were subsampled for each verb from each modified context set (play and dinner), excluding the unacceptable sentences. These sentences form the *real* level

of the LEXICAL CONTEXT factor.

To create the items in the *nonce* level of the LEXICAL CONTEXT factor, all nouns, verbs, adjectives, and adverbs for the above sentences were replaced with nonce words with morphology matching the ones found on the real words found in the original sentence. All determiners, prepositions/particles (of, to, at, up, etc.), and complementizers (that, if, for, to) were retained. Among the determiners were included quantificational determiners/quantifiers (*every(thing)*, *any(thing)*, etc.) and WH words (*who*, *what*, *where*, etc.). The intention here was to retain only words from functional categories.[5] To this end, some nouns and verb-like elements were also retained.

These noun exceptions included all personal pronouns (*I*, *you*, *(s)he*, *me*, *mine*, etc.) as well as temporal (*now*, *then*) and locative indexicals (*here*, *there*).[6] These exceptions seem reasonable since under many theories, they fall into the determiner class or are at least partially constituted by a determiner-like meanings.

Verb exceptions included all auxiliary verbs: all forms of *be*, perfect auxiliary forms of *have*, all modal auxiliaries (*can*, *might*, *must*, etc.). Semi-modals (*have to*, *ought to*) were treated as lexical verbs in this respect—i.e. *have* or *ought* would be

[5]There is a question here whether all prepositions are purely functional. This seems unlikely, but the replacement of prepositions can severely degrade participants ability to access the syntactic structure of a sentence. This is likely due to the fact that prepositions are relatively closed-class—at least compared to nouns, verbs, and adjectives. It is standard in human simulation paradigm experiments to retain prepositions (cf. Gillette et al., 1999).

[6]Under this criterion might fall temporal expressions like *today*, *yesterday*, and *tomorrow*, since they seem indexical in ways similar to *now* and *then*. The problem is that many complex temporal expressions, like *last night* or *next week* are similarly indexical, and it is unclear where to draw the line. One criterion could be to retain only single word indexical expressions, like *yesterday*, *today*, and *tomorrow*, but this privileges expressions that involve days over those that involve other time intervals, and thus some amount of arbitrariness is necessary. The current methods seems to me the most conservative if indexical expressions are to be retained.

replaced with a nonce word but *to* retained.

Finally, four common adverbs were excepted from replacement by a nonce word: *too*, *either*, and *else*. This seems reasonable since, in contrast to derived adverbs like *carefully* or *intentionally*, these adverbs' meanings are logical in nature and thus naturally fall into a class with, e.g., the pronouns. Indeed, all are anaphoric.

4.2.1.2.2 Memory task In many psycholinguistic experiments—e.g. acceptability judgment tasks—it is possible to analyze the distribution of reaction times across participants to assess whether they were in fact doing the task. (See Chapter 2 for an example of such a filtering procedure.) In this case, however, much more variability is expected in response speed due to differences in typing speed, meaning that reaction time analysis alone may be insufficient for detecting bad responders. (Indeed, we exclude no participants in this section based on the same reaction time criterion used in Chapter 2.) To this end, an additional measure was gathered to assess whether participants are in fact doing the task: a lexical memory task.[7]

Of the 20 real word items participants received, half were followed by the question "which word was in the previous sentence?" along with five words, only one of which was actually in the previous sentence. (None of the nonce word sentences were followed by this memory task.) For instance, participants who saw the sentence in (5) received the memory question along with the set of words in (6).

(5) I think what we should do is try to florp what we took apart last and put that together first.

[7]Much thanks to Ellen Lau for suggesting this.

(6) a. Which word was in the previous sentence?

b. {boy, mothers, together, never, family}

Participants' response accuracy for each item was then collected and analyzed for the purposes of data validation.

4.2.1.3 Participants

Participants were recruited until each item had at least 20 observations associated with it after the data validation procedure described in the last section. Participants were allowed to respond to up to three lists. 577 unique participants were recruited through Amazon Mechanical Turk (AMT) using a standard Human Intelligence Task (HIT) template designed for externally hosted experiments and modified for the specific task. Of these unique participants, 483 responded to a single list, 88 responded to two lists, and 6 responded to three lists. No participant that responded to multiple lists responded to the same list twice.

Prior to viewing the HIT, participants were required to score seven or better on a nine question qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. After finishing the experiment, participants received a 15-digit hex code, which they were instructed to enter into the HIT. Once this submission was received, participants were paid $1.

4.2.1.4 Data validation

Three data validation techniques were used. First, for each participant, the number of correct responses to the memory task were tabulated (by list for participants that responded to more than one list). The vast majority of participants (0.793) obtain perfect scores, with almost all of the remainder answering incorrectly only once (0.168) or twice (0.031). Given this distribution, only participants that scored 8 (of 10) or better on the memory task were retained. This resulted in the exclusion of 3 participants: 1 who responded with only 3 correct, 1 that responded with only 6 correct, and 1 that responded with only 7 correct.[8]

Next, participants' log reaction times (log RTs) were analyzed. First, each participant's median log RT and the interquartile range(IQR)—the difference between 25^{th} and 75^{th} precentiles)—of their log RTs were computed. The median and IQR of each of these statistics was then computed over participants. Participants were excluded using Tukey's method, wherein the Tukey interval ([Q1-1.5*IQR, Q3+1.5*IQR]) is constructed for both by-participant medians and IQRs and participants excluded if their median log RT or IQR log RT fell outside this interval. No participant's median log RT fell outside the Tukey interval of median log RT over participants and no participant's IQR log RT fell outside the Tukey interval of IQR log RT over participants; thus no participants were excluded under these criteria.

The median log RT-based exclusion procedure was also conducted for par-

[8]Even if the participant with 7 correct were retained at this stage of validation, most of that participant's responses would be excluded in the third stage due to the fact that that participant gave almost solely nonce word responses.

ticular responses. For each participant, the IQR of the log RT for that participant's responses was computed. Responses were then excluded if they fell below the participant-specific Tukey interval. 6 responses (across participants) were excluded in this way.

The final filtering step was to exclude all nonword responses. As an initial approximation, nonword was defined as any word not occurring at least once in the PukWaC corpus. Of the 26953 total response tokens and 1517 response types, 258 response tokens and 85 response types were marked as nonwords in this way. Those words that were marked as nonwords were then handchecked. Many cases either involved the participant responding with the placeholder verb—i.e. *florp*, *florps*, *florping*, or *florped*—a random string[9]—e.g. *lmpw* or *toxat*—or a multiword string[10]—e.g. *where were you* or *he asked me*.

Other responses, however, were clear typos. (Indeed, multiple participants emailed to apologize for having made a typo somewhere in the experiment.) For instance, *know* had three typo variants—*knlw*, *knkow* and *knokw*. When their correct variant was clear, these typos were corrected manually. For instance, *knkow* would be changed to *know*.[11] These corrections results in 16 of the datapoints original marked as nonword responses to become word responses. The remaining 242 were then excluded.

One problem with using corpus counts to filter nonwords is that, while filtered

[9]Some participants' strategy in this case was to type a nonword from the sentence itself. For instance, two of the nonwords used in the experiment were *spurply* and *slargle*, and these were both given as responses.

[10]Participants were explicitly instructed not to do this at multiple points in the instructions.

[11]These changes were not made to the raw data itself, but rather in the analysis script, and are documented in the analysis scripts made available on my github.

words will tend to be nonwords (the method has high precision) some nonfiltered words may still be nonwords, since common typos will be counted (the method has lower recall). A subsequent filtering step was thus conducted by hand. Of the 26711 response tokens and 1432 remaining after the first nonword filtering step, 75 response tokens from 62 response types were deemed nonwords that were not typo variants of a real word.[12] Of the 26636 response tokens and 1372 response levels remaining after this filtration, 44 were clear typo variants of true words, which were corrected. The final number of response tokens after filtering was thus 25636 and the final number of response types after filtering was 1328.

4.2.1.5 Results

In this section, I begin with a mixed effects regression analysis of accuracy as conditioned by LEXICAL CONTEXT and CONTEXT OF UTTERANCE, controlling variability due to participants, verbs, and items (nested within verb). I show that only LEXICAL CONTEXT, not CONTEXT OF UTTERANCE reliably conditions higher accuracy. This is something of a sanity check to ensure that the task fits with previous findings—it does—but it is also of wider interest, since to my knowledge no one has tested CONTEXT OF UTTERANCE in HSP before. Further, it allows for the explicit quantification of variability in informativity within a verb using random effects.

I then turn to three analyses focused on which properties of particular items

[12]In fact, for some of these nonwords, the intent was clear. For instance, *practic* is likely a typo of *practice*. These typos were only corrected if the true variant already showed up at least once elsewhere (not as a typo).

give rise to the overall accuracies I show in the first section. In the first analysis, I assess the extent to which participants are able to recover the correct syntactic category of a word. This is important, since if they can't recover the syntactic category, other higher order properties of the linguistic context, such as the syntactic structure, probably won't be accessible either. In the second analysis, I delve briefly into which aspects of the syntactic structure (tense information, complementizer information, etc.) predict accuracy. This analysis is somewhat limited in scope, since as I note in the final of these three analyses: a response may be inaccurate while still being quite close to the true response semantically. In this final analysis, I give a qualitative characterization of this "closeness." This characterization in turn motivates a quantitative method for assessing similarity of a response to the true word as a measure of the informativity of a particular item.

4.2.1.5.1 Accuracy I begin with an analysis of accuracies as they are conditioned by the two factors in the design. To repeat the original example from above, if the true sentence were (7a), the stimulus created from that sentence would be (7b), with *florped* the placeholder within a text box. The accuracy for this particular item (1 of 20 from its particular condition) would be calculated as the number of times *told* were given as a response to (7b) over the total responses for (7b).

(7) a. I **told** you I'm not having a new baby now.

b. I florped you I'm not having a new baby now.

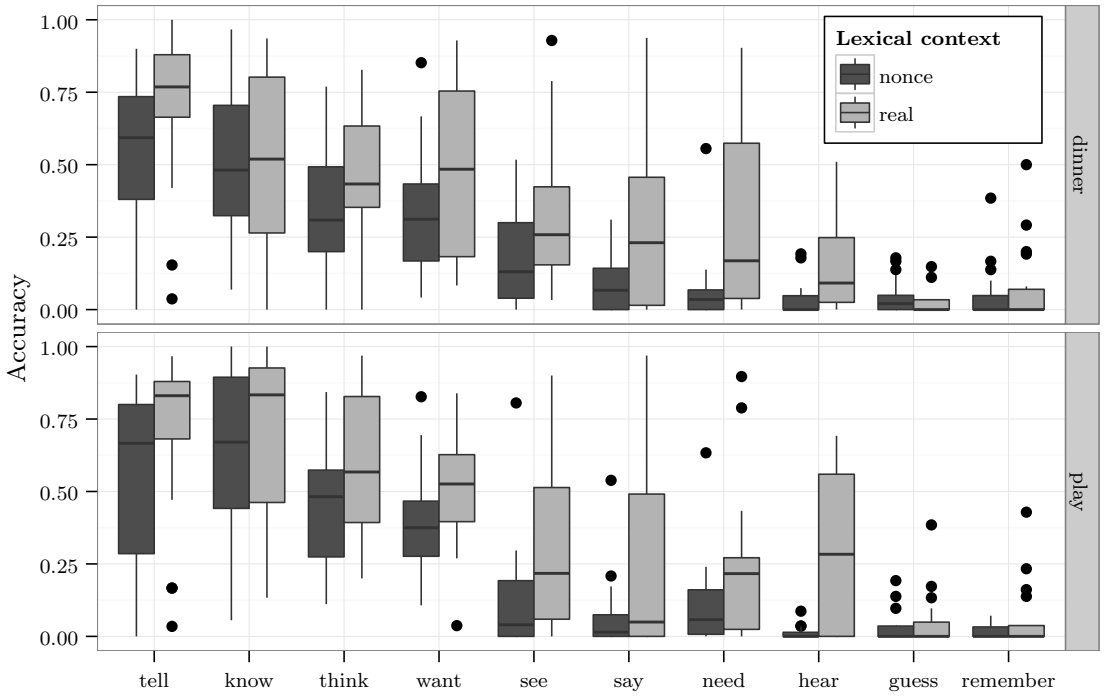


Figure 4.1: Distribution of by-item accuracy given verb, lexical context, and context of utterance. Each box represents the distribution of accuracy over the 20 items in that condition.

To visualize the distribution of accuracy for particular items within particular conditions, the proportion of times an item was responded to with the word that actually occurred in the corpus was computed. Figure 4.1 shows the distribution of these accuracies across verbs as well as LEXICAL CONTEXT, given by the fill on the boxplot, and CONTEXT OF UTTERANCE, given by the facet. Each box thus represents the accuracies for 20 items in the standard way: boxes given Q1-Q3 and whiskers give the Tukey fence ([Q1-1.5*IQR, Q3+1.5*IQR]). We see that, on the whole, sentences whose content words were replaced with nonce words, had lower accuracy. This is not surprising given that these sentences were designed to remove some information that might help participants infer the meaning of the word in the blank.

Table 4.1: Fixed effects for mixed effects logistic regression accuracy model.

	Dependent variable:
	Accuracy
Intercept	-2.404^{***}
	(0.573)
LEXICAL CONTEXT: *real*	1.112^{***}
	(0.134)
Observations	26,636
Log Likelihood	$-12{,}620.700$
Note:	*p<0.1; **p<0.05; ***p<0.01

This trend is corroborated by the results of fitting a mixed effects logistic regression to these accuracy data with accuracy as the dependent variable; fixed effects of LEXICAL CONTEXT, CONTEXT OF UTTERANCE, and their interaction; random intercepts for participants, verbs, and items (nested under verbs). This model shows high correlation among the fixed effects, suggesting that some of the fixed effects may not be necessary. To assess this, the same model without the interaction was also fit and a likelihood ratio test conducted. This test did not reach significance ($\chi^2(1) = 0.129, p = 0.719$), and thus the interaction term was dropped.[13] The same procedure was carried out for LEXICAL CONTEXT ($\chi^2(1) = 64.077, p < 0.001$) and CONTEXT OF UTTERANCE ($\chi^2(1) = 0.110, p = 0.741$). Only LEXICAL CONTEXT was significant under this criterion and was thus retained while the others were dropped.

[13]Note that, though it is standard to talk in terms of likelihood ratio tests—firmly within a null hypothesis testing paradigm, this procedure will also almost always coincide with a reduction in the Akaike Information Criterion (AIC) as well, and so this model building procedure can be thought of as a procedure aimed at reducing overfitting.

4.2.1.5.1.1 Fixed effects Table 4.1 shows the fixed effect terms of the resulting model. This model is given in terms of reference coding, where the reference level is LEXICAL CONTEXT: *nonce*. Thus the intercept term gives the expected log-odds of an accurate response against an inaccurate response in the *nonce* condition. The upshot of this table is that the effect of the lexical context having real words as opposed to nonce words is reliably positive, thus confirming the apparent trend for participants to respond more accurately when the lexical context is real words. This, again, is unsurprising, since the whole point of including a nonce word condition was to remove some information relevant to making an inference about a word's meaning.

4.2.1.5.1.2 Random effects Turning to the random effects, the variance for the participant random intercepts is 0.173 (sd: 0.416) (in log-odds space); the variance for the verb random intercepts is much larger at 3.183 (sd: 1.784); and the variance for the item random intercepts was similarly large at 2.946 (sd: 1.716). This means that participants varied little in their ability to answer accurately—for comparison, the participant random intercept standard deviation is less than half the size of the estimated fixed effect of *lexical context*. Verbs and items on the other hand show much higher variability. This can be seen in Figure 4.1 in the fact that some verbs—e.g. *tell*, *know*, *think*, and *want*—have much higher accuracy over most of their items than other verbs—like *remember*, *guess*, and *hear*. The item variability can be seen in the size of the boxes for each verb; even the verbs with the highest median accuracies show a lot of variability in those accuracies.

4.2.1.5.1.3 Frequency effects? One possible explanation for the verb variability may be a frequency effect: verbs with higher frequencies may come to mind more readily during the task and so participants might also respond with these more readily. But if participants are more willing to respond with higher frequency verbs, those verbs might have a higher accuracy due to their frequency. To control for this, a second mixed model was fit with true word log frequency (obtained from the counts available from the ukWaC website) as a predictor alongside LEXICAL CONTEXT. To ensure that the intercept corresponds to the lowest frequency word in the set of true words (*guess*) instead of a log frequency of 0, the predictor was entered as the true word log frequency minus the minimum true word log frequency over all words. This predictor was significant ($\chi^2(1) = 3.847$, $p < 0.05$). A third model was fit with the interaction between TRUE WORD LOG FREQUENCY and LEXICAL CONTEXT, but this term was not significant.

Table 4.2 shows the fixed effects estimates for this model. Note that the effect of LEXICAL CONTEXT remains the same and the intercept lowers. Note that this lowering has to do with the fact that now the intercept correspond to the reference level LEXICAL CONTEXT: *nonce* at TRUE WORD LOG FREQUENCY: 0, which corresponds to the specific verb *guess* instead of the average at LEXICAL CONTEXT: *nonce*. The effect of TRUE WORD LOG FREQUENCY is approximately the same size as that of LEXICAL CONTEXT. This means that for every order of magnitude increase in frequency of the true word, participants were more accurate (on average) to about the same extent as if they had gotten a *real* item instead of

Table 4.2: Fixed effects for mixed effects logistic regression accuracy model with the addition of the true word's log frequency as a predictor.

	Dependent variable:
	Accuracy
Intercept	-4.518^{***}
	(1.077)
LEXICAL CONTEXT: *real*	1.112^{***}
	(0.134)
TRUE WORD LOG FREQUENCY	1.014^{**}
	(0.465)
Observations	26,636
Log Likelihood	−10,345.660
Note:	*p<0.1; **p<0.05; ***p<0.01

a *nonce* item.

Surprisingly, however, controlling for frequency does not affect the variance estimates for verb random intercepts all that much, though there is a reduction.[14] The new estimate of the variance for verb intercepts is 2.156 (sd: 1.469), which is about a 0.25 reduction in the standard deviation from the previous estimate. This suggests that not all variability in verb responses is driven by frequency effects.

4.2.1.5.1.4 What else might drive accuracy? To get an accurate response, it must be clear from a particular item which verb fits in that item. On the one hand, as I show in Section 4.2.1.5.4, many of the most common responses to particular items are verbs that, while not accurate because they don't match the true item, nonetheless share a semantic component with the true item and are thus intuitively

[14]One wouldn't expect it to affect the participant or item random intercepts, since frequency is a property of verbs, not participants or items, and indeed, those estimates remain constant.

closer in meaning than some random item. This would suggest an item that is somewhat informative about the semantics of a word, but not fully informative. On the other hand, some items may be extremely uninformative—to the point that even the syntactic category of the true response is unclear. These two states of affairs are quite different in nature, and ideally there would be some way of pulling them apart.

In the next section, I investigate the question of how easy it is to recover the correct syntactic category (verb) across items (Section 4.2.1.5.2). In the subsequent section, I then turn to a preliminary analysis of which syntactic features are most useful in giving an accurate response. Finally, I turn to the inaccurate responses that participants give in order to assess how informative items are about the meaning a verb has (Section 4.2.1.5.4).

4.2.1.5.2 Nonverb responses Going back to at least Brown 1957, it has been known that syntactic category is a useful cue to word meaning. Indeed, to use the syntactic distribution a verb occurs in to help infer its meaning, one first needs to know that the word they are dealing with is a verb in the first place. One possibility for at least some of the inaccuracy for some of the above verbs then could be uncertainty about the syntactic category that the word falls into. Here, I assess this as binary outcome: either the participant knew that an item was a verb or not.

To assess the overall uncertainty about syntactic category, responses were labeled for whether they were a verb or not by hand. As in Figure 4.1, Figure 4.2 shows the distribution over items of the proportion of nonverb responses to that

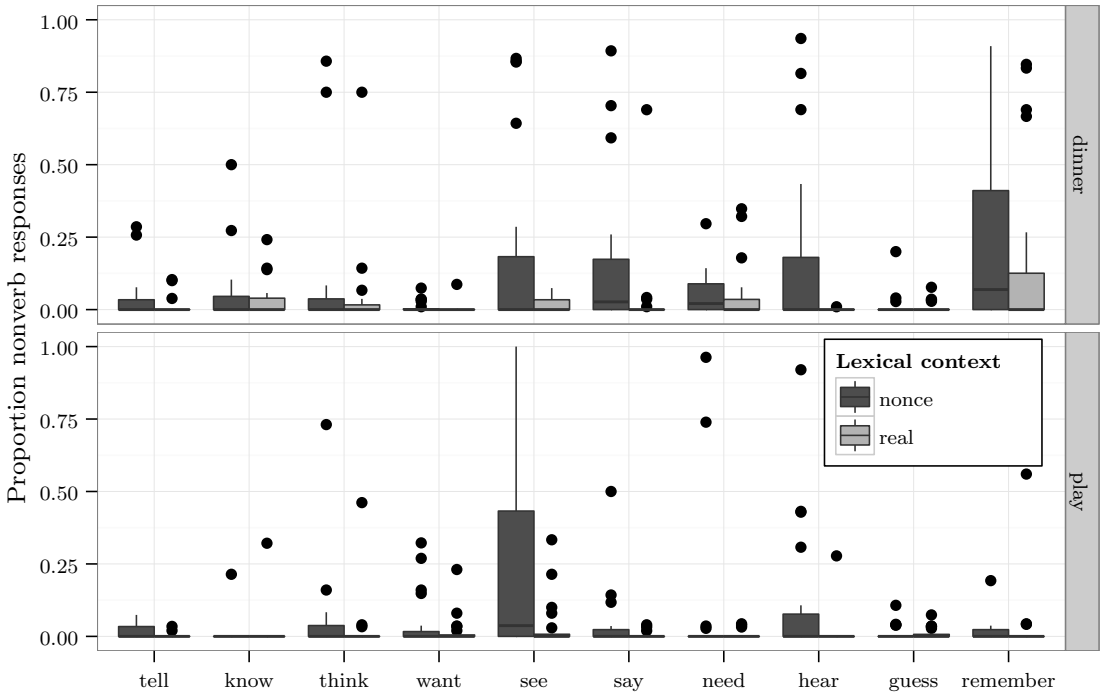


Figure 4.2: Distribution of by-item proportion of nonverb responses. Verbs are ordered as in Figure 4.1 (by median accuracy).

item. A model with the same structure as the one discussed in the last section was fit with NONVERB as the dependent variable, and the likelihood ratio test procedure repeated. In this case, as before, the interaction between LEXICAL CONTEXT and CONTEXT OF UTTERANCE is not significant; but in contrast to the previous case, both main effects of LEXICAL CONTEXT and CONTEXT OF UTTERANCE are significant.

Table 4.3 shows the fixed effect for this model. On the whole, the probability of nonverb responses is quite low, even at the reference level (LEXICAL CONTEXT: *nonce* × CONTEXT OF UTTERANCE: *dinner*). As one might expect, the chances of a nonverb response go further down when the items contain real words, but they also go down in the play context. This appears to be driven by more verbs having

Table 4.3: Fixed effects for mixed effects logistic regression nonverb model.

	Dependent variable:
	Nonverb
Constant	-6.413^{***}
	(0.457)
LEXICAL CONTEXT: *real*	-1.396^{***}
	(0.342)
CONTEXT OF UTTERANCE: *play*	-0.762^{*}
	(0.339)
Observations	26,636
Log Likelihood	$-2{,}981.346$
Note:	*p<0.05; **p<0.01; ***p<0.001

at least a few somewhat uncertain items in the dinner context, especially when they involve nonce words.

Dinner sentences were longer on average, and so one plausible driver of the increase in nonword responses is that, in longer sentences, participants get overloaded with the number of nonce words whose category they are uncertain about and then shut down, making a random guess.[15] To test this, a model including item word count alongside the other predictors as well as one including both item word count and its interaction with LEXICAL CONTEXT alongside CONTEXT OF UTTERANCE was constructed. Neither model showed significant improvement in likelihood ratio

[15]To establish that dinner sentences were longer on average, a mixed effects poisson regression was conducted on only the LEXICAL CONTEXT:*real* sentences with sentence length as the dependent variable, a fixed effect for CONTEXT OF UTTERANCE, and random intercepts for VERB. The CONTEXT OF UTTERANCE effect is significant under a likelihood ratio test comparing this model to one without the fixed effect ($\chi^2(1) = 5.08, p < 0.05$). The coefficient for the *dinner* level is furthermore positive (log increase in length: 0.084), suggesting that dinner sentences are longer on average. (Only LEXICAL CONTEXT:*real* were used in this regression since the LEXICAL CONTEXT:*nonce* sentences will necessarily have the same length as their corresponding LEXICAL CONTEXT:*real* sentence.)

tests ($\chi^2(2) = 1.546, p = 0.462$).

4.2.1.5.3 Syntactic features Once a participant knows that the word they are trying to recover is a verb, the question arises what aspect of the syntactic structure they might use in recovering the word. To assess this, I use an off-the-shelf method for measuring variable importance: mean gini decrease in a random forest (Breiman, 2001).[16]

First, the syntactic features of every item were hand-coded using the same feature set described in Chapter 3 (see that chapter for a description of each feature level): COMPLEMENTIZER (*none*, *finite*, *polar question*, *WH question*), EMBEDDED TENSE (*none/no embedding*, *infinitival*, *bare*, *gerund*, *tensed*), MATRIX SUBJECT (*referential*, *it*, *there*), EMBEDDED SUBJECT (*none/no embedding*, *nominative*, *accusative*, *case unknown*), MATRIX OBJECT 1 (*true*, *false*), MATRIX OBJECT 2 (*true*, *false*), MATRIX OBLIQUE (*true*, *false*). These features were then merged with their corresponding items in the HSP dataset, and along with VERB, LEXICAL CONTEXT, CONTEXT OF UTTERANCE, and TRUE WORD LOG FREQUENCY, were entered into a random forest classifier with 1000 trees and three variables tried at each split and ACCURACY as the dependent variable.

Table 4.4 shows the importance of each feature as measured by the mean decrease in the Gini obtained when including that feature. Higher numbers mean better predictability. The best predictor by far is VERB, followed closely by TRUE

[16]Initially, an analysis in the same family as the previous accuracy model was attempted, since each syntactic feature might have been entered in as a predictor and then its significance tested in a likelihood ratio test. However, these models showed poor convergence, likely due to gross imbalances in the distribution of particular feature values, and so a more robust method was used.

Table 4.4: Variable importance as measured by mean decrease in Gini

	mean decrease Gini
VERB	1,255.597
TRUE WORD LOG FREQUENCY	768.296
COMPLEMENTIZER	498.921
EMBEDDED TENSE	273.295
EMBEDDED SUBJECT	187.749
LEXICAL CONTEXT	175.070
MAIN OBJECT 1	131.388
CONTEXT OF UTTERANCE	71.683
MAIN SUBJECT	60.163
MAIN OBJECT	29.048

WORD FREQUENCY. This is somewhat unsurprising given that verbs show high variability in the accuracies associated with them (see Figure 4.1) and it was already noted that the log frequency of the true word also predicts higher accuracy. With regard to other predictors already discussed, the somewhat low importance score assigned to LEXICAL CONTEXT is interesting given its significance in the previous analysis.

Turning to the syntactic, COMPLEMENTIZER is the most useful of the syntactic features. This is likely driven by question complementizers, since overt reflexes of the nonquestion complementizers are rare; speakers almost always drop the complementizer in nonquestion complements. The second most important syntactic feature is EMBEDDED TENSE, which as noted in Chapter 2, is consistently associated (at least in English) with robust semantic distinctions. EMBEDDED SUBJECT and MAIN OBJECT 1 follows these. Embedded subject, which encodes the case of the embedded subject if it is ambiguous, may be a useful cue for distinguishing between tensed and untensed complements if that particular distinction is ambiguous in a particular

instance. MAIN OBJECT 1 is likely useful for determining whether something is a speech verb—in particular, *tell*.

4.2.1.5.4 Response distributions This is useful, but it's not quite the end of the story. One thing that the previous analyses do not take into account is that, even if participants don't give the exact verb that occurred in a particular position, they might nevertheless answer with one that is semantically similar. For instance, note that the overall accuracy for *remember* is quite low, and to some extent, this could be a product of nonverb responses. But even among verb responses, the sorts of responses participants give are far from random. As I will show, the most common response to *remember* sentences was not *remember*, but *know*. Considering that *know* seems to be involved in the meaning of *remember* at some level of representation—*x remembered p at t* seems to presuppose that *x knew p at* $s < t$—it's quite interesting that participants give this response.

To get a feel for how prevalent inaccurate-yet-semantically-similar responses are, I now turn to a qualitative analysis of the distribution of responses to each verb's items. In the next subsection (Section 4.2.2), this qualitative analysis will be augmented with a quantitative analysis that uses a similarity judgment task akin to those presented in Chapter 2. To delve into these responses, it will be useful to first find the proportion of times a word was given for a particular item, then look at the distribution of those proportions. This is analogous to looking at the distribution of accuracy over items shown in Figure 4.1, where the outcomes are binary (accurate v. nonaccurate); the differences is that here, the outcomes are treated as many-valued.

Prior to looking at these distributions, it's necessary to perform some preprocessing on the responses. Because some items will involve an inflected version of the verb, the correct response will be the verb's inflected form. For the purposes of computing response accuracy, a match between the true word in its inflected form and the response is ideal. For the current analysis, the inflectional morphology is likely not important. Therefore, it will be useful to map the true verb and the responses to their root form. One way to obtain these root forms is to use a stemmer. Stemmers make frequent mistakes, however, and since there are so few response types (see above for counts), it is easy to lemmatize by hand. This hand lemmatization was done for all response types, regardless of syntactic category, except for adverbs.[17]

After this lemmatization, the proportion of times a particular root form was given as a response to a particular item was computed. This results in a relative frequency distribution over the root form of responses for each item. Because (i) for current purposes we care about general trends over items for particular verbs and (ii) it is difficult to visualize each item's distribution in an easy to digest way, I graph these distributions by response type in Figure 4.3. This graph shows, for each verb, the response roots with the highest median relative frequency over items. Thus, as in 4.1, each bar represents datapoints for 20 items in the relevant condition. The bar itself gives the median and the error bar gives the range.

We see that for verbs *think*, *want*, *see*, *tell*, and *know*, the most common responses in both the *real* and *nonce* conditions are the true verb itself. This

[17]Two issues arose for this lemmatization: (i) all past participal forms were converted into their root verb, even though this might be problematic when the response was intended to be a deverbal adjective; and (ii) semi-modals like *supposed (to)* would be lemmatized to *suppose*.

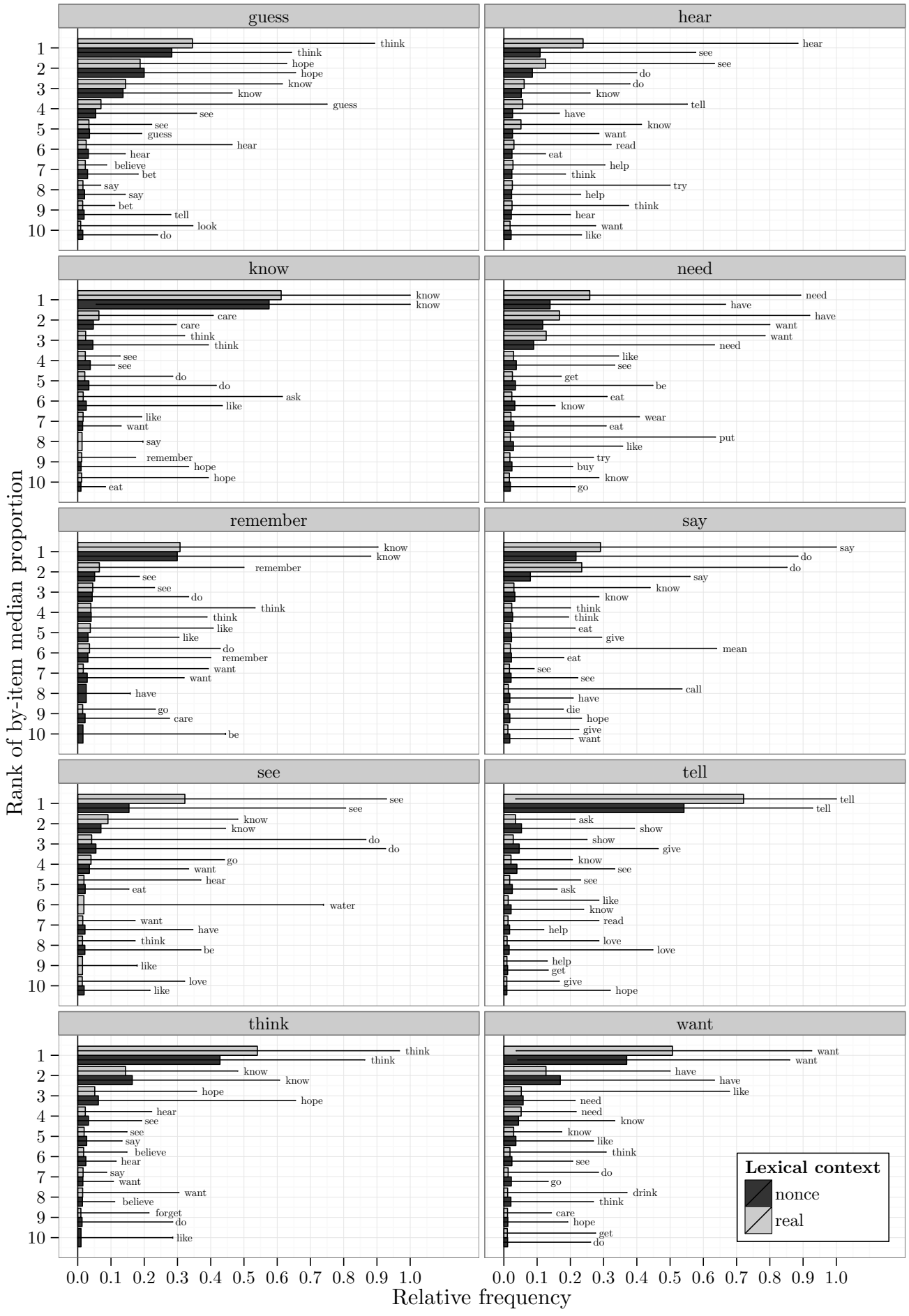


Figure 4.3: Distribution of by-item response root relative frequencies given verb, lexical context, and context of utterance. Each bar+error bar represents the distribution of relative frequency for the labeled response root over the 20 items in that condition.

suggests that many of the sentences these verbs are found in are highly informative about their semantics and, further, that this seems to be a result of their syntax to a great extent—since these verbs are guessed even without lexical context. Among the other responses for these verbs, *think* and *know*, notably, have responses that tend to involve representational attitudes: the second and third most common responses to *think* sentences are *know* and *hope*, respectively, in both *real* and *nonce* conditions and the third most common response to *know* sentences is *think*.[18] *Tell* has responses that tend to be communicative (*ask* and *show*) or that at least involve a transfer semantics (*give*). *See* has responses that are cognitive but nonperceptual (*know*) but also some very bleached responses (*do*). And *want* has a secondary response that is quite bleached (*have*) but other responses that are closer to its preferential semantics (*need*)—though these responses are not given much more often than representational responses like *know* and *think*, which is interesting.

For the verbs *say*, *need*, and *hear*, the most common response in the *real* conditions is the true verb but not in the *nonce* conditions. This may suggest that much of participants' ability to guess the correct verb in these three cases is somehow dependent on the lexical context. Each has a slightly different response profile in terms of how (intuitively) close the responses are to the true verb's meaning. For instance, like *want*, *need* receives many bleached responses (*have*), but it also receives quite a few responses with preferential semantics (*want*) in both the *real* and *nonce* conditions. Similarly, *hear* receives quite a few bleached response (*do*),

[18]It is unclear whether the use of *care*, the second most common response to *know* sentences, is representational or not.

but also quite a few perceptual responses (*see*). This may suggest that the fact that *hear* is perceptual is encoded in at least some of the items—e.g. in small clause items, like *John heard Mary leave*—but that what kind of perceptual verb it is is not gleanable.[19] *Say* differs from *need* and *hear* in this respect in the sense that, while it has common representational responses (*think* and *know*), it doesn't seem to have any common speech responses, except for maybe *call*. One possibility is that *say*—like many of the speech verbs—has more perceptual correlates than the other representationals, and thus that a syntax-based learner would require more nonlingistic context than for the other representationals.

The final group of verbs—*remember* and *guess* are interesting for the fact that, though their most common responses in either the *nonce* or *real* conditions are technically incorrect, they are both quite close semantically. In the case of *remember*, the vast majority of responses are *know*. This is interesting since *remember* seems to encode *know* as a subpart of its meaning (as mentioned above). Further, the other responses to *remember* are also broadly representational, with both cognitive (*think*) and perceptual (*see*) representationals. *Guess* similarly has a most common response that is representational (*think*) and most of the other responses are also representational (*hope*, *know*, *see*).

One interesting aspect of both of these cases is that the representationals cross not only the cognitive-perceptual divide, but also the factive-nonfactive divide. Factive *remember* gets *think* responses and nonfactive *guess* gets factive responses *know*.

[19]The fact that *see* is a common response to both *see* and *hear* sentences could suggest a frequency effect, which seems likely since *see* is about two orders of magnitude more frequent (as measured by the frequency in ukWaC) in the *present* tense than *hear*. (They are about equally frequent in the past.)

Indeed, this extends to even the previous two groups of verbs, where nonfactive *think*, *say*, and *hear* got *know* responses and factive *know* got *think* responses. One reason this may be is that the purported syntactic cue to factivity—that the factive verb occurs with both polar question and nonquestion complements—cannot be contained within a single subcategorization frame; it is fundamentally an aspect of a verb's distribution.

4.2.1.6 Discussion

In this section, I explored various aspects of participants' responses to the HSP norming task. I showed that accuracy in this task is predicted by both lexical context, corroborating previous findings in this domain, and true word frequency. I then investigated further drivers of this accuracy: participants ability to detect that the target word was a verb and, once participants correctly detect syntactic category, the syntactic features that predict accuracy. I then moved on to a more fine-grained analysis of participants response distributions, and found that many common responses to a particular verb's items tended to be somewhat semantically close to that verb. With the next norming task, I aim to quantify this semantic closeness explicitly.

4.2.2 Similarity norming

In the previous task, I noted that despite the wide variability in item accuracy across verbs, even inaccurate responses are not completely random. Indeed, this

has been noted since the inception of HSP. In this task, I aim to quantify this semantic closeness using the same sort of similarity task employed in Chapter 2. I then conduct analyses of these similarities akin to the accuracy analyses from the last section.

4.2.2.1 Design

All response roots that were (i) marked as verbs in Section 4.2.1.5.2 and (ii) were inaccurate were paired with the true verb that occurred in that position. (This is why only inaccurate pairs were retained. An accurate pair is just the same verb twice.) There were 2429 such pairs. For each of the 10 verbs sampled from the corpus, the pairs involving that word as the true word were then randomized and inserted into lists, with amount of pairs proportional to the number of unique response types to a particular word. With the criterion that each list should contain around 60 pairs, 37 lists were created in this way.

4.2.2.2 Participants

155 participants were recruited through Amazon Mechanical Turk (AMT)—five for 36 of the lists and 10 for the last[20]—using a standard Human Intelligence Task (HIT) template designed for this particular experiment.[21] Prior to viewing the HIT, participants were required to score seven or better on a nine question

[20]The five extra participants were recruited because an off-by-one error that affected only one list was discovered after the first five were run for that list.

[21]A separate experiment script was created in Ibex for each list. The javascript and HTML for this script were then scraped and loaded into an AMT HIT template designed for this task.

qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. Once a participant's submission was received, they were paid $1.

4.2.2.3 Data validation

As in the previous task, a log reaction time-based data validation procedure was conducted. First, each participant's median log RT was computed. The median of these median log RTs as well as the interquartile range (IQR)—the difference between 25^{th} and 75^{th} precentiles)—was then computed. Participants were excluded using Tukey's method applied to both participant medians (described above) and IQRs. 5 participants' median log RTs fell below Q1 log RT minus 1.5 times the IQR and were thus excluded. No participant's IQR fell outside of the Tukey interval of IQRs across participants.

The same RT-based exclusion procedure was also conducted for particular responses. For each participant, the IQR of the log RT for that participant's responses was computed. Responses were then excluded if they fell below that participant's median log RT minus 1.5 times that participant's specific IQR. 18 responses (across participants) were excluded in this way. This yielded a total of 12320 observations with the minimum number of observations per item being 4. (Post-filtering, 7 items had 4 responses and the remaining 1737 had 5 or more.)

4.2.2.4 Results

Prior to analysis, two standardization procedures were applied to similarity judgments: a ridit scoring and a z-scoring. Ridit scoring involves constructing for each participant the empirical cumulative distribution function (CDF) of their responses, then mapping each discrete response level to its corresponding quantile, thus accounting for differences in participants' use of the ordinal scale and forcing the ratings onto the unit interval. Z-scoring involves first mean-centering the ordinal responses (as though they were interval responses) and dividing each by the standard deviation of the responses (again, as though they were interval responses). This ridit scoring is used for the purposes of visualizing the data as well as for later construction of the task this is a norming study for. The z-scored responses are used in the statistical analysis, since they allow for the use of standard linear models, which tend to be much easier to fit.

The average of both the ridit score and z-score transformed variants of the judgments was then taken (separately) and each result associated with each of the true word-response pairs from the previous experiment. Following the lead of the accuracy graph (Figure 4.1) from the last section, the by-item mean of these ridit scored judgments was then taken. Figures 4.4 and 4.5 show the distribution of these by-item similarity means by verb and context of utterance. Figure 4.4 includes accurate response in this mean as 1s and nonverb responses as 0s (neither of which were included in the similarity task). Figure 4.5 shows only the similarity distributions for inaccurate verb responses to each item.

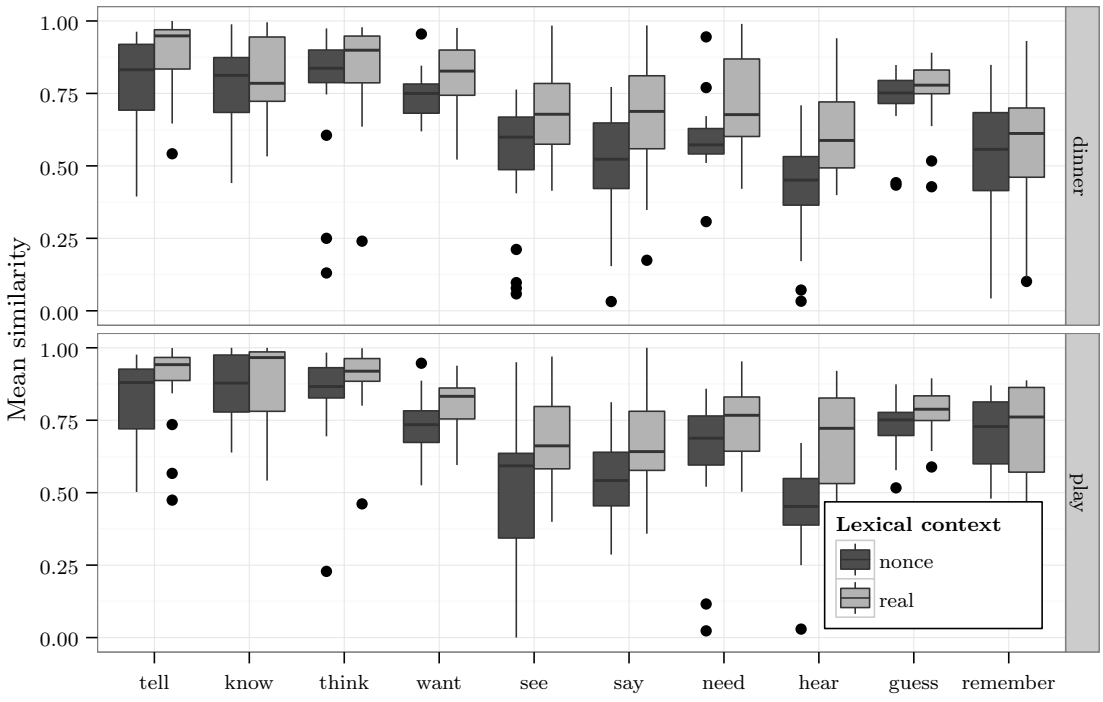


Figure 4.4: Distribution of ridit-scored similarity across items with accurate items set to 1.

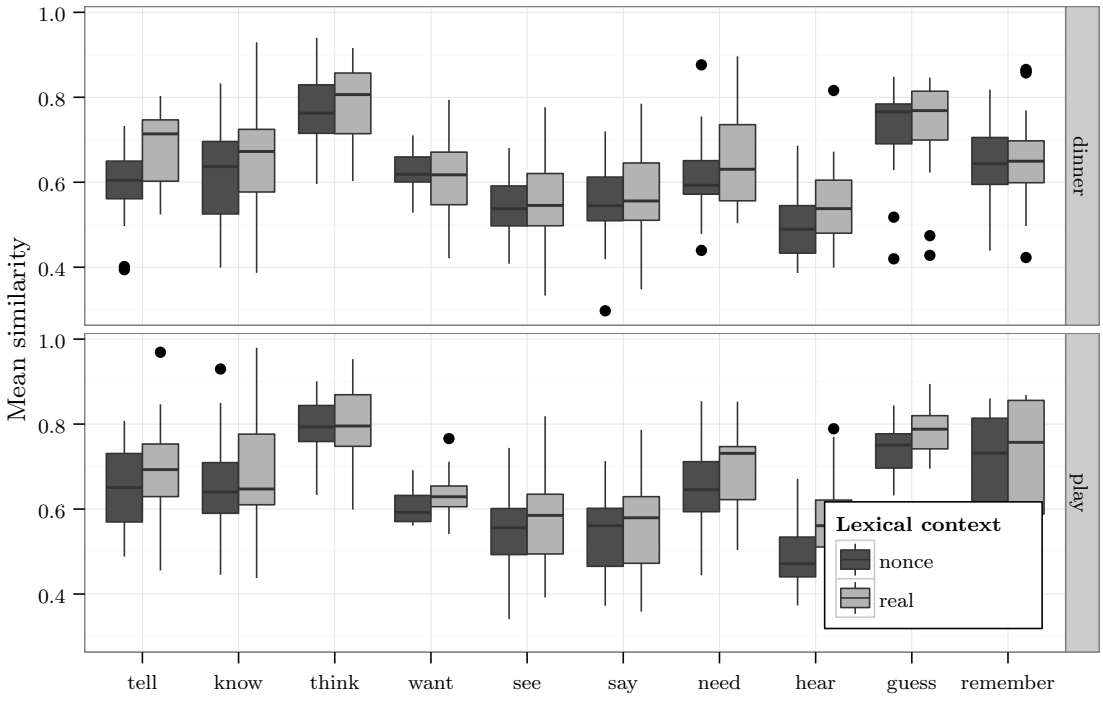


Figure 4.5: Distribution of ridit-scored similarity across items with accurate items set to 1.

4.2.2.4.1 Fixed effects To assess the effects of LEXICAL CONTEXT and CONTEXT OF UTTERANCE on the similarity between an inaccurate verb response and the true word, a linear mixed effects model was fit with mean z-score similarity for each true word-response word as the dependent variable; LEXICAL CONTEXT, CONTEXT OF UTTERANCE, and their interaction as fixed effects; and random intercepts for participant, verb, and item (nested under verb). Thus, this model has the same structure as the original accuracy model, but instead of being fit to accuracy as the dependent variable, it is fit to similarity as indexed by the mean of participants' z-scored similarity responses to a particular response-true word pair—only when the response was inaccurate.[22] And as before, likelihood ratio tests were conducted to assess the significance of particular predictors. As before, the interaction term was not significant ($\chi^2(1) = 0.002, p = 0.968$) and was thus dropped. The two main effect terms LEXICAL CONTEXT ($\chi^2(1) = 14.961, p < 0.001$) and CONTEXT OF UTTERANCE ($\chi^2(1) = 7.807, p < 0.01$) were significant and were thus kept.

Table 4.5 shows the fixed effect estimates for the resulting model. As in the previous section, estimates are given in terms of reference coding with a reference level LEXICAL CONTEXT: *nonce* × CONTEXT OF UTTERANCE: *dinner*. The positive effect of LEXICAL CONTEXT: *real* suggests that participants were able to get closer to the true word's semantics when they had both syntactic and lexical information. This is yet another corroboration of the utility of combining structural and

[22]Indeed, the current model can be thought of as the continuous component of a two-stage zero-inflation model: one that first considers whether the response given by a participant will be a verb or not; then decides whether that response will be accurate; then if inaccurate, decides how similar the response is to the true response. The accuracy and nonverb models from the last section would serve as the first two components.

Table 4.5: Fixed effects for linear mixed effects similarity model.

	Dependent variable:
	INACCURATE RESPONSE SIMILARITY
Intercept	0.445^{***}
	(0.120)
LEXICAL CONTEXT: *real*	0.126^{***}
	(0.032)
CONTEXT OF UTTERANCE: *play*	0.091^{**}
	(0.032)
Observations	17,104
Log Likelihood	−18,301.710
Note:	*p<0.05; **p<0.01; ***p<0.001

lexical information. The positive effect of CONTEXT OF UTTERANCE suggests that participants were able to get closer to the true word's semantics when they were responding to sentences that came from the *play* context.

This latter effect is quite interesting, since in the accuracy model, no significant effect of CONTEXT OF UTTERANCE obtains. This means that though participants are not reliably more likely to respond with the true word based alone on the context in which a sentence was uttered, they do get reliably closer to that word. This may in turn arise due to the same reason that participants are better able to grasp the true syntactic category of the word in the *play* contexts. It further suggests that *dinner* contexts may be a particularly interesting test case, since on the whole they provide somewhat less information per-occurrence about an attitude verbs semantics. I return to this in the discussion.

4.2.2.4.2 Random effects Turning now to the random effects, the variance for the participant random intercepts is 0.012 (sd: 0.111); the variance for the verb random intercepts is 0.136 (sd: 0.369); and the variance for the item random intercepts is 0.174 (sd: 0.417). The size of the participant random intercept standard deviation is about the size of the effect of LEXICAL CONTEXT. This suggests that participants showed variability in their ability to produce high similarity inaccurate verb responses that dwarfs the consistently robust effect of LEXICAL CONTEXT; said another way, it would be unsurprising from the point of view of this model if some participants did about as well with *nonce* contexts as others did with *real* contexts (holding CONTEXT OF UTTERANCE fixed). Given that the fixed effect of CONTEXT OF UTTERANCE is slightly smaller than that of LEXICAL CONTEXT, this also holds, *mutatis mutandis*, for CONTEXT OF UTTERANCE.

As in the accuracy model, verb and item variance far outstrip participant variance with standard deviations $3-4$ times the size of that estimated for participant effects. Looking at the Best Linear Unbiased Predictors for the verb intercepts, this variability appears to affect different classes of verbs differently. Inaccurate verb responses to perception verbs (*hear*, *see*) and (*tell*, *say*) tend to be much less similar to the true verb than those to, e.g., cognitive representations (*remember*, *know*, *think*, *guess*). This seems likely driven by the fact that the perception and speech verbs very often take noun phrase complements, which is likely only a vague cue to their semantics, whereas the cognitive representationals do so less often, tending to take full clausal complements.

Table 4.6: Variable importance as measured by increase in node purity predicting inaccurate verb response similarity judgments

	INCREASE NODE PURITY
VERB	1,499.350
EMBEDDED TENSE	744.547
MAIN OBJECT 1	308.085
COMPLEMENTIZER	209.125
MAIN SUBJECT	149.053
CONTEXT OF UTTERANCE	63.127
LEXICAL CONTEXT	49.653
MAIN PREP	35.782
MAIN OBJECT 2	3.429

4.2.2.4.3 Syntactic features To investigate this conjecture further, I carry out a variable importance analysis similar to the one presented in the last section. There, I entered the hand-coded syntactic features associated with each item into a random forest classifier. In this case, I enter the inaccurate verb response z-score similarities analyzed above along with VERB, LEXICAL CONTEXT, and CONTEXT OF UTTERANCE into a random forest regression with 1000 trees and 3 variables tried at each split. Table 4.6 shows the variable importance measure INCREASE IN NODE PURITY for this model.

Here, again, VERB comes out as an important predictor, as one might expect given the high variability in accuracy across verbs seen in the analysis of the random effects in the last section. Also as in the last section, EMBEDDED TENSE, MAIN OBJECT 1, and COMPLEMENTIZER are important predictors.

4.2.2.5 Discussion

In this section, I presented a norming task aimed at gathering a measure of similarity of participants' responses to the true verb that occurred in a particular item. I then explored various aspects of participants' inaccurate verb responses to the HSP norming task. I showed that accuracy in this task is predicted by both lexical context and context of utterance. I then investigated further drivers of these inaccurate verb response similarities: in particular, the syntactic features that predict similarity. Here, I showed that the same predictors that predict accuracy are also important predictors of similarity.

4.3 Spatial human simulation

In this section, I use the two norming tasks presented above to construct a third experiment aimed at investigating participants' ability to recover the semantics of an attitude verb. This paradigm is close to a standard HSP experiment in the sense that, unlike the first norming task above, participants are told that they were learning the same word over multiple items. It differs, however, in the sense that, instead of giving a free choice response after each item is presented, participants are asked for similarity judgments after the entire set of items.

4.3.1 Design

The task has two main parts: a training phase, in which participants receive a set of sentences containing the same novel word, and a test phase, in which partici-

pants are asked to make ordinal scale similarity judgments. In the first part of the test phases, participants make similarity judgments between the novel word they were just trained on and all real words from the ordinal scale experiment presented in Chapter 2. In the second part of the test phase, participants make similarity judgments between two known words drawn from this same group and selected so as to span the similarity range.

The experiment has four factors: VERB (the same 10 verbs tested in the norming studies), LEXICAL CONTEXT (*real* v. *nonce*), INFORMATIVITY (*high* v. *low*) and TRAINING SIZE (*big* v. *small*). These latter two factors are explained in more detail below; the former two are the same as from the norming studies.

4.3.2 Materials

Training sets were constructed by partitioning the sentences corresponding to each verb from the previous experiment into two sets by-verb. For each of the ten verbs from the norming studies, the median ridit similarity value was obtained by averaging over the values for responses to each item, including accurate responses as 1 and nonverb responses as 0. Items with scores below this median were labeled low informativity (LI) for that verb and items with scores above the median were labeled high informativity (HI). Thus, for each verb at each level of LEXICAL CONTEXT (*real* and *nonce*), there were 10 LI and 10 HI items.

Training sets were then constructed from either solely HI or solely LI items.[23]

[23]This diverges from Medina et al. (2011) in the sense that their training sets involved a mix of the two sets. The reason this was done here was to attempt to draw out the largest possible difference from the sets, which as I show, even this stark partitioning is barely able to do. People

Two of the training sets use all of the items in the informativity partition (TRAINING SIZE: *big*). The other uses only half the items (TRAINING SIZE: *small*). In the LI + small case, the lowest informativity items were used—i.e. those in the first quartile of informativity scores for their particular verb—and in the HI + small case, the highest informativity items were used—i.e. this in the fourth quartile of the informativity scores for their particular verb.

10 different nonce-real test sets were constructed. This test set consisted of 31 pairs: the nonce verb participants were trained on paired with the 31 verbs in the ordinal scale task presented in Chapter 2 and used again in Chapter 3. A real-real test list that remained constant across training sets was also constructed. This list was selected from all pairs in the original ordinal scale task by ordering those pairs based on their mean z-scored rating across participants and then taking every 30^{th} pair. This selection was hand-checked to ensure that a few verbs didn't show up a disproportionate amount of times under this procedure. None did. The reasoning behind this selection procedure was the ensure that the pairs come from across the similarity space, so that any contraction or expansion in the mapping governing participants similarity responses due to the training could be detected.

4.3.3 Participants

2400 participants (515 unique) were recruited through Amazon Mechanical Turk (AMT) using a Human Intelligence Task (HIT) template designed for this

perform extremely well at the task even with LI sets.

particular experiment.[24] Prior to viewing the HIT, participants were required to score seven or better on a nine question qualification test assessing whether they were a native speaker of American English. Along with this qualification test, participants' IP addresses were required to be associated with a location within the United States, and their HIT acceptance rates were required to be 95% or better. Once this submission was received, participants were paid $1.

Participants were allowed to do as many of the lists as they liked, though they were not allowed to do the same list more than once. Lists were deployed in batches of 10, each containing a training set for a particular true verb. This was done to ensure that any participant who did two lists in quick succession would not have gotten two lists pertaining to the same true verb. The median number of lists that each participant did was 1 and the mean was 4.7.

4.3.4 Data validation

Three separate filtering stages were conducted prior to analysis: (i) participant filtering based on memory task accuracy; (ii) participant filtering based on median and IQR of (log) reaction times to the similarity task; and (iii) response filtering based on median and IQR of (log) reaction times by-participant.

For the first stage of filtering a mixed effects logistic regression with random intercepts for participant and verb was built with accuracy on the memory task as the dependent variable and all experimental conditions LEXICAL CONTEXT, ITEM

[24] A separate experiment script was created in Ibex for each list. The javascript and HTML for this script were then scraped and loaded into an AMT HIT template designed for this task.

INFORMATIVITY, and TRAINING SIZE) as well as all possible two- and three-way interactions as fixed effects. The inclusion of these fixed effects was meant to control for the fact that the memory task in certain conditions may be harder—e.g. those conditions with LEXICAL CONTEXT: *nonce*, where participants had to remember nonce words—or less well estimated—e.g. an error in the conditions with TRAINING SIZE:*small* counts more than one in TRAINING SIZE:*big*. This full interaction model was tested against a model without the three-way interaction but with the two-way interactions using a likelihood ratio test, and the three-way interaction was found to be significant ($\chi^2(1) = 17.01$, $p < 0.001$), so the full model was kept.

Table 4.7 shows the fixed effects of this model. We see here that—somewhat unsurprisingly—participants do better at the memory task when the lexical context is real words as opposed to nonce words. More surprisingly, they also appear to do significantly better when the lexical context is both real words and the items are high informativity or the training size is larger. These three positive effects are slightly tamped down by the significant three-way interaction, which is negative and essentially works to cancel the two two-way interactions just mentioned.

Participants were excluded based on the Best Linear Unbiased Predictors (BLUPs) of the participant random intercepts inferred by the model—fit using Restricted Maximum Likelihood (REML) as implemented in the R package `lme4`. These BLUPs for the participant intercepts were then mean-centered and standardized by their standard deviation. All participants whose standardized intercept fell below -2—i.e. two standard deviations below the mean accuracy—were then excluded. This results in the exclusion of 25 total participants, and the loss of 9920

Table 4.7: Fixed effects of mixed effects logistic regression with random intercepts for participant and verb. The reference level is LEXICAL CONTEXT:*nonce* x INFORMATIVITY:*low* x TRAINING SIZE:*small*.

	Dependent variable:
	Accuracy
Intercept	2.094^{***}
	(0.122)
LEXICAL CONTEXT:*real*	0.942^{***}
	(0.158)
INFORMATIVITY:*high*	0.046
	(0.136)
TRAINING SIZE:*big*	0.105
	(0.119)
LEXICAL CONTEXT:*real* x INFORMATIVITY:*high*	0.526^{**}
	(0.249)
LEXICAL CONTEXT:*real* x TRAINING SIZE:*big*	0.522^{**}
	(0.213)
INFORMATIVITY:*high* x TRAINING SIZE:*big*	0.093
	(0.170)
LEXICAL CONTEXT:*real* x INFORMATIVITY:*high* x TRAINING SIZE:*big*	-1.298^{***}
	(0.312)
Observations	15,780
Log Likelihood	−3,852.397
Akaike Inf. Crit.	7,724.793
Bayesian Inf. Crit.	7,801.458
Note:	*p<0.1; **p<0.05; ***p<0.01

total datapoints.

Next, as in previous sections, participants were excluded based on a reaction time analysis (see above for procedure). Only reaction time to the similarity judgment task was considered. In the median-based filtering, 9 participants had a median log RT below the Tukey interval of participant medians and were excluded, resulting in the enclusion of 10240 datapoints. In the IQR-based filtering, 2 participants had IQRs of log RTs outside the Tukey interval, resulting in the exclusion of 310 datapoints.

Finally, as in previous sections, particular responses were excluded based on a reaction time (see above for procedure). 685 observations (across participants) were excluded for a log RT falling outside the Tukey interval for the participant that gave that response.

The final dataset size after this filtering was 125412 observations (60568 nonce-real judgments) with each list having at least 18 response sets (out of the 30 total collected). Note that this results in a number of responses to any particular similarity judgment item that is still more than three times greater than the number of responses even to the original likert scale task, presented in Chapter 2, which only had 5 responses per verb pair.

4.3.5 Results

In this section, I present two types of analysis: one that assesses the overall effects of manipulating the training set size and informativity on participants' ability

to recover the true word from the training items corresponding to that word; and another that assesses what representation participants learned.

To investigate the first, I compare the ordinal scale similarities presented in Chapter 2 to the similarity responses reported by participants in the learning task. As is standard with ordinal scale judgments, responses for both the original task and the current task were z-score-transformed by participant. The mean over these transformed responses was then taken over pairs in the original task, thus yielding a single similarity value for each possible (unordered) pairing.

The similarity values for each of the true verbs for which training sets were built in the current study were then paired with each set of responses given by participant. For instance, for training sets built from sentences that originally contained *think*, the (transformed) similarity judgments given after being trained on those sentences were paired with the (transformed and averaged) similarity judgments from the original experiment for which *think* was one of the verbs in the pair. The idea behind the current analysis is to then assess the correlation between the true word similarity judgments, obtained from the original similarity task carried out in Chapter 2, and what I refer to as the learned word judgments, as obtained from the current task. More specifically, I ask how predictable the learned word similarity judgments are given the true word similarity and the various factors present in the experimental design

To carry this analysis out, I begin with a mixed effects model with fixed effects for TRUE WORD SIMILARITY, LEXICAL CONTEXT, INFORMATIVITY, and TRAINING SIZE as well as all two-way, three-way, and four-way interactions between these vari-

ables; random intercepts for participant and verbs as well as random slopes for TRUE WORD SIMILARITY for both participants and verbs. The fixed effects allow for the assessment of the overall effects of the experimental factors on participants' ability to recapitulate the true verbs similarity ratings from the training set they received, and the random effects allow for the assessment of variability across participants and verbs in accuracy of this recapitulation.

As in previous sections, I carry out a series of likelihood ratio tests to assess which fixed effects should be kept in the model. First, the same model as above was fit without the one four-way interaction and compared to the full model with that interaction. This four-way interaction does not come out significant ($\chi^2(1) = 2.537, p = 0.111$) and was thus dropped. Next, three models excluding all three-way interactions containing a particular categorical variable (LEXICAL CONTEXT, INFORMATIVITY, or TRAINING SIZE). For instance, excluding the three-way interactions involving LEXICAL CONTEXT would entail excluding TRUE WORD SIMILARITY × LEXICAL CONTEXT: *real* × INFORMATIVITY: *high*, TRUE WORD SIMILARITY × LEXICAL CONTEXT: *real* × TRAINING SIZE: *big*, and LEXICAL CONTEXT: *real* × INFORMATIVITY: *high* × TRAINING SIZE: *big*. Under this criterion, the three-way interactions containing INFORMATIVITY ($\chi^2(3) = 25.126, p < 0.001$) and TRAINING-SIZE ($\chi^2(3) = 25.197, p < 0.001$) come out significant, while those containing LEXICAL CONTEXT do not ($\chi^2(3) = 0.632, p = 0.889$). Thus, all three-way interactions containing LEXICAL CONTEXT were dropped, leaving a single three-way interaction: TRUE WORD SIMILARITY × INFORMATIVITY: *high* × TRAINING SIZE: *big*. This interaction was also tested alone and came out significant ($\chi^2(1) = 24.832, p < 0.001$).

The final model's fixed effects can be found in Table 4.8.

4.3.5.1 Fixed effects

I now delve into these fixed effect estimates. These estimates are given in reference coding with the following reference level: LEXICAL CONTEXT:*nonce* × INFORMATIVITY:*low* × TRAINING SIZE:*small*. (As is standard for continuous variables, the intercept represents TRUE WORD SIMILARITY at 0). Coefficients for TRUE WORD SIMILARITY or any of the interactions that contain it represent slopes or changes in slope; all other variables can be thought of as shifts in the intercept for the relevant condition (moving the line up or down along the y-axis wholesale). Since the interesting aspect of these data is the correlation between the true word similarity and the learning word similarity, I focus solely on this former sort of coefficient. (It is unclear to me how one should interpret for the latter with respect to participants' ability to learn a word's semantics from its linguistic contexts.)

The first important coefficient is the one corresponding to TRUE WORD SIMILARITY. This coefficient gives the relationship between the true word similarity and learned word similarity in the reference level (LEXICAL CONTEXT:*nonce* × INFORMATIVITY:*low* × TRAINING SIZE:*small*), which should intuitively be the hardest. Nonetheless, on average, participants recapitulate the similarity judgments of the true word reliably, shown by the positive slope. As one might expect from results in the norming studies, participants do reliably better when given lexical information—even in the low informativity contexts with small training sizes—which can be seen

Table 4.8: Fixed effects of mixed effects regression with random intercepts for participant and verb. The reference level is LEXICAL CONTEXT:*nonce* × INFORMATIVITY:*low* × TRAINING SIZE:*small*. (As is standard for continuous variables, the intercept represents TRUE WORD SIMILARITY: 0)

	Dependent variable:
	LEARNED WORD SIMILARITY
Intercept	-0.095^{**}
	(0.042)
TRUE WORD SIMILARITY	0.375^{***}
	(0.030)
LEXICAL CONTEXT: *real*	-0.096^{***}
	(0.024)
INFORMATIVITY: *high*	-0.019
	(0.023)
TRAINING SIZE: *big*	0.050^{**}
	(0.023)
TRUE WORD SIMILARITY × LEXICAL CONTEXT: *real*	0.068^{***}
	(0.011)
TRUE WORD SIMILARITY × INFORMATIVITY: *high*	0.005
	(0.016)
TRUE WORD SIMILARITY × TRAINING SIZE: *big*	-0.035^{**}
	(0.016)
LEXICAL CONTEXT: *real* × INFORMATIVITY: *high*	-0.026
	(0.027)
LEXICAL CONTEXT: *real* × TRAINING SIZE: *big*	0.028
	(0.027)
INFORMATIVITY: *high* × TRAINING SIZE: *big*	-0.025
	(0.027)
TRUE WORD SIMILARITY × INFORMATIVITY: *high* × TRAINING SIZE: *big*	0.114^{***}
	(0.023)
Observations	60,568
Log Likelihood	$-71{,}436.690$
Note:	*p<0.05; **p<0.01; ***p<0.001

in the reliably positive interaction between TRUE WORD SIMILARITY and LEXICAL CONTEXT.

Interestingly, the other two two-way interactions—TRUE WORD SIMILARITY with INFORMATIVITY and TRAINING SIZE show different trends. As can be seen from the interaction of TRUE WORD SIMILARITY and INFORMATIVITY, participants do no better when given high informativity items with small training sets containing nonce words than if they're given low informativity items. And, as can be seen from the reliable negative iteraction between TRUE WORD SIMILARITY and TRAINING SIZE, if they are given more low informativity nonce items, they actually do worse. The last makes intuitive sense: low informativity items were low informativity because participants didn't get very close to the relevant verb's semantics, compared to other items containing that verb, so giving a participant more of those items would likely only hinder their ability to find the correct semantics. With this said, however, the size of this interaction relative to the main effect of TRUE WORD SIMILARITY is actually quite small (less than 10% of the size), and so participants' inferences actually still look quite robust.

Finally, the positive three-way interaction TRUE WORD SIMILARITY × INFORMATIVITY: *high* × TRAINING SIZE:*big* suggest that getting more contextual information only helps when the instances received are high informativity. This also makes intuitive sense: if one gets a lot of really good information about the semantics of a word, they should do better in recovering those semantics. What's interesting about this effect is its size interpreted in the context of its constitutive two-way interactions. Note that the size of the interaction between TRUE WORD

SIMILARITY and TRAINING SIZE a third and a quarter the size of the three-way interaction, meaning that getting more good items seems to help more than getting more bad items hurts. This is interesting in the current context, since real distributions will include some high informativity and some low informativity items, and so even if a learner were to get a string of low informativity items, a string of high informativity items might well quickly return them to the correct part of the semantic space.

4.3.5.2 Random effects

I now turn to the random effects estimates, which allow for the quantification of uncertainty across verbs and participants. I again focus on the slope estimates—here, the by-verb and by-participant random slopes—since the intercepts don't have clearly interesting interpretation. The variance estimate for the verb slopes was 0.007 (sd: 0.085) and the variance estimate for the participant slopes was 0.033 (sd: 0.182). This state of affairs is the complete flip of that seen in the norming studies where there tended to be low variability due to participants and high variability due to verbs.

For a rough comparison to the fixed effects just mentioned, this variability among verbs might make a randomly selected verb look about as good (for the average participant) in the *lexical context*: *real* × INFORMATIVITY: *low* × TRAINING SIZE:*small* condition as an average verb in the *lexical context*: *real* × INFORMATIVITY: *high* × TRAINING SIZE:*big* condition. That is, this model would not be

very surprised if there are some verbs that could be learned twice as fast as others even with bad data, or conversely, for which it would take twice as long to learn even with good data. In contrast, a randomly selected participant might do just as well (for an average verb) in the *lexical context*: *nonce* × INFORMATIVITY: *low* × TRAINING SIZE:*small* condition as an average participant in the *lexical context*: *real* × INFORMATIVITY: *high* × TRAINING SIZE:*big* condition.

Delving into participant variance within particular conditions, the rough generalization seems to be that participants show lower variability when they are in one of the "extreme" conditions—nonce words with small low informativity training sets or real words with big high informativity training sets—but higher variability in the "middling conditions." One reason for this might be that the extreme conditions give so little or so much information to work with that participants tend to come to the same conclusions about the semantics, but as more or better information comes in, participants have to work harder to incorporate that information into their representation of the word.

4.3.5.3 Relationship to known words

One thing the above analysis does not tell us is what the representations that participants learn might look like. To assess this, I now analyze the actual similarity judgments that participants gave. As an initial stab, the number of times each participant from each condition gave a particular verb the highest rating that participant gave to any response was extracted. The maximum count for these

was then computed within the condition and the verb with this maximum count extracted. For the verbs *want*, *think*, *tell*, and *know*, this verb was the true verb across conditions. For *guess*, the verb was *think* for every condition except for the *lexical context*: *real* × INFORMATIVITY: *high* × TRAINING SIZE:*big* condition. For *say*, the verb was *say* for all but some of the low informativity conditions, where *think*, *want*, and *know* were given. For *remember*, *remember* was given in about half the conditions and *know*, *think*, and *understand* were given in the others. For *hear*, *see* was given in about half the conditions, with *tell*, *think*, *hear*, and *feel* filling in the rest. On the whole then, participants appear to be quite accurate in recovering at the least the correct space of verbs—e.g. representational v. nonrepresentational or even, to some extent, factive v. nonfactive—if not the true verb itself.

This, however, is only a rough measure and throws away quite a bit of information latent in the similarity judgments. To access this information, a more fine-grained analysis is necessary. I give a preliminary one here, but leave a more sophisticated one for future work.

4.4 Discussion

In Chapters 2 and 3, I showed that there is quite fine-grained semantic information present in propositional attitude verbs' syntactic distributions—both their competence distributions (Chapter 2) and their performance distributions (Chapter 3). One question that remains even after showing this is whether it is reasonable to assume that learners have such access to verbs' entire syntactic distribution when

making inferences about those verbs' meanings. The answer to this is almost surely no. Learners receive data incrementally, and it seems likely that the inferences that underlie word-learning also operate incrementally. The question that naturally arises, then, is whether learners can take advantage of attitude verb syntactic distributions in an incremental setting.

In this chapter, I presented three experimental studies aimed at answering this question. The first two experiments, norming studies for the third, were aimed at assessing the informativity of particular sentences in children's input using a variant of the Human Simulation Paradigm (HSP). I did this by extending a previous norming methodology, pioneered by Medina et al. (2011), to allow the similarity between a true word and a response to be quantified.

I then used the results of these norming studies to construct an experiment that manipulated the informativity of items in different training sets given to participants. In this experiment, participants were taught a novel word using these training sets, as in the standard HSP task, but instead of asking participants to make guesses about the word that occurred in each item of the training set, they were asked, after viewing the entire set, to make similarity judgments between the word they just learned and words they already knew. The idea here was that this might allow for the quantification of uncertainty in participants' grasp of the novel words semantic features. Corroborating previous experiments, I showed that participants can utilize syntactic information quite robustly, even using only low informativity learning instances, to learn the meanings of a novel word.

Chapter 5: A strategy for solving the labeling problem

In Chapter 1, I defined two of the main problems of syntactic bootstrapping approaches to verb learning. On the one hand, the syntactic bootstrapping model must define a method by which to discover regularities in verbs' syntactic distributions (the *clustering problem*) and on the other it must have some way of linking these regularities with the facets of meaning they are associated with (the *labeling problem*). One strength of the syntactic bootstrapping approach is that it gives a natural solution to both problems. Verbs are clustered based on how many of the same syntactic contexts they occur in and these clusters are labeled based on rules relating semantic features and syntactic contexts.

One weakness of this approach, at least in its traditional instantiation, is that it is brittle to cross-linguistic variation. This brittleness does not show for many commonly studied classes of verbs—e.g. vanilla transitive verbs like *hit* or ditransitive verbs like *give*—since those classes tend to have fairly cross-linguistically stable syntactic distributions. But as noted briefly in Chapter 1 and as elaborated more fully below, this brittleness is potentially damning for traditional solutions to the labeling problem, since they rely on fixed rules mapping syntax to semantics. This is particularly problematic in the domain of propositional attitude verbs, since

these verbs show much more cross-linguistic variability with respect to the kinds of syntactic contexts they occur in.

In this chapter, I show one way that the traditional approach might be adapted to solve this problem that relies in a crucial way on the properties of the model of syntactic bootstrapping developed in this dissertation. Like traditional approaches to syntactic bootstrapping, the model I propose in Chapters 2 and 3 encodes the abstract notion of a projection rule—a rule that maps from semantic features to syntactic contexts/features. Unlike the traditional syntactic bootstrapping model, however, the rules themselves undergo change; they are inferred at the same time verbs' semantic features are. It is this flexibility in inferring the mapping rules that I seize on to solve the labeling problem for attitude verbs.

To concretize this proposal, I focus on a particular distinction among attitude verbs: the representational-preferential distinction. As I have noted throughout the preceding chapters, this distinction is robustly attested in participants' similarity judgments, both for words they already know (Chapters 2 and 3) and for words they have just learned (Chapter 4). And in English it appears to be robustly tracked by tense, evidenced both in acceptability judgments (Chapter 2) and corpus distributions (Chapter 3). But as I noted briefly in Chapter 1, tense does not robustly track this distinction cross-linguistically. In the next section, I review the syntactic correlates of this distinction in English—suggested in previous literature and corroborated in previous chapters—noting in particular that tense appears to be a robust correlate in English. I then present two problem cases for tense in particular as a cue to the representational-preferential distinction.

Following previous work in this domain, I then suggest that the syntactic correlates of the representational-preferential distinction in each of the three cases discussed can be given an abstract characterization in the following terms: representational verbs, like *think* and *know*, take subordinate clauses that are more closely matched to main clauses than those taken by (pure) preferential verbs, like *want* and *prefer*. This relativizes the mapping between the representationality feature and syntactic context to a language while retaining the abstract notion of projection. This, in turn, makes it possible to construct such a learner within the model I have been developing, which I give a preliminary sketch of in English. I then turn to a small experiment in implementing this proposal before concluding with some future directions.

5.1 The representational-preferential distinction

5.1.1 Representationals and preferentials in English

I now review the representational-preferential distinction discussed in Chapter 1. This distinction is one among propositional attitude verbs is that between verbs that express beliefs—or represent "mental pictures" or "judgments of truth" (Bolinger, 1968)—and those that express desires—or more generally, orderings on states of affairs induced by, e.g. commands, laws, preferences, etc. (Bolinger, 1968; Stalnaker, 1984; Farkas, 1985; Heim, 1992; Villalta, 2000, 2008; Anand and Hacquard, 2013, a.o.). Within the first class—the representationals—fall verbs like *think* and *know*—and within the second class—the preferentials—fall verbs like *want* and

order.

There appear to be various aspects of the syntactic distribution that roughly track this distinction in English. One well-known case—the parade case—is finiteness: representationals tend to allow finite subordinate clauses (1a) but not nonfinite ones (1b) while preferentials tend to allow nonfinite subordinate clauses (2b) but not finite ones (2a).

(1) a. John thinks that Mary went to the store.

b. *John thinks Mary to go to the store.

(2) a. *John wants that Mary went to the store.

b. John wants Mary to go to the store.

There are two important things to note about this distinction. First, though the representationality distinction is often talked about as though it were mutually exclusive, some verbs appear to fall into both categories. For instance, as noted in the last section, *hope p* involves both a desire that *p* come about and the belief that *p* is possible (Portner, 1992; Scheffler, 2009; Anand and Hacquard, 2013, but see also Portner and Rubinstein 2013). Ideally, then, a model of syntactic bootstrapping would discover that *hope* has both a representational and a preferential semantics. Such a discovery seems plausibly since *hope* shows up in both finite (3a) and nonfinite (3b) frames.

(3) a. John hopes that Mary went to the store.

b. John hopes to go to the store.

Second, the link between representationality and finiteness is just a tendency. Some verbs plausibly classed as representationals allow nonfinite subordinate clauses (17a)/(17b), and others plausibly classed as preferentials allow subordinate clauses that look finite (17c). In spite of this, as I show in Chapters 2 through 4, finiteness is a useful cue in distinguishing representationals and preferentials.

5.1.2 Representationals and preferentials outside English

The roughness of this correlation is perhaps not surprising since not all languages track representationality with tense. I focus on two cases of this: ones where the distinction is roughly tracked by mood—in the Romance languages, representationals tend to take indicative mood and preferentials tend to take subjunctive mood (Bolinger, 1968; Hooper, 1975; Farkas, 1985; Portner, 1992; Giorgi and Pianesi, 1997; Giannakidou, 1997; Quer, 1998; Villalta, 2000, 2008, a.o.)—and others where the distinction is tracked by the availability of verb second (V2) syntax (Truckenbrodt, 2006; Scheffler, 2009).

An instance of the correlation with mood can be seen in Spanish. In Spanish both the representational (belief) verb *creer* (*think*/*believe*) and the preferential (desire) verb *querer* (*want*) take finite subordinate clauses. The difference between these subordinate clauses is that, whereas verbs like *creer* (*think*) take subordinate clauses with verbs inflected for indicative mood (4a), verbs like *querer* (*want*) take subordinate clauses with verbs inflected for subjunctive mood (4b).

(4) a. Creo que Peter va a la casa.
think.1S.PRES that Peter go.PRES.IND to the house.

b. Quiero que Peter vaya a la casa.
want.1S.PRES that Peter go.PRES.SBJ to the house.

This makes the subordinate clause under *creer* look more like the declarative main clause in Spanish, whose tensed verb is inflected for indicative mood.

(5) Peter va a la casa.

Peter go.PRES.IND to the house.

An instance of the correlation with V2 can be seen in German and other Germanic languages—e.g. Dutch. V2, which is generally found in main clauses, is a phenomenon in which a clause's tensed verb appears as the second word in a sentence. For instance, (6) shows a German main clause with the tensed form of the auxiliary verb *sein* (*be*) occurring as the second word of the sentence (in second position).

(6) Peter ist nach Hausen gegangen
Peter is to home gone

In subordinate clauses headed by the complementizer *dass* (*that*), this verb occurs clause-finally, which evidences the fact that German is underlyingly a subject-object-verb (SOV) language. Both the verb *glauben* (*think*) and the verb *wollen* (*want*) can take such clauses, in which the main verb is tensed.

(7) a. Ich glaube, dass Peter nach Hausen gegangen ist.
I think that Peter to home gone is.

b. Ich will, dass Peter nach Hausen geht.
I want that Peter to home goes.

Only *glauben* (*think*), however, allows a second sort of structure more akin to the main clause in the position of the tensed verb (Scheffler, 2009). If the complementizer

dass (*that)* is not present, *glauben* (*think*) can take a subordinate clause with syntax that looks exactly like that of the main clause—compare the main clause in (6) with the subordinate clause in (8a). *Wollen* does not allow this (8b).

(8) a. Ich glaube, Peter ist nach Hausen gegangen.
I think Peter is to home gone.

b. *Ich will, Peter geht nach Hausen.
I want Peter goes to home.

Thus, though both Spanish and German take tensed complements, militating against a hard-coded link between tense and representationality, they still show language-internal correlations between representationality and some more abstract aspect of the clausal syntax. Further, the aspect of the clausal syntax that occurs with only the representational verbs—indicative mood in Spanish and V2 in German—also tends to show up in declarative main clauses.

5.1.3 Main clause syntax

This apparent language-internal correlation has led some authors to conclude that, rather than there being a relationship directly between representationality and tense, as is evidenced in English, the relationship needs to be specified more abstractly. One idea is that this more abstract mapping between semantics and syntax should be specified in terms of *main clause syntax* (Dayal and Grimshaw, 2009; Hacquard, 2014).

Under this view, then, the apparent relationship between tense in English, mood in Spanish (and the rest of Romance), and V2 in German (and other Germanic

languages besides English) is really the outgrowth of a more abstract relationship between some cluster of syntactic features—call them MAIN CLAUSE features—that are language-specific but likely highly constrained. The way in which they are constrained is that they tend to be associated with properties of the subordinate clause's that are "close" to the attitude verb. For instance, both complementizers and mood tend to be assumed to be quite high within the clausal structure (cf. Cinque, 1999; Speas, 2004), which in turn seems to make them amenable to selection by particular semantic classes of verbs—e.g. representationals or preferentials. Indeed, ideally, one could pin the relevant feature to some particular type of head which carries the relevant selection information—e.g. the complementizer— and is "as high as possible" within the subordinate clause so as to make selection maximally local.

Suggestive of this possibility is that the standard analysis of German V2, which has that V2 is a particular kind of complementizer-driven movement akin to that seen in English WH-movement (Den Besten, 1983). English may be amenable to such an analysis in the sense that complementizer drop with finite subordinate clauses tends to only occur with representationals (Dayal and Grimshaw, 2009), as discussed in Chapter 1.

(9) a. Bo {thinks, believes, knows} (that) Jo is out of town.

b. Bo {loves, hates} *(that) Jo is out of town.

This latter fact is furthermore suggestive, since of course English main clauses do not have complementizers, bolstering the relationship between main clause syntax and representationality, at least in English. This, however, also raises a potential

problem for languages like Spanish, which lack complementizer drop in any subordinate clauses but whose declarative main clauses do not have complementizer. I return to this in this chapter's discussion.

But regardless of whether main clause syntax information can be carried solely in the complementizers themselves—thus allowing for an extremely local form of selection giving rise to the relationship between representationality and main clause syntax—or whether somewhat longer distance relationships need to be posited, there is nonetheless a potential relationship between the representational-preferential distinction and this language-specific-yet-highly-constrained MAIN CLAUSE feature.

The importance of this for current purposes is that, if such a correlation between representational and main clause syntax exists, it may signal a possible candidate for a hard-coded-yet-flexible projection rule that allows for a solution to the labeling problem in this particular case. Further, since the main clause syntax itself is presumably observable to the same extent that subordinate clause syntax is, the language specific instantiation of the MAIN CLAUSE feature may well itself be learnable. And if this can be made to work in this particular case, one might seek further cases where, though a particular mapping between semantics and syntax appears unstable cross-linguistically, there is nonetheless a more abstract feature that correlates with said mapping and which itself might be learner from some observable features of the input.

In the next section, I show how this insight about the correlation between main clause syntax and representationality might be incorporated into the model of syntactic bootstrapping developed throughout the dissertation to solve one piece of

the labeling problem for propositional attitude verbs.

5.2 Leveraging main clause syntax

In this section, I show how one might incorporate the abstract relationship between main clause syntax and representationality into the model of syntactic bootstrapping I develop throughout the dissertation. The essential idea is that the learner should construct a particular projection rule or set of rules over the course of learning, which—unlike the other rules they construct—is directly linked to a particular semantic feature—in this case, representationality. The model as it currently stands only has a way of constructing sets of rules that are unlabeled, so what needs to be added is some way of singling out a rule or set of rules that project onto the main clause syntax features. The problem is that what these main clause syntax features are must themselves be learned. Luckily, however, these features should be quite easily learnable; they are just the ones that are seen every time a declarative main clause is seen.

This suggests a quite simple addition to the current model of syntactic bootstrapping—one that requires only a minor change to the structure or algorithm that the model employs. This solution is to add declarative main clauses to the data set as though they were subordinate clauses taken by a particular attitude verb that is never heard,[1] and then force the model to explain this verb's distribution using only a single feature. This in turn means that the model has to have at least one feature

[1]It is sufficient to use this only to implement the proposal here. I remain agnostic about whether this is the correct syntactic or semantic analysis of any particular sentence.

that projects onto the main clause features.

The idea that main clauses are in fact subordinate clauses to a particular kind of verb (or set of verbs) is an old idea instantiated most famously by Ross's (1970) Performative Hypothesis (see also Rizzi, 1997; Ambar, 1999; Krifka, 2001; Ginzburg and Sag, 2001; Speas and Tenny, 2004; Hacquard, 2010). I follow Hacquard (2010), and others, in calling this special element ASSERT.

Why should this minor addition of a special verb ASSERT along with a rule that labels ASSERT's feature as representational work to solve the labeling problem? The intuition here is that (i) the syntactic contexts that ASSERT occurs in are extremely constrained—to one context: a finite clause with no complementizer—and (ii) ASSERT is extremely frequent—every declarative sentence counts as evidence for the distribution of ASSERT. The second property makes it expensive—in terms of likelihood—for the model to ignore ASSERT. This means that the model should ensure that ASSERT's distribution matches up with the features that the model posits and the projection rules for those features. The first property—in concert with the second—will ensure that at least one feature projects onto main clause syntax. This feature should presumably be the representational feature associated with ASSERT and hopefully other verbs. In the next section, I implement this proposal and give some preliminary results.

5.3 Experiment

In this section, I fit the nonnegative projection model proposed in Chapter 3 to the modified dataset suggested in the last section. I show that this model discovers a small core of high frequency representational verbs that share a feature with ASSERT.

5.3.1 Data

The dataset used here is the same one used in the experiment in Chapter 3, which was extracted from the PukWaC corpus (see that chapter for dataset construction).[2] To this dataset was added approximately 3.5 million observations of main clauses, represented as a subordinate clause embedded under a special verb ASSERT.

Main clauses were identified within the corpus by checking that a particular verb was a dependent of a ROOT node in the dependency parse. Some clauses identified as main clauses were misparses, e.g., of constituents like purpose clauses. These were filtered out by only allowing main clauses that (i) had a subject in the dependent parse; (ii) were tensed in the dependency parse; and (iii) had no complementizer. (This last criterion excludes question main clauses, but, though these may be useful to include in future experiments.)

Finally, all main subject values for ASSERT were set to *referential* (see Chapter

[2]This corpus is not necessarily ideal for testing a learning model, but due to the fact that annotations in CHILDES—the standard collection of child-directed speech corpora in English—are extremely noisy, similar automatic extraction of subcategorization frames is difficult.

3 for description). This was done to mimic the fact that performatives are claimed to involve covert first person subjects. A version of this dataset was also constructed in which the subject features was not included in the construction of the frames, but this made no discernible difference on the results.

5.3.2 Model fitting

The model fitting procedure was the same given in Chapter 3 except for two things. First, the number of features was set at 2. The idea here is to force the model to make a choice of either giving a verb the same feature as ASSERT or not. (The other feature will, in essence, be a "waste bin" feature, collapsing all other semantic features besides representationality into one.) As in Chapter 3, the model fitting was restarted multiple times with random initializations to ensure that a high likelihood point was discovered.

Second, as mentioned in the previous section, ASSERT was only allowed one feature which remained constant across the model fitting. That is, the model had only one feature with which to explain the syntactic contexts that ASSERT occurs with. Thus, this feature will have an associated projection rule that picks out at least the main clause syntax.

5.4 Results

Of the 232 verbs plus ASSERT, 10 verbs share a feature with ASSERT: *ask*, *consider*, *find get*, *know*, *say*, *see*, *show*, *tell*, and *think*. This list is interesting because

it includes verbs from across the range of representationals. *Ask*, *tell*, *say*, and *show* involve communication, while *think*, *know*, *see*, *find*, and *get* involve cognition and perception. Furthermore, no preferentials—besides perhaps *ask*, *tell*, and *say*, at least with some frames—are represented in this list.

One interesting thing about this list is the prevalence of question-taking verbs. As noted above, all of the question main clauses were removed from the dataset, and so it's interesting that these verbs are included. This is especially surprising, since for none of these verbs does the model posit a second feature. One way this may have happened is through a process of generalization for the projection rule associated with the feature associated with ASSERT. For instance, *think*, *say*, and *know* both take subordinate clauses that look like main clauses ("subordinate clauses" of ASSERT) with high frequency. But *know* takes question complements with fairly high frequency as well—as does *say* in certain circumstances—and so the model may have adjusted the projection rule to include some weights on question complements. This may in turn heightened the likelihood that verbs like *ask* would find their way in.

Another interesting thing about this list is the notable absence of many representational verbs, such as *understand*, *realize*, *suppose*, *point out*, etc. These verbs for some reason end up in the "waste bin" category. One possibility for why this occurs in that these verbs are slightly lower frequency than the ones that make it into the above list and are plausibly acquired later. This may in turn have given the fitting procedure less impetus to associate them with the more targeted feature associated with ASSERT.

5.5 Discussion

In this chapter, I reviewed the labeling problem for syntactic bootstrapping. I noted that the standard approach to this problem—reliance on hard-coded links between particular projection rules and particular semantic features—runs into problems with cross-linguistic variation in the mappings from semantics to syntax. If languages vary with respect to how they map semantic features into the syntax, then those mappings seemingly couldn't be hard-coded. I showed that this was particularly pernicious within the domain of propositional attitude verbs, since even the distinctions that appeared most robustly in participants' semantic similarity judgments, show little cross-linguistic stability in their mappings to the syntax—at least on the face of it.

I then turned to a discussion of what these mappings look like within particular languages. Following recent work, I noted that though the particular syntactic syntactic features a semantic distinction like representationality maps to differ across languages, there appears to be a family resemblance between these cross-linguistically active syntactic features. The particular family resemblance relevant to representationality appears to be whether or not a verb takes main clause syntax.

In the latter part of the chapter, I then showed how this family resemblance might be incorporated into a syntactic bootstrapping learner of the kind proposed throughout the dissertation. I showed in a preliminary experiment that, when implemented this sort of learner shows promising results, though there is much more work yet to be done on this problem.

There are two particular directions that seem likely to be fruitful. First, these sorts of models could—indeed, should—be deployed on languages other than English to truly test their robustness to different sorts of input conditions. For instance, would the sort of model developed here be able to detect that subjunctive rather than tense is the property important to the representational-preferential distinction in Spanish? Would it similarly be able to detect that V2 is relevant in German? Second, since the model was fit to a dataset that likely does not reflect the child's input, a dataset derived from a corpus of child-directed speech is desirable.

Chapter 6: Conclusion

I began this dissertation by laying out the central problems of learning what Gleitman et al. (2005) dub the *hard words*, focusing in particular on the *propositional attitude verbs* like *think*, *know*, and *want*. I noted two main problems for learning these verbs: (i) the eventualities they describe tend not to have sensory correlates, and (ii) their meanings are both fine-grained and multi-faceted, thus presenting problems for accounts based on learning from nonlinguistic context (or even discourse context) alone.

I then turned to a discussion of learning from linguistic context, noting two particular kinds of linguistic contexts that have been discussed as possible learning cues: lexical context and syntactic context. I noted that, while lexical context is likely useful for certain distinction among verbs—indeed, it may be useful even for some distinctions among propositional attitude verbs—it likely does not track other distinctions of central interest. This led me to turn to the use syntactic context as a word-learning cue—a strategy exemplified most notably in *syntactic bootstrapping* approaches to word learning.

I noted two problems that any syntactic bootstrapping approach must solve: (i) it must explain how learners cluster verbs based on the syntactic contexts they

occur with—the *clustering problem*—and (ii) it must explain how learners label these clusters with the facets of meaning they correspond to—the *labeling problem*. The ability of a syntactic bootstrapping account to solve either of these problems for any particular type of verb is dependent on the (i) the granularity with which that particular verb type's semantics is mirrored by the syntactic distribution and (ii) the availability of principles that would allow a learner to label the semantic features. I raised doubts about this second prospect having to do with the cross-linguistic stability of the mapping principles, particularly in the attitude domain, arguing that the labeling problem quite plausibly could be solved via other means, and so the first problem should be attacked first in isolation.

I then turned to an overview of what is known about this relationship in the domain of propositional attitude verb. I showed that the results are quite promising but also that the correlations are not perfect. This raises the need for a more fine-grained investigation of these correlations, which I carried out.

In **Chapter 2**, I began the investigation by showing how to quantify the relationship between naïve speakers' knowledge of the syntactic contexts a propositional attitude verb can occur in—what I refer to as the *competence distribution*—and their knowledge of that verb's semantics. To do this, I deployed a methodology that Fisher et al. (1991) used to probe such relationships as they obtain for verbs across the lexicon, here focusing in on the propositional attitude verb domain in order to test the limits of this relationship. The main result of this chapter was that there is a significant correlation between the syntax measure and the semantics measure. This omnibus result, however, tells us little about the relationship between partic-

ular syntactic contexts and particular facets or features of the meaning. To delve into this, I developed a model, which I dubbed the *nonnegative model of projection*, to investigate this relationship. The benefit of this model is that it furthermore implements part of a solution to the clustering problem. I showed that this model discovers the sorts of fine-grained features discussed in Chapter 1.

In **Chapter 3**, I investigated to what extent the same sort of relationship found between verbs' competence distributions and their semantics also obtains between the distribution of syntactic contexts a propositional attitude verb occurs in in a corpus, what I refer to as its *performance distribution*, and participants' knowledge of those same verb's semantics. To do this, I developed a model that augments the nonnegative model of projection presented in the previous chapter with a model of corpus count data. This model simultaneously discovers competence distributions using the corpus distributions, while at the same time solving the clustering problem. The main result of this chapter is that performance distributions also carry a significant amount of information about propositional attitude verb semantics and that this information is comparable with that found in the direct measures of competence distribution employed in Chapter 2.

In **Chapter 4**, I investigated whether the information in performance distributions is in fact accessible to learners and how robustly represented this information is. To do this, I adapted recently developed methodologies related to the Human Simulation Paradigm (HSP) to (i) measure the informativity of particular items in the performance distribution about the semantics of the word that occurs in them and (ii) measure the informativity of the distribution itself. The main result of this

chapter is that, even if items are manipulated in such a way to give participants as little information as possible, inference to all propositional attitude verbs meanings are extremely robust, even down to extremely fine-grained facets of those verbs' meanings.

In **Chapter 5**, having focused for the majority of the dissertation on solving the clustering problem, I presented a novel proposal for how to approach the labeling problem. This proposal starts with the observations that, particularly in the propositional attitude verb domain, the relationship between particular aspects of the semantics and particular syntactic contexts seems to be cross-linguistically unstable. This does not raise problems for the model presented in previous section necessarily, since as long as those languages exhibit roughly the same patterns of correlations between meaning and syntactic context, this model should similarly succeed in solving the clustering problem. The problem arises if labels are somehow associated *a priori* with particular syntactic contexts—for instance, if tense were somehow associated with the representationality distinction—since not all languages show this correlation. The proposal presented in this chapter was that, while not all languages associate particular facets of the semantics with particular syntactic contexts, at least some particular facets may be associated with families of syntactic contexts and that the learner's job is to select the appropriate syntactic context to associate with that facet using the data. I then show how this might be encoded in a model like the one I develop in the previous chapters.

I now conclude the dissertation with some further directions for this work.

6.1 Future directions

6.1.1 Quantifying meanings

In Chapters 2 and 3, I validate the nonnegative model of projection and the nonnegative model of syntactic bootstrapping against semantic similarity judgments. Semantic similarity judgments are useful for validation in that many disparate models of semantics can be tested against the same dataset. This is because most computational models of semantics—be they category-based, vector-based, ontology-based, etc.—provide some way(s) of measuring the distance between two meanings (or at least the divergence of one meaning from the other).

This generality presents a problem, however, in assessing what these similarity judgments are actually indexing, since they are designed to some extent to provide omnibus measures of the semantics. To be sure, such omnibus measures seem to be differentially sensitive to certain semantic features, as I showed in Chapter 2; but it is very hard to tell *a priori* which features a particular methodology will be sensitive to.

As such, one potential future direction is to utilize the method presented in Chapter 2 of comparing the outcomes of two similarity tasks against each other to pinpoint exactly where they disagree. This line runs two risks, however. First, it is possible that particular methods tap necessarily vague concepts. In Chapter 2, the difference between the two methods lie in how participants responded to some vague notion of antonym, where many semantic distinctions that are different in

principle are collapsed into one.[1] This might be so no matter how many words are tested. Second, and relatedly, one risks studying the particular methods themselves as opposed to the underlying space they are meant to tap. That is, it is always possible that with so few verbs, the apparent differences between the methods had little to do with the underlying space itself, but were artifacts of some other process involved in making the sorts of decisions the task demands.

Another potential direction in the vein of methods for quantifying meaning are tasks aimed more explicitly at particular features. These tasks come in roughly two flavors: those that explicitly ask about semantic features of the verb (Hartshorne et al., 2013) and those that explicitly ask about semantic properties of a verb's arguments (Kako, 2006).[2] These sorts of methods are useful since, as long as they are well normed, they give much more direct access to particular features, thus bypassing a step of inducing features from similarities.

The main problem with these latter sorts of methods is that they may not exhaust the space of features, where semantic similarity judgments might be more successful. Thus, these methods, like the traditional methods employed by linguists fall prey to the criticism that our discoveries are limited by the space of features that readily come to the mind of the investigator. But this was just the problem that quantitative assessment was intended to solve (Fisher et al., 1991).

A potential remedy here is to combine these more explicit methods with se-

[1]The hope in this case would be that, even though the notion of antonymy in the "residual space" of the two methods is vague, other aspects of the judgments in each method would help to parcel different sorts of antonymy out based on other aspects of a word's content.

[2]The distinction between these two kinds of properties is sometimes difficult or impossible to discern. The fact that *break* involves a change of state implies that one of its arguments undergoes a change of state.

mantic similarity-based methods to assess to what extent the more explicit methods exhaust the information in the similarities—e.g. by predicting the similarities from the explicit feature judgments. One way of generating new explicit feature questions might then be to take similarity values that are not well predicted, find clusters within those badly predicted values, ask annotators what commonality that group has, and then construct a question based on that commonality.

A final potential direction is to validate directly onto psycholinguistic data.

6.1.2 Mapping from syntax to meaning

As discussed throughout the dissertation, the main job of a syntactic bootstrapping mechanism is to map from a word's syntactic distribution to the concept associated with that word. In Chapter 2, I showed that, though a learner can conclude by similarity in distribution that there is likely a similarity in meaning but not that, if there is a dissimilarity in distribution, there is dissimilarity in meaning. I then incorporated this idea into the syntactic bootstrapping model in Chapter 3 by using the particular combination of prior and likelihood I did.

One question for future research is why this property should exist. Why shouldn't being different in distribution also implied difference in meaning? One intuition is that, in generalizing about two words based on their distribution, it is easier to ask which frames they share than which frames they don't share. But this intuition is vague and possibly wrong, since of course linguists do such comparisons on a conscience level all the time. What lower-level aspect of cognition—perhaps

specifically linguistic cognition—gives rise to this state of affairs?

A second question in this same vein is what, if anything, the particular functional relationship between similarity in distribution space and similarity in semantic similarity space means about cognition. In Chapter 2, I showed that similarities gathered in the generalized discrimination task appeared to be logarithmically related to the syntactic similarities, whereas those gathered in the ordinal scale task appeared to be sigmoidal.[3] Are these logarithmic relationships fundamental to the syntactic bootstrapping mechanism? Or are they merely artifacts of the similarity tasks?

Finally, and relatedly, how does the learner, whose whole job is to find the correct mapping from words to concepts use distributional similarity to construct the mapping? One potential future direction aimed at investigating this question is to construct a model which has access to the concepts—e.g. proxied by similarity judgments—and the distributions, but does not have the mapping from verbs to concepts. This model would then need to learn this mapping by discovering a permutation (of indices) that is optimal under some loss. Relevant to the previous discussion in this section, this loss should encode that similarity in distribution should match similarity in meaning. How divergence in such similarity are penalized is relevant to the question of what status (if any) the exponential properties mentioned above have in the learning mechanism.

[3]The best fitting model for the generalized discrimination task was the diffusion kernel/exponential model, whereas the best fitting one for the ordinal scale task was the linear model. (The sigmoid would arise from the latter in due to the presence of a logistic function implicit in that mapping.)

6.1.3 Finer-grained incremental conjectures

In Chapter 4, I defined a way of looking at fine-grained aspects of participants' final conjectures about a word's meaning after different amounts of training. This provides a view of their incremental conjecures with a grain-size proportional to the number of different training set sizes—in that chapter five or ten sentences. Ideally, however, one would have such fine-grained resolution about the participant's conjecture after each sentence in the training set—not just at the end—and it would be prohibitively expensive to test all possible set sizes in a systematic way.

One way this might be remedied is to instead run the spatial human simulation experiments from Chapter 4 to get a series of fine-grained final conjectures, and then attempt to "backtrack through" the conjectures after each training item. This might be done in two complementary ways.

The first is a modeling approach. If a model is set up that assumes (i) a unique starting point for conjectures and (ii) that each item draws a participant's conjecture toward a particular point in similarity space more or less strongly (depending on its informativity), then assuming full randomization of the training sets, a rough idea of the conjecture path over the course of the training set might be attainable.

In concert with this modeling approach, one might also collect data from a standard human simulation paradigm run on the same training sets. Though the standard paradigm gives less fine-grained results than the similarity judgments gathered after each training set, it still gives an indication of nearby words in similarity space. With this in mind, the backtracking model sketched above might be aug-

mented to incorporate data from a standard human simulation paradigm.

6.1.4 Main clause syntax and beyond

Chapter 5 gives a preliminary sketch of a model that takes advantage of main clause syntax features to label verbs as either representational or preferential. There are multiple ways this sketch could be expanded.

The first involves the corpus used in training the model. As an initial stab, the use of child-directed speech corpora is crucial for testing the efficacy of this approach. The current problem with such corpora is the poor state of the parses associated with their sentences. Good parses are crucial for extracting good subcategorization frames, so either producing better parses or working around the current parses in an innovative way is necessary. Second, and perhaps more importantly, corpora from other languages must be tested. A similar hurdle arises for such corpora—particularly child-directed speech corpora.

Beyond logistical questions are questions regarding the model itself. In particular, how quickly does labeling happen if this model is converted into an online version? And what other sorts of labels might be learned in a similar fashion? For instance, could different sorts of non-main clause features be helpful in labeling distinctions among preferentials?

Appendix A: Appendix A

A.1 Figures and tables

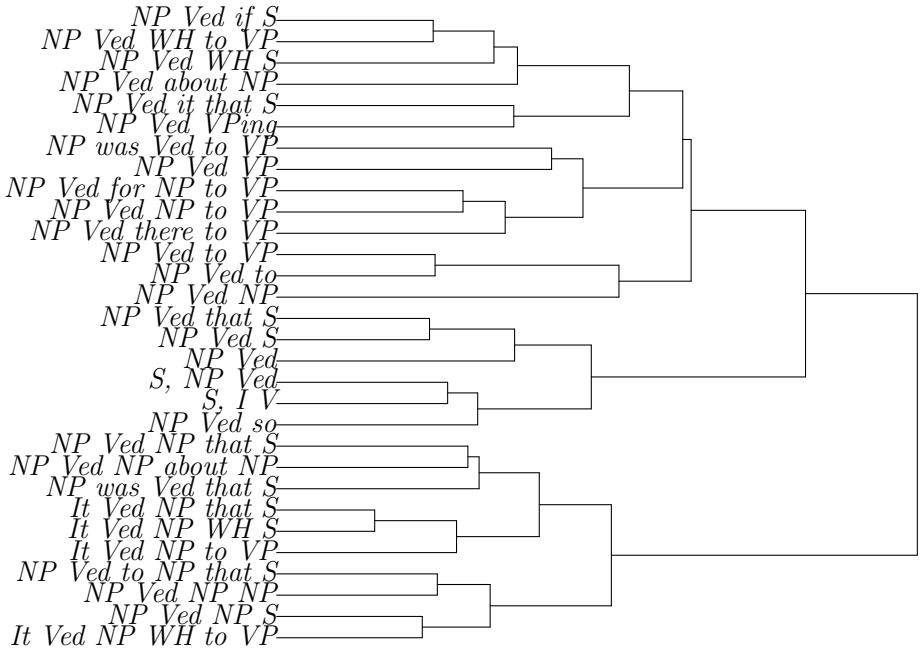


Figure A.1: Hierarchical clustering of frames based on data in Figure 2.2.

verb 1	verb 2	Standardized residual
want	love	3.39
want	imagine	3.32
remember	see	3.32
think	promise	3.15
feel	doubt	3.15
want	think	2.99
think	deny	2.91
suppose	hate	2.79
think	hope	2.77
doubt	amaze	2.73
see	imagine	2.65
understand	deny	2.60
think	expect	2.59

Table A.1: Pairs rated more highly in the likert scale task than in the generalized discrimination task.

A.2 Non-negative projection model

A.2.1 Parametric binary feature model

$$\pi_j \mid \alpha \sim \text{Beta}(\alpha, 1)$$

$$r_{ij} \mid \pi \sim \text{Bernoulli}(\pi_j)$$

A.2.2 Nonparametric binary feature model

In footnote 30 in Section 2.2, I note that a nonparametric version of the non-negative projection model was implemented. To do this, I use an Indian Buffet Process prior on the verb binary feature space (Griffiths and Ghahramani, 2006). This allows us to simultaneously infer the number of features at the same time as we

infer what those features are. For ease of implementation, a stick-breaking process representation was used (Teh et al., 2007). A truncation of 100 features appears to be more than sufficient.

$$\pi \mid \alpha \sim \text{IBPStick}(\alpha)$$

$$r_{ij} \mid \pi \sim \text{Bernoulli}(\pi_j)$$

A.2.3 Projection principles (feature loading) model

$$p_{jk} \sim \text{Exponential}(1)$$

A.3 Response models

A.3.1 Ordinal logit mixed model

Because both the non-negative projection model $\hat{\mathbf{D}} = \mathbf{ZB}$ (Section 2.2.6) and the similarity kernels K (Section 2.4.2) both have strictly non-negative codomains, an ordinal logit mixed model with strictly positive cutpoints was used. The model for all participants i, for all but one likert scale response level j, for all verbs m, n is then

$$\lambda_i \sim \text{Gamma}(1,1) = \text{Exponential}(1)$$

$$c_{i1} \mid \lambda \sim \text{Exponential}(\lambda_i)$$

$$c_{ij} \mid \lambda, \mathbf{C}^{\top}_{1:j-1} \sim c_{i(j-1)} + \text{Exponential}(\lambda_i)$$

$$l_{imn} \mid \mathbf{Q} \sim \text{Multinomial}(q_{i1mn}, q_{i2mn} - q_{i1mn}, \ldots)$$

where $\mathbf{Q}$ is defined by

$$q^{\text{acceptability}}_{ijmn} \equiv \text{logit}^{-1}(c_{ij} - r_{mn})$$

$$q^{\text{similarity}}_{ijmn} \equiv \text{logit}^{-1}(c_{ij} - K(m,n))$$

for verb m and frame n or verb m and verb n, respectively.

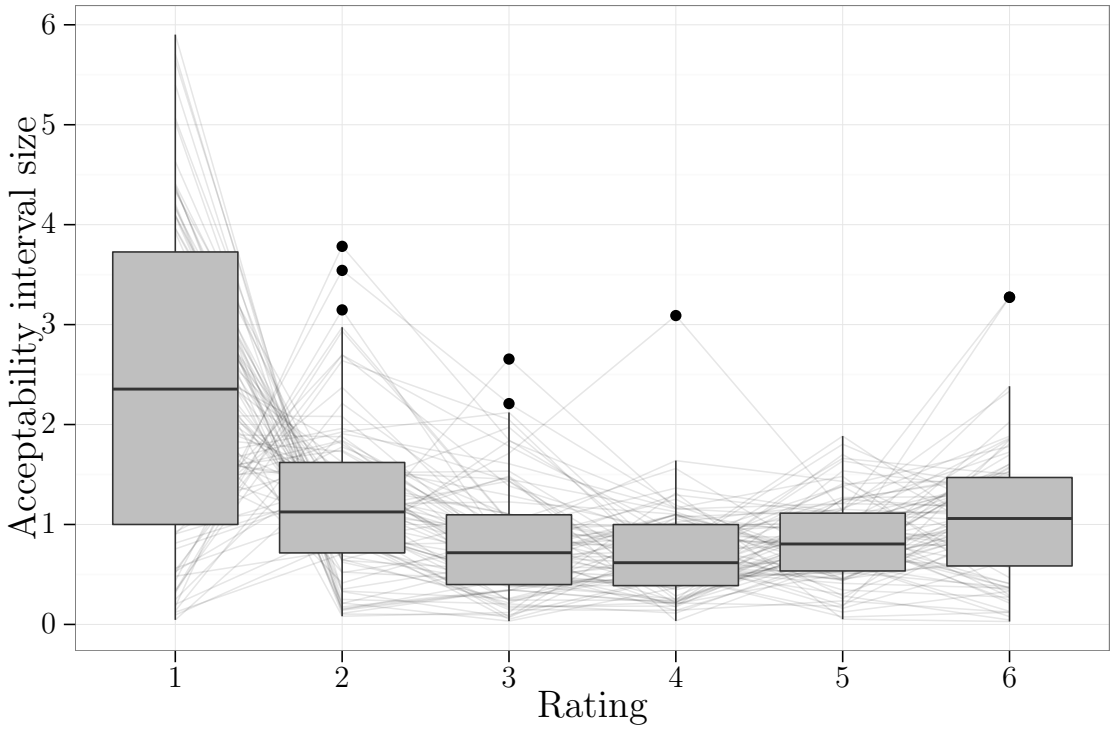


Figure A.2: Distribution over participants of size of acceptability interval mapped to each rating.

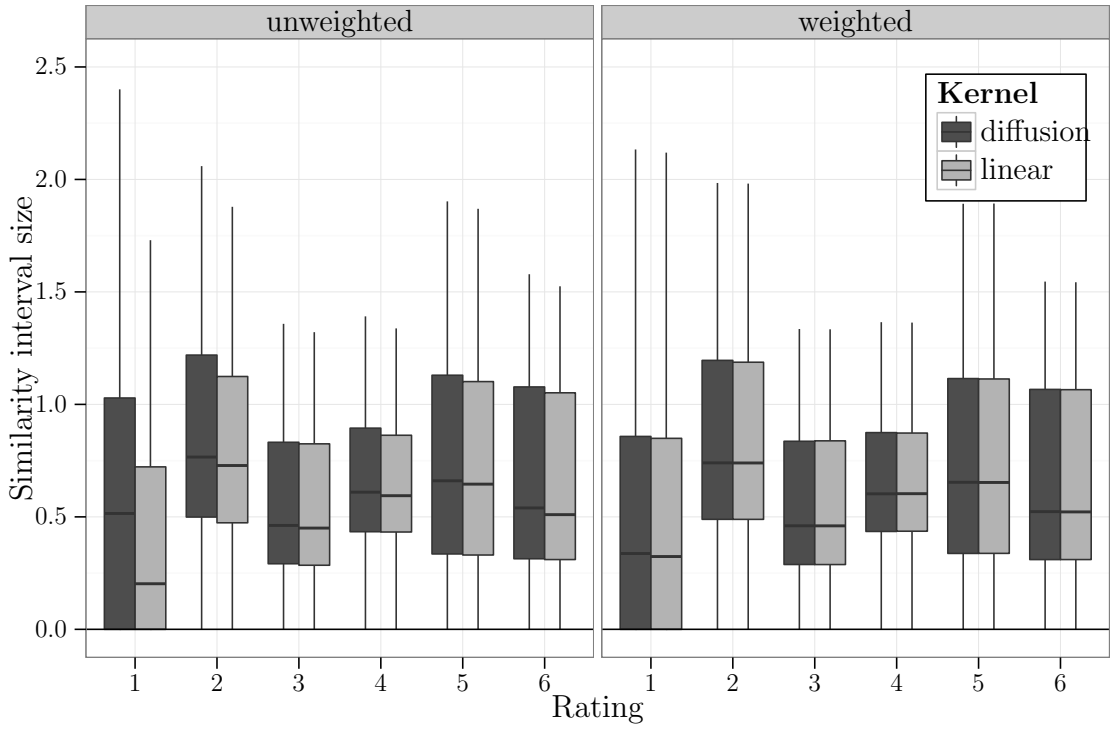


Figure A.3: Distribution over participants of size of similarity interval mapped to each rating.

A.3.2 Multinomial logit mixed model

A standard multinomial logit model with a softmax link was used.

$$\text{softmax}(\mathbf{x}) = \left[\frac{\exp(x_1)}{\sum_i \exp(x_i)}, \frac{\exp(x_2)}{\sum_i \exp(x_i)}, \ldots\right]$$

The idea behind this model is that participants choose a verb in the generalized discrimination task based on the similarity of the other two verbs according to some kernel K. The more similar those two verbs are the more likely the participant is to choose the other verb. The model also encodes a participant-specific bias vector $\mathbf{b}_i$ to account for random variation in how much participant i likes to choose a particular response based on which position it had on the display, independent of its semantics.

$$\begin{aligned}
\delta_i &\sim \text{Gamma}(1,1) = \text{Exponential}(1) \\
b_{ij} \mid \delta &\sim \text{Exponential}(\delta_i) \\
t_{imno} \mid \mathbf{B} &\sim \text{Categorical}\left(\text{softmax}\begin{pmatrix} K(n,o) + b_{i1} \\ K(m,o) + b_{i2} \\ K(m,n) + b_{i3} \end{pmatrix}\right)
\end{aligned}$$

A.3.2.1 Distribution of bias

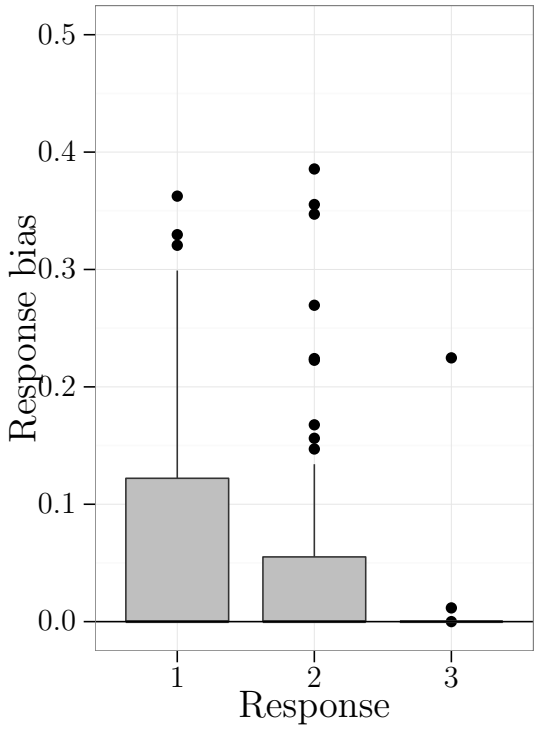


Figure A.4: Distribution over participants of bias for response based on placement in response list.

Appendix B: Appendix B

B.1 Generative story for IBP prior

In Chapter 3, I give the generative story for $\mathbf{S}$ with finite columns (features). This generative story does not work for the infinite feature case because the first outer for-loop would never terminate and thus no observed counts would be generated. To describe the infinite case, we can use the Indian Buffet Process metaphor, wherein the feature probabilities are implicitly integrated out.

Choose $K \sim \text{Poisson}(\alpha) + 1$
for verb i in $1 : V$ **do**
  for feature k in $1 : K$ **do**
    Choose a feature value $s_{ik} \sim \text{Bernoulli}\left(\frac{\sum_{v=1}^{i-1} s_{vk}}{i}\right)$
  end for
  Choose a number of new features $K_{\text{new}} \sim \text{Poisson}(\frac{\alpha}{K})$
  for new feature k in $1 : K_{\text{new}}$ **do**
    Set $s_{i(K+k)} \equiv 1$
    for verb v in $1 : (i-1)$ **do**
      Set $s_{v(K+k)} \equiv 0$

11: **end for**

12: **end for**

13: Set $K \equiv K + K_{\text{new}}$

14: **end for**

15: **for** feature k in $1 : K$ **do**

16: **for** syntactic context j in $1 : F$ **do**

17: Choose a projection strength $p_{kj} \sim \text{Exponential}(\lambda)$

18: **end for**

19: **end for**

20: **for** verb i in $1 : V$ **do**

21: Choose a verb prevalence $g_i \sim \text{Gamma}(\gamma, \delta)$

22: **for** syntactic context j in $1 : F$ **do**

23: Choose a competence distribution strength $d_{ij} \sim \text{Beta}([\mathbf{SP}]_{ij}, 1)$

24: Choose a cooccurrence count $x_{ij} \sim \text{Poisson}(g_i d_{ij})$

25: **end for**

26: **end for**

B.2 Sampler derivation

In Chapter 3, I give the derivation of the log-likelihood and log-posterior for $\mathbf{D}$. This in turn yields a relatively simple form for the log-likelihood.

$$\log \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D} \mid \mathbf{X}; \Psi) \propto \log \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi) + \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \Psi)$$

$$= \log \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi) + \log \left[\prod_{i=1}^{V} \frac{\prod_{j=1}^{F} d_{ij}^{x_{ij}}}{\left(\delta + \sum_{j=1}^{F} d_{ij}\right)^{\gamma + \sum_{j=1}^{F} x_{ij}}} \right]$$

$$= \log \mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi) + \sum_{i=1}^{V} \left[\sum_{j=1}^{F} x_{ij} \log d_{ij} \right] - \left(\gamma + \sum_{j=1}^{F} x_{ij} \right) \log \left(\delta + \sum_{j=1}^{F} d_{ij} \right)$$

The Jacobian (gradient) $\nabla \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)$ has cells.[1]

[1]The Hessian is also quite easy to compute, but since $\mathbf{D}$ is bounded, we need a constrained optimization method. Most of these methods require only Jacobians.

$$
\begin{aligned}
\frac{\partial}{\partial d_{mn}} \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta) &= \frac{\partial}{\partial d_{mn}} \sum_{i=1}^{V} \left[\sum_{j=1}^{F} x_{ij} \log d_{ij} \right] - \left(\gamma + \sum_{j=1}^{F} x_{ij} \right) \log \left(\delta + \sum_{j=1}^{F} d_{ij} \right) \\
&= \frac{\partial}{\partial d_{mn}} \left[\sum_{j=1}^{F} x_{mj} \log d_{mj} \right] - \left(\gamma + \sum_{j=1}^{F} x_{mj} \right) \log \left(\delta + \sum_{j=1}^{F} d_{mj} \right) \\
&= \left[\sum_{j=1}^{F} \frac{\partial}{\partial d_{mn}} x_{mj} \log d_{mj} \right] - \frac{\partial}{\partial d_{mn}} \left(\gamma + \sum_{j=1}^{F} x_{mj} \right) \log \left(\delta + \sum_{j=1}^{F} d_{mj} \right) \\
&= \frac{x_{mn}}{d_{mn}} - \frac{\partial}{\partial d_{mn}} \left(\gamma + \sum_{j=1}^{F} x_{mj} \right) \log \left(\delta + \sum_{j=1}^{F} d_{mj} \right) \\
&= \frac{x_{mn}}{d_{mn}} - \left(\gamma + \sum_{j=1}^{F} x_{mj} \right) \frac{\partial \log \left(\delta + \sum_{j=1}^{F} d_{mj} \right)}{\partial d_{mn}} \\
&= \frac{x_{mn}}{d_{mn}} - \frac{\left(\gamma + \sum_{j=1}^{F} x_{mj} \right)}{\left(\delta + \sum_{j=1}^{F} d_{mj} \right)}
\end{aligned}
$$

This is useful because it can be used in the initialization of a sampler—e.g. using an optimization procedure to set $\mathbf{D}$ to its MLE—and/or in the sampler itself—e.g. by including the prior over $\mathbf{D}$ to get the Jacobian of the log-posterior $\nabla \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta) \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})$, given samples for $\mathbf{S}$ and $\mathbf{P}$. Finding the Jacobian of the posterior might be useful if one does not care to quantify the distribution over $\mathbf{D}$ and are satisfied with a point estimate. For instance, the Jacobian of the log-posterior might be employed in lieu of an Markov Chain Monte Carlo (MCMC)

approach. In the next section, we derive this Jacobian explicitly in the same section that we present the necessary equations to construct a Gibbs sampler for $\mathbf{D}$.

B.2.1 Inference equations

I now show how to discover $\mathbf{S}$, $\mathbf{P}$, and $\mathbf{D}$. One way to do this is to construct a Gibbs sampler for these variables. I show how to construct two mixed approaches: one that samples $\mathbf{S}$ and $\mathbf{P}$ using Gibbs, then optimizes $\mathbf{D}$ using $\nabla \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})$; and one that samples $\mathbf{S}$, then optimizes $\mathbf{D}$ using $\nabla \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})$ and $\mathbf{P}$ using $\nabla \log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})\mathbb{P}(\mathbf{P}; \lambda)$. The last of these approaches is used the experiment below, though all approaches were implemented, and the code is available on my github.

I begin by deriving the joint $\mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi)$, which, due to the way it factors, will provide us with the raw ingredients for both the Gibbs transition probabilities and the optimization gradients. First, note that

$$\mathbb{P}(\mathbf{S}, \mathbf{P}, \mathbf{D}; \Psi) = \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})\mathbb{P}(\mathbf{P}; \lambda)\mathbb{P}(\mathbf{S}; \alpha, \beta)$$

B.2.1.1 Inferring $\mathbf{D}$

The (log-)prior on $\mathbf{D}$ has the following form.

$$\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P}) = \prod_{i=1}^{V} \prod_{j=1}^{F} \mathbb{P}(d_{ij} \mid \mathbf{S}, \mathbf{P})$$

$$= \prod_{i=1}^{V} \prod_{j=1}^{F} \frac{\Gamma([\mathbf{SP}]_{ij} + 1)}{\Gamma([\mathbf{SP}]_{ij})\Gamma(1)} d_{ij}^{[\mathbf{SP}]_{ij}-1} (1 - d_{ij})^{1-1}$$

$$= \prod_{i=1}^{V} \prod_{j=1}^{F} \frac{\Gamma([\mathbf{SP}]_{ij} + 1)}{\Gamma([\mathbf{SP}]_{ij})} d_{ij}^{[\mathbf{SP}]_{ij}-1}$$

$$= \prod_{i=1}^{V} \prod_{j=1}^{F} [\mathbf{SP}]_{ij} d_{ij}^{[\mathbf{SP}]_{ij}-1}$$

$$\log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P}) = \log \left[\prod_{i=1}^{V} \prod_{j=1}^{F} [\mathbf{SP}]_{ij} d_{ij}^{[\mathbf{SP}]_{ij}-1} \right]$$

$$= \sum_{i=1}^{V} \sum_{j=1}^{F} \log \left([\mathbf{SP}]_{ij} d_{ij}^{[\mathbf{SP}]_{ij}-1} \right)$$

$$= \sum_{i=1}^{V} \sum_{j=1}^{F} \log [\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij}$$

Interestingly, the likelihood places a very similar pressure on $\mathbf{D}$:

One can now find the gradient $\nabla \log[\mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})]$. Because the log-posterior is the sum of the log-likelihood and the log-prior, one can simply add the gradient of the log-prior to the gradient of the log-likelihood to get the log-posterior. But the gradient of the log-likelihood has already derived, so one need merely derive the gradient of the log-prior.

$$\frac{\partial}{\partial d_{mn}} \log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P}) = \frac{\partial}{\partial d_{mn}} \sum_{i=1}^{V} \sum_{j=1}^{F} \log[\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij}$$

$$= \frac{\partial}{\partial d_{mn}} ([\mathbf{SP}]_{mn} - 1) \log d_{mn}$$

$$= \frac{[\mathbf{SP}]_{mn} - 1}{d_{mn}}$$

$$\frac{\partial}{\partial d_{mn}} \log[\mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta)\mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})] = \frac{\partial}{\partial d_{mn}} \log \mathbb{P}(\mathbf{X} \mid \mathbf{D}; \gamma, \delta) + \frac{\partial}{\partial d_{mn}} \log \mathbb{P}(\mathbf{D} \mid \mathbf{S}, \mathbf{P})$$

$$= \frac{x_{mn}}{d_{mn}} - \frac{\left(\gamma + \sum_{j=1}^{F} x_{mj}\right)}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)} + \frac{[\mathbf{SP}]_{mn} - 1}{d_{mn}}$$

$$= \frac{x_{mn} + [\mathbf{SP}]_{mn} - 1}{d_{mn}} - \frac{\left(\gamma + \sum_{j=1}^{F} x_{mj}\right)}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)}$$

This gradient might be used in optimizing $\mathbf{D}$ within the internal loop of a Gibbs sampler (instead of, or prior to, sampling). Another option is to sample $\mathbf{D}$ as well. The relevant equations:

$$\mathbb{P}(d_{mn} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi) \propto \mathbb{P}(\mathbf{X} \mid d_{mn}, \mathbf{D}_{-(mn)}; \gamma, \delta)\mathbb{P}(d_{mn}, \mathbf{D}_{-(mn)} \mid \mathbf{S}, \mathbf{P})$$

$$\propto \mathbb{P}(x_{mn} \mid d_{mn}; \gamma, \delta)\mathbb{P}(d_{mn} \mid \mathbf{S}, \mathbf{P})$$

$$= \frac{d_{mn}^{x_{mn}}}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)^{\gamma + \sum_{j=1}^{F} x_{mj}}} [\mathbf{SP}]_{mn} d_{mn}^{[\mathbf{SP}]_{mn}-1}$$

$$= \frac{[\mathbf{SP}]_{mn} d_{mn}^{x_{mn}+[\mathbf{SP}]_{mn}-1}}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)^{\gamma + \sum_{j=1}^{F} x_{mj}}}$$

$$\propto \frac{d_{mn}^{x_{mn}+[\mathbf{SP}]_{mn}-1}}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)^{\gamma + \sum_{j=1}^{F} x_{mj}}}$$

$$\log \mathbb{P}(d_{mn} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi) \propto \log \frac{d_{mn}^{x_{mn}+[\mathbf{SP}]_{mn}-1}}{\left(\delta + \sum_{j=1}^{F} d_{mj}\right)^{\gamma + \sum_{j=1}^{F} x_{mj}}}$$

$$= (x_{mn} + [\mathbf{SP}]_{mn} - 1) \log d_{mn} - \left(\gamma + \sum_{j=1}^{F} x_{mj}\right) \log \left(\delta + d_{mn} + \sum_{j \neq n}^{F} d_{mj}\right)$$

This leaves us to define a proposal distribution for d_{mn}. Using a symmetric proposal distribution (given the current d_{mn}) is preferable in order to reduce computations; if the proposal is symmetric, we need not compute the ratio of proposal probabilities. Since d_{mn} must be bounded on $(0, 1)$, one simple symmetric proposal distribution is Uniform$(0, 1)$. The problem with this distribution is that it will often result in poor proposals, since it is insensitve to the current d_{mn}—henceforth, d_{mn}^{old}.

Instead, one might condition the parameters of the proposal based on the closest bound. For instance, one might use

$$\text{Uniform}\left(d_{mn}^{\text{old}} - \frac{\min\left(d_{mn}^{\text{old}}, 1 - d_{mn}^{\text{old}}\right)}{b}, d_{mn}^{\text{old}} + \frac{\min\left(d_{mn}^{\text{old}}, 1 - d_{mn}^{\text{old}}\right)}{b}\right)$$

where $b \geq 1$. The size of b determines the size of the possible jumps. For instance, if $d_{mn}^{\text{old}} = 0.5$, $b = 1$ yields Uniform$(0, 1)$, $b = 2$ yields Uniform$(.25, .75)$, etc. This is dynamic in the sense that, as d_{mn} approaches either bound, the interval over which the proposals are drawn gets smaller, regardless of b. This is nice, because intuitively, if d_{mn} is closer to the bounds already, we are probably fairly sure it should be there (though this means we will need a good initialization strategy).

This dynamicity presents a problem, however, in that the proposals will no longer be symmetric in the majority of cases. (The only time the proposals will be symmetric is if $d_{mn}^{\text{new}} = 1 - d_{mn}^{\text{old}}$.) Luckily, though, the Uniform PDF is simple, so the ratio of proposal densities is just

$$\frac{\mathbb{P}(d_{mn}^{\text{new}} \to d_{mn}^{\text{old}})}{\mathbb{P}(d_{mn}^{\text{old}} \to d_{mn}^{\text{new}})} = \frac{\frac{b}{2\min(d_{mn}^{\text{new}}, 1 - d_{mn}^{\text{new}})}}{\frac{b}{2\min(d_{mn}^{\text{old}}, 1 - d_{mn}^{\text{old}})}} = \frac{\min\left(d_{mn}^{\text{old}}, 1 - d_{mn}^{\text{old}}\right)}{\min\left(d_{mn}^{\text{new}}, 1 - d_{mn}^{\text{new}}\right)}$$

The full log acceptance probability is then

$$\log \mathrm{A}(d_{mn}^{\text{old}} \to d_{mn}^{\text{new}}) = \log \min \left(1, \frac{\mathbb{P}(d_{mn}^{\text{new}} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi)}{\mathbb{P}(d_{mn}^{\text{old}} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi)} \frac{\mathbb{P}(d_{mn}^{\text{new}} \to d_{mn}^{\text{old}})}{\mathbb{P}(d_{mn}^{\text{old}} \to d_{mn}^{\text{new}})}\right)$$

$$= \min \left(0, \log \frac{\mathbb{P}(d_{mn}^{\text{new}} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi)}{\mathbb{P}(d_{mn}^{\text{old}} \mid \mathbf{D}_{-(mn)}, \mathbf{S}, \mathbf{P}, \mathbf{X}; \Psi)} \frac{\mathbb{P}(d_{mn}^{\text{new}} \to d_{mn}^{\text{old}})}{\mathbb{P}(d_{mn}^{\text{old}} \to d_{mn}^{\text{new}})}\right)$$

$$= \min \left(\begin{array}{l} \\ (x_{mn} + [\mathbf{SP}]_{mn} - 1)\left(\log d_{mn}^{\text{new}} - \log d_{mn}^{\text{old}}\right) - \\ \\ 0, \left(\gamma + \sum_{j=1}^{F} x_{mj}\right) \left[\begin{array}{l} \log\left(\delta + d_{mn}^{\text{new}} + \sum_{j \neq n}^{F} d_{mj}\right) - \\ \\ \log\left(\delta + d_{mn}^{\text{old}} + \sum_{j \neq n}^{F} d_{mj}\right) \end{array} \right] + \\ \\ \log \min\left(d_{mn}^{\text{old}}, 1 - d_{mn}^{\text{old}}\right) - \log \min\left(d_{mn}^{\text{new}}, 1 - d_{mn}^{\text{new}}\right) \\ \\ \end{array} \right)$$

Though this equation looks quite expensive, much of the necessary computations can be done once and saved.

B.2.1.2 Inferring $\mathbf{P}$

The simplicity of our prior on $\mathbf{P}$ makes its gradient and sampling equations extremely easy to derive. Whether sampling or optimizing $\mathbf{P}$, we assume that $\mathbf{D}$, $\mathbf{S}$, and $\mathbf{X}$ are known.

$$\log \mathbb{P}(\mathbf{P} \mid \mathbf{D}, \mathbf{S}, \mathbf{X}; \lambda) \propto \log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \log \mathbb{P}(\mathbf{P}; \lambda)$$

We already know the first term, so we need merely derive the second. (In this section, we leave the restrictor of the sums over k undefined. The reason for this has to do with the way $\mathbf{S}$ is sampled, so we defer discussion until the next section.)

$$\begin{aligned}
\log \mathbb{P}(\mathbf{P}; \lambda) &= \sum_k \sum_{j=1}^{F} \log \mathbb{P}(p_{kj}; \lambda) \\
&= \sum_k \sum_{j=1}^{F} \log \left(\lambda \exp[-\lambda p_{kj}]\right) \\
&\propto \sum_k \sum_{j=1}^{F} -\lambda p_{kj} \\
&= -\lambda \sum_k \sum_{j=1}^{F} p_{kj} \\
&= -\lambda \|\mathbf{P}\|_1
\end{aligned}$$

Note that this is just an L1 regularizer, as mentioned above. The Jacobian of the prior is quite simple; the Jacobian of the likelihood (prior on $\mathbf{D}$) is a bit more complicated.

$$\frac{\partial}{\partial p_{ln}} \log \mathbb{P}(\mathbf{P}; \lambda) = \frac{\partial}{\partial p_{ln}} \left[-\lambda \sum_k \sum_{j=1}^{F} p_{kj} \right]$$

$$= -\lambda$$

$$\frac{\partial}{\partial p_{ln}} \left[\log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \log \mathbb{P}(\mathbf{P}; \lambda) \right] = \frac{\partial}{\partial p_{ln}} \log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \frac{\partial}{\partial p_{ln}} \log \mathbb{P}(\mathbf{P}; \lambda)$$

$$= -\lambda + \frac{\partial}{\partial p_{ln}} \log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S})$$

$$= -\lambda + \frac{\partial}{\partial p_{ln}} \sum_{i=1}^{V} \sum_{j=1}^{F} \log [\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij}$$

$$= -\lambda + \sum_{i=1}^{V} \sum_{j=1}^{F} \frac{\partial}{\partial p_{ln}} \log [\mathbf{SP}]_{ij} + \frac{\partial}{\partial p_{ln}} ([\mathbf{SP}]_{ij} - 1) \log d_{ij}$$

$$= -\lambda + \sum_{i=1}^{V} \sum_{j=1}^{F} \frac{\partial}{\partial p_{ln}} \log \sum_k s_{ik} p_{kj} + \frac{\partial}{\partial p_{ln}} (-1 + \sum_k s_{ik} p_{kj}) \log d_{ij}$$

$$= -\lambda + \sum_{i=1}^{V} \sum_{j=1}^{F} \frac{s_{il} p_{ln}}{\sum_k s_{ik} p_{kj}} + s_{il} p_{ln} \log d_{ij}$$

$$= -\lambda + \sum_{i=1}^{V} \sum_{j=1}^{F} s_{il} p_{ln} \left[\frac{1}{\sum_k s_{ik} p_{kj}} + \log d_{ij} \right]$$

$$= -\lambda + p_{ln} \sum_{i=1}^{V} s_{il} \sum_{j=1}^{F} \frac{1}{[\mathbf{SP}]_{ij}} + \log d_{ij}$$

The Gibbs sampling equation is given by

$$\log \mathbb{P}(p_{ln} \mid \mathbf{P}_{-(ln)}, \mathbf{D}, \mathbf{S}, \mathbf{X}; \lambda) \propto \log \mathbb{P}(\mathbf{D} \mid p_{ln}, \mathbf{P}_{-(ln)}, \mathbf{S}) + \log \mathbb{P}(p_{ln}, \mathbf{P}_{-(ln)}; \lambda)$$

$$= \sum_{i=1}^{V} \sum_{j=1}^{F} \log[\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij} - \lambda \sum_{k} \sum_{j=1}^{F} p_{kj}$$

$$\propto -\lambda p_{ln} + \begin{array}{l} \sum_{i=1}^{V} \log[\mathbf{SP}]_{in} + ([\mathbf{SP}]_{in} - 1) \log d_{in} \\ \\ + \sum_{i=1}^{V} \sum_{j \neq n}^{F} \log[\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij} \end{array}$$

$$\propto -\lambda p_{ln} + \sum_{i=1}^{V} \log[\mathbf{SP}]_{in} + ([\mathbf{SP}]_{in} - 1) \log d_{in}$$

$$= -\lambda p_{ln} + \sum_{i=1}^{V} \log[\mathbf{SP}]_{in} + \left(\left[\sum_{k} s_{ik} p_{kn} \right] - 1 \right) \log d_{in}$$

$$\propto -\lambda p_{ln} + \sum_{i=1}^{V} \log[\mathbf{SP}]_{in} + (s_{il} p_{ln} - 1) \log d_{in}$$

B.2.1.3 Inferring $\mathbf{S}$

As with $\mathbf{P}$, we ultimately want to compute

$$\log \mathbb{P}(\mathbf{S} \mid \mathbf{D}, \mathbf{P}, \mathbf{X}; \lambda) \propto \log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S}) + \log \mathbb{P}(\mathbf{S}; \alpha, \beta)$$

And since we already know the form of $\log \mathbb{P}(\mathbf{D} \mid \mathbf{P}, \mathbf{S})$, we need merely define the prior over $\mathbf{S}$. I specified two related priors on $\mathbf{S}$: one that requires a set number of features K and another that allows for an unbounded number of features. Both take advantage of the Beta-Bernoulli conjugacy. We'll derive the finite (parametric) case first, then return to the infinite (nonparametric) case. (A less explicit derivation of the finite case can be found in Griffiths and Ghahramani 2011, which is more concerned with the infinite case.)

The finite prior is given by

$$
\begin{aligned}
\mathbb{P}(\mathbf{S} \mid \alpha, \beta, K) &= \int_{(0,1)^K} \mathrm{d}\pi \;\; \mathbb{P}(\mathbf{S}, \pi \mid \alpha, \beta, K) \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \mathbb{P}(\mathbf{S} \mid \pi) \mathbb{P}(\pi \mid \alpha, \beta, K) \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \prod_{k=1}^{K} \mathbb{P}(\pi_k \mid \alpha, \beta) \prod_{i=1}^{V} \mathbb{P}(s_{ik} \mid \pi_k) \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \prod_{k=1}^{K} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi_k^{\alpha-1}(1-\pi_k)^{\beta-1} \prod_{i=1}^{V} \pi_k^{s_{ik}}(1-\pi_k)^{1-s_{ik}} \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \prod_{k=1}^{K} \pi_k^{\alpha-1}(1-\pi_k)^{\beta-1} \prod_{i=1}^{V} \pi_k^{s_{ik}}(1-\pi_k)^{1-s_{ik}} \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \prod_{k=1}^{K} \pi_k^{\alpha-1}(1-\pi_k)^{\beta-1} \pi_k^{\sum_{i=1}^{V} s_{ik}}(1-\pi_k)^{V-\sum_{i=1}^{V} s_{ik}} \\
&= \int_{(0,1)^K} \mathrm{d}\pi \;\; \prod_{k=1}^{K} \pi_k^{\alpha-1}(1-\pi_k)^{\beta-1} \pi_k^{\sum_{i=1}^{V} s_{ik}}(1-\pi_k)^{V-\sum_{i=1}^{V} s_{ik}} \\
&= \prod_{k=1}^{K} \int_{(0,1)} \mathrm{d}\pi \;\; \pi_k^{\alpha-1+\sum_{i=1}^{V} s_{ik}}(1-\pi_k)^{\beta-1+V-\sum_{i=1}^{V} s_{ik}} \\
&= \prod_{k=1}^{K} \int_{(0,1)} \mathrm{d}\pi \;\; \pi_k^{\alpha-1+\sum_{i=1}^{V} s_{ik}}(1-\pi_k)^{\beta-1+V-\sum_{i=1}^{V} s_{ik}} \\
&= \prod_{k=1}^{K} \frac{\Gamma\left(\alpha + \sum_{i=1}^{V} s_{ik}\right) \Gamma\left(\beta + V - \sum_{i=1}^{V} s_{ik}\right)}{\Gamma(\alpha+\beta+V)} \\
&\propto \prod_{k=1}^{K} \Gamma\left(\alpha + \sum_{i=1}^{V} s_{ik}\right) \Gamma\left(\beta + V - \sum_{i=1}^{V} s_{ik}\right)
\end{aligned}
$$

And the log-prior is

$$\log \mathbb{P}(\mathbf{S} \mid \alpha, \beta) \propto \log \prod_{k=1}^{K} \Gamma\left(\alpha + \sum_{i=1}^{V} s_{ik}\right) \Gamma\left(\beta + V - \sum_{i=1}^{V} s_{ik}\right)$$

$$= \sum_{k=1}^{K} \log \Gamma\left(\alpha + \sum_{i=1}^{V} s_{ik}\right) + \log \Gamma\left(\beta + V - \sum_{i=1}^{V} s_{ik}\right)$$

Since it will be useful in deriving the Gibbs equation, we now derive the conditional distribution of s_{il} given all other cells of $\mathbf{S}$.

$$\mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) \propto \mathbb{P}(s_{ml}, \mathbf{S}_{-(ml)}; \alpha, \beta)$$

$$\propto \prod_{k=1}^{K} \Gamma\left(\alpha + \sum_{i=1}^{V} s_{ik}\right) \Gamma\left(\beta + V - \sum_{i=1}^{V} s_{ik}\right)$$

$$\propto \Gamma\left(\alpha + s_{ml} + \sum_{i \neq m}^{V} s_{il}\right) \Gamma\left(\beta + V - s_{ml} - \sum_{i \neq m}^{V} s_{il}\right)$$

Note that this can be simplified when taking the ratio of the probability that $s_{ml} = 1$ to the probability that $s_{ml} = 0$.

$$\frac{\mathbb{P}(s_{ml}=1 \mid \mathbf{S}_{-(ml)};\alpha,\beta)}{\mathbb{P}(s_{ml}=0 \mid \mathbf{S}_{-(ml)};\alpha,\beta)} \propto \frac{\Gamma\left(\alpha+1+\sum_{i\neq m}^{V} s_{il}\right)\Gamma\left(\beta+V-1-\sum_{i\neq m}^{V} s_{il}\right)}{\Gamma\left(\alpha+0+\sum_{i\neq m}^{V} s_{il}\right)\Gamma\left(\beta+V-0-\sum_{i\neq m}^{V} s_{il}\right)}$$

$$= \frac{\alpha+\sum_{i\neq m}^{V} s_{il}}{\beta+V-1+\sum_{i\neq m}^{V} s_{il}}$$

(Note that inverting the ratio results in inverting the term involving α with respect to the term involving β.)

we now move onto the infinite case. Due to some complexities that arise from taking limits, the prior in the infinite case must be specified in terms of equivalence classes over binary matrices. Griffiths and Ghahramani (2011) discuss this extensively. The gist is that, if variates (instantiations) of $\mathbf{S}$ are treated as sequences of binary numbers, one can define a mapping $[\cdot]$ that maps from unordered sequences to ordered sequences. The net effect of this is to reduce the support of the prior considerably. The prior over these equivalence classes can then be defined in terms of a sum over the image of $[\mathbf{S}]$.

$$\mathbb{P}([\mathbf{S}] \mid \alpha) = \sum_{\mathbf{S} \in [\mathbf{S}]} \mathbb{P}(\mathbf{S} \mid \alpha)$$

$$= \frac{\alpha^{K_+} \exp[-\alpha H_V]}{\prod_{h=1}^{2^V-1} K_h!} \prod_{k=1}^{K_+} \frac{\Gamma(\sum_{i=1}^{V} s_{ik})\Gamma(V + 1 - \sum_{i=1}^{V} s_{ik})}{V!}$$

where K_+ is the number of columns k for which $\sum_{i=1}^{V} s_{ik}$ is nonzero and K_h is the number of columns that correspond to the h^{th} binary number. (Only K_+ is important in the actual sampling, since it gives the number of rows of $\mathbf{P}$ we must keep track of.) This makes the relationship to the finite case clear ($\frac{\alpha}{K} \to 0$ and $\beta \equiv 1$). The only difference is in the fact that K_+ may vary. (Indeed, this is the point of using the nonparametric prior in the first place.)

Interestingly, this equation simplifies substantially when computing the same conditionally probability as above. Indeed, the infinite version of $\mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta)$ is just like the finite version.

$$\frac{\mathbb{P}(s_{ml} = 1 \mid \mathbf{S}_{-(ml)}; \alpha, \beta)}{\mathbb{P}(s_{ml} = 0 \mid \mathbf{S}_{-(ml)}; \alpha, \beta)} \propto \frac{\sum_{i \neq m}^{V} s_{il}}{V - \sum_{i \neq m}^{V} s_{il}}$$

The trade-off is that some of the complexity inherent to the infinite version is moved into other procedures in the sampler, as we specify below.

we can now move onto computing $\log \mathbb{P}(\mathbf{S} \mid \mathbf{D}, \mathbf{P}, \mathbf{X}; \lambda)$. For both $\mathbf{D}$ and $\mathbf{P}$, we found both the gradient for the relevant quantity and the Gibbs equation. In this case, computing a gradient is not useful since $\mathbf{S}$ is discrete. And since the constraints would not be linear, we cannot use a method like Integer Linear Programming (ILP). Thus, we will compute only the Gibbs equations.

General to both the finite and infinite cases (where, as for $\mathbf{P}$, $[\mathbf{SP}]_{mj}$ assumes the replacement of s_{ml} with the proposed):

$$\log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}, \mathbf{D}, \mathbf{P}, \mathbf{X}; \Psi) \propto \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \log \mathbb{P}(\mathbf{D} \mid s_{ml}, \mathbf{S}_{-(ml)}, \mathbf{P})$$

$$= \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \begin{array}{l} \left[\sum_{j=1}^{F} \log[\mathbf{SP}]_{mj} + ([\mathbf{SP}]_{mj} - 1) \log d_{mj}\right] \\ + \left[\sum_{i \neq m}^{V} \sum_{j=1}^{F} \log[\mathbf{SP}]_{ij} + ([\mathbf{SP}]_{ij} - 1) \log d_{ij}\right] \end{array}$$

$$\propto \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \sum_{j=1}^{F} \log[\mathbf{SP}]_{mj} + ([\mathbf{SP}]_{mj} - 1) \log d_{mj}$$

$$= \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \sum_{j=1}^{F} \log[\mathbf{SP}]_{mj} + \left(\left[\sum_{k} s_{mk} p_{kj}\right] - 1\right) \log d_{mj}$$

$$\propto \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \sum_{j=1}^{F} \log[\mathbf{SP}]_{mj} + (s_{ml} p_{lj} - 1) \log d_{mj}$$

$$= \log \mathbb{P}(s_{ml} \mid \mathbf{S}_{-(ml)}; \alpha, \beta) + \sum_{j=1}^{F} \log \left[s_{ml} p_{lj} + \sum_{k \neq l} s_{mk} p_{kj}\right] + (s_{ml} p_{lj} - 1) \log d_{mj}$$

The log posterior odds for the finite case is then given by

$$\log \frac{\mathbb{P}(s_{ml}=1 \mid \mathbf{S}_{-(ml)}, \mathbf{D}, \mathbf{P}, \mathbf{X}; \Psi)}{\mathbb{P}(s_{ml}=0 \mid \mathbf{S}_{-(ml)}, \mathbf{D}, \mathbf{P}, \mathbf{X}; \Psi)} \propto \log \frac{\alpha + \sum_{i \neq m}^{V} s_{il}}{\beta + V - 1 + \sum_{i \neq m}^{V} s_{il}} + \begin{array}{l} \sum_{j=1}^{F} \log[\mathbf{SP}]_{mj,1} + (p_{lj} - 1) \log d_{mj} \\ \\ - \sum_{j=1}^{F} \log[\mathbf{SP}]_{mj,0} - \log d_{mj} \end{array}$$

$$\propto \log \frac{\alpha + \sum_{i \neq m}^{V} s_{il}}{\beta + V - 1 + \sum_{i \neq m}^{V} s_{il}} + \sum_{j=1}^{F} \begin{array}{l} \log[\mathbf{SP}]_{mj,1} - \log[\mathbf{SP}]_{mj,0} \\ \\ + p_{lj} \log d_{mj} \end{array}$$

$$= \log \frac{\alpha + \sum_{i \neq m}^{V} s_{il}}{\beta + V - 1 + \sum_{i \neq m}^{V} s_{il}} + \sum_{j=1}^{F} \begin{array}{l} \log \left[p_{lj} + \sum_{k \neq l} s_{mk} p_{kj} \right] \\ \\ - \log \left[\sum_{k \neq l} s_{mk} p_{kj} \right] \\ \\ + p_{lj} \log d_{mj} \end{array}$$

And the log posterior odds for the infinite case is given by

$$\log \left[p_{lj} + \textstyle\sum_{k \neq l} s_{mk} p_{kj} \right]$$

$$\log \frac{\mathbb{P}(s_{ml} = 1 \mid \mathbf{S}_{-(ml)}, \mathbf{D}, \mathbf{P}, \mathbf{X}; \Psi)}{\mathbb{P}(s_{ml} = 0 \mid \mathbf{S}_{-(ml)}, \mathbf{D}, \mathbf{P}, \mathbf{X}; \Psi)} \propto \log \frac{\sum_{i \neq m}^{V} s_{il}}{V - \sum_{i \neq m}^{V} s_{il}} + \sum_{j=1}^{F} - \log \left[\textstyle\sum_{k \neq l} s_{mk} p_{kj} \right]$$

$$+ p_{lj} \log d_{mj}$$

Finally, note that the log posterior odds is equivalent to the acceptance probability since the only reasonable proposal distribution (up to isomorphism), Bernoulli($1 - s_{ml}^{\text{old}}$), will always cancel out.

Bibliography

Abbott, Barbara. 2006. Where have some of the presuppositions gone. *Drawing the boundaries of meaning: Neo-Gricean studies in pragmatics and semantics in honor of Laurence R. Horn* 1–20.

Abusch, Dorit. 2002. Lexical alternatives as a source of pragmatic presuppositions. In *Proceedings of SALT*, volume 12, 1–19.

Ackley, David H., Geoffrey E. Hinton, and Terrence J. Sejnowski. 1985. A learning algorithm for boltzmann machines*. *Cognitive science* 9:147–169.

Alishahi, Afra, and Suzanne Stevenson. 2008. A computational model of early argument structure acquisition. *Cognitive science* 32:789–834.

Ambar, Manuela. 1999. Aspects of the syntax of focus in Portuguese. *The grammar of focus* 24:23.

Anand, Pranav, and Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6:1–59.

Anand, Pranav, and Valentine Hacquard. 2014. Factivity, Belief and Discourse. In *The Art and Craft of Semantics: A Festschrift for Irene Heim*, ed. Luka Crnič and Uli Sauerland, volume 1, 69–90. Cambride, MA: MIT Working Papers in Linguistics.

Asher, Nicholas. 2000. Truth conditional discourse semantics for parentheticals. *Journal of Semantics* 17:31–50.

Baker, Carl L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 533–581.

Baker, Mark C. 1988. *Incorporation: A theory of grammatical function changing*. University of Chicago Press Chicago.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2014a. Gradual Acquisition of Mental State Meaning: A Computational Investigation. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society*.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2014b. Learning verb classes in an incremental model. *ACL 2014* 37.

Barak, Libby, Afsaneh Fazly, and Suzanne Stevenson. 2013. Acquisition of Desires before Beliefs: A Computational Investigation. In *Proceedings of the 17th Conference on Computational Natural Language Learning (CoNLL)*.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation* 43:209–226.

Bartsch, Renate. 1973. Negative transportation" gibt es nicht. *Linguistische Berichte* 27.

Bengio, Yoshua, Holger Schwenk, Jean-Sébastien Senécal, Fréderic Morin, and Jean-Luc Gauvain. 2006. Neural probabilistic language models. In *Innovations in Machine Learning*, 137–186. Springer.

Berwick, Robert C. 1985. *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3:993–1022.

Bolinger, Dwight. 1968. Postposed main phrases: an English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14:3–30.

Breiman, Leo. 2001. Random forests. *Machine learning* 45:5–32.

Briscoe, Ted, and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the fifth conference on Applied natural language processing*, 356–363. Association for Computational Linguistics.

Brown, Roger. 1957. Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology* 55:1.

Carey, Susan, and Elsa Bartlett. 1978. Acquiring a single new word. *Papers and Reports on Child Language Development* 15:17–29.

Carter, Richard. 1976. Some linking regularities. *On Linking: Papers by Richard Carter Cambridge MA: Center for Cognitive Science, MIT (Lexicon Project Working Papers No. 25)* .

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems*, 288–296.

Chomsky, Noam. 1981. *Lectures on Government and Binding: The Pisa Lectures*. Walter de Gruyter.

Church, Kenneth W., and William A. Gale. 1995. Poisson mixtures. *Natural Language Engineering* 1:163–190.

Church, Kenneth Ward, and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Comput. Linguist.* 16:22–29. URL `http://dl.acm.org/citation.cfm?id=89086.89095`.

Cinque, Guglielmo. 1999. *Adverbs and functional heads: A cross-linguistic perspective*. Oxford University Press.

Coates, Adam, Andrew Y. Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *International conference on artificial intelligence and statistics*, 215–223.

Collobert, Ronan, and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, 160–167. ACM.

Connor, Michael, Cynthia Fisher, and Dan Roth. 2013. Starting from scratch in semantic role labeling: Early indirect supervision. In *Cognitive aspects of computational language acquisition*, 257–296. Springer.

Dayal, Veneeta, and Jane Grimshaw. 2009. Subordination at the interface: the Quasi-Subordination Hypothesis.

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *JAsIs* 41:391–407.

Den Besten, Hans. 1983. On the interaction of root transformations and lexical deletive rules. *On the formal syntax of the Westgermania* 47–131.

Depiante, Marcela Andrea. 2000. The syntax of deep and surface anaphora: a study of null complement anaphora and stripping/bare argument ellipsis. Doctoral Dissertation, University of Connecticut.

Diessel, Holger, and Michael Tomasello. 2001. *The acquisition of finite complement clauses in English: A corpus-based analysis*.

Dowty, David. 1991. Thematic proto-roles and argument selection. *Language* 67:547–619.

Drummond, Alex. 2014. Ibex. URL `https://github.com/addrummond/ibex`.

Dudley, Rachel, Naho Orita, Valentine Hacquard, and Jeffrey Lidz. 2015. Three-

year-olds' understanding of know and think. In *Experimental Perspectives on Presuppositions*, 241–262. Springer.

Dunning, Ted. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics* 19:61–74.

Egré, Paul. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77:85–125.

Farkas, Donka. 1985. *Intensional descriptions and the Romance subjunctive mood*. Taylor & Francis.

Fillmore, Charles J. 1963. The position of embedding transformations in a grammar. *WORD-JOURNAL OF THE INTERNATIONAL LINGUISTIC ASSOCIATION* 19:208–231.

Fillmore, Charles John. 1970. The Grammar of Hitting and Breaking. In *Readings in English Transformational Grammar*, ed. R.A. Jacobs and P.S. Rosenbaum, 120–133. Waltham, MA,: Ginn.

Fisher, Cynthia, Henry Gleitman, and Lila R. Gleitman. 1991. On the semantic content of subcategorization frames. *Cognitive psychology* 23:331–392.

Fodor, Jerry, and Ernie Lepore. 1999. Impossible Words? *Linguistic Inquiry* 30:445–453. URL `http://www.jstor.org/stable/4179071`.

Fodor, Jerry A., and Ernie Lepore. 1998. The emptiness of the lexicon: reflections on James Pustejovsky's The Generative Lexicon. *Linguistic Inquiry* 29:269–288.

Frank, Michael C., Noah D. Goodman, and Joshua B. Tenenbaum. 2009. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science* 20:578–585.

Fyshe, Alona, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2014. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of ACL*.

Fyshe, Alona, Leila Wehbe, Partha P. Talukdar, Brian Murphy, and Tom M. Mitchell. 2015. A Compositional and Interpretable Semantic Space. In *Proceedings of the NAACL-HLT*. Denver.

Gajewski, Jon Robert. 2007. Neg-raising and polarity. *Linguistics and Philosophy* 30:289–328.

Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. 2013. *Bayesian data analysis*. CRC press.

Giannakidou, Anastasia. 1997. The landscape of polarity items. Doctoral Dissertation, University of Groningen.

Gillette, Jane, Henry Gleitman, Lila Gleitman, and Anne Lederer. 1999. Human simulations of vocabulary learning. *Cognition* 73:135–176.

Ginzburg, Jonathan. 1995. Resolving questions, II. *Linguistics and Philosophy* 18:567–609.

Ginzburg, Jonathan, and Ivan Sag. 2001. *Interrogative investigations*. Stanford: CSLI publications.

Giorgi, Alessandra, and Fabio Pianesi. 1997. *Tense and Aspect: Form Semantics to Morphosyntax*. Oxford: Oxford University Press.

Gleitman, Lila. 1990. The structural sources of verb meanings. *Language acquisition* 1:3–55.

Gleitman, Lila R., Kimberly Cassidy, Rebecca Nappa, Anna Papafragou, and John C. Trueswell. 2005. Hard words. *Language Learning and Development* 1:23–64.

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

Goldwater, Sharon, Thomas L. Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *The Journal of Machine Learning Research* 12:2335–2382.

Goodman, Nelson. 1955. *Fact, fiction, and forecast*. Harvard University Press.

Gormley, Matthew R., Mark Dredze, Benjamin Van Durme, and Jason Eisner. 2012. Shared components topic models. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 783–792. Association for Computational Linguistics.

Griffiths, Thomas, and Zoubin Ghahramani. 2006. Infinite latent feature models and

the Indian buffet process. In *Advances in Neural Information Processing Systems 18*, ed. Y. Weiss, B. Schölkopf, and J.C. Platt, 475–482. Cambridge, MA: MIT Press.

Griffiths, Thomas L., and Zoubin Ghahramani. 2011. The indian buffet process: An introduction and review. *The Journal of Machine Learning Research* 12:1185–1224.

Griffiths, Thomas L., Mark Steyvers, and Joshua B. Tenenbaum. 2007. Topics in semantic representation. *Psychological review* 114:211.

Grimshaw, Jane. 1979. Complement selection and the lexicon. *Linguistic inquiry* 10:279–326.

Grimshaw, Jane. 1990. *Argument structure*. Cambridge, MA: MIT Press.

Grimshaw, Jane. 1994. Lexical reconciliation. *Lingua* 92:411–430.

Grimshaw, Jane. 2009. That's nothing: the grammar of complementizer omission.

Grishman, Ralph, Catherine Macleod, and Adam Meyers. 1994. COMLEX syntax: Building a computational lexicon. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, 268–272. Association for Computational Linguistics.

Gruber, Jeffrey Steven. 1965. Studies in lexical relations. Doctoral Dissertation, Massachusetts Institute of Technology.

Guttman, Louis. 1954. Some necessary conditions for common-factor analysis. *Psychometrika* 19:149–161.

Hacquard, Valentine. 2010. On the event relativity of modal auxiliaries. *Natural language semantics* 18:79–114.

Hacquard, Valentine. 2014. Bootstrapping attitudes. In *Semantics and Linguistic Theory*, volume 24, 330–352.

Hacquard, Valentine, and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics* 5:1–29.

Hale, Ken, and Samuel Jay Keyser. 2002. *Prolegomena to a Theory of Argument Structure*. Cambridge, MA: MIT Press.

Hankamer, Jorge, and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry* 391–428.

Harrigan, Kaitlyn. 2015. Syntactic bootstrapping in the acquisition of attitude verbs. Doctoral Dissertation, University of Maryland.

Hartshorne, Joshua K., Claire Bonial, and Martha Palmer. 2013. The VerbCorner Project: Toward an Empirically-Based Semantic Decomposition of Verbs. In *EMNLP*, 1438–1442.

Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of semantics* 9:183–221.

Hintikka, Jaakko. 1975. Different Constructions in Terms of the Basic Epistemological Verbs: A Survey of Some Problems and Proposals. In *The Intentions of Intentionality and Other New Models for Modalities*, 1–25. Dordrecht: D. Reidel.

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313:504–507.

Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. 2009. Replicated softmax: an undirected topic model. In *Advances in neural information processing systems*, 1607–1614.

Hooper, Joan B. 1975. On Assertive Predicates. In *Syntax and Semantics*, ed. John P. Kimball, volume 4, 91–124.

Horn, Laurence. 1978. Remarks on neg-raising. *Syntax and semantics* 9:129–220.

Horn, Laurence R. 1971. Negative transportation: Unsafe at any speed. In *Chicago Linguistic Society*, volume 7, 120–133.

Horn, Laurence R. 1975. Neg-raising predicates: Toward an explanation. *CLS* 11:94.

Horn, Laurence R. 1989. *A natural history of negation*, volume 960. University of Chicago Press Chicago.

Horn, Laurence Robert. 1972. On the semantic properties of logical operators in English. Doctoral Dissertation, UCLA.

Hoyer, Patrik O. 2002. Non-negative sparse coding. In *Neural Networks for Signal*

Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on, 557–565. IEEE.

Hoyer, Patrik O. 2004. Non-negative matrix factorization with sparseness constraints. *The Journal of Machine Learning Research* 5:1457–1469.

Hurley, Niall, and Scott Rickard. 2009. Comparing measures of sparsity. *Information Theory, IEEE Transactions on* 55:4723–4741.

Jackendoff, Ray. 1972. *Semantic interpretation in generative grammar*. MIT press Cambridge, MA.

Jackendoff, Ray. 1990. *Semantic structures*, volume 18. MIT press.

Jackson, Donald A. 1993. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. *Ecology* 2204–2214.

Johnson-Laird, Philip N. 1983. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press.

Jones, Eric, Travis Oliphant, and Pearu Peterson. 2001. {SciPy}: Open source scientific tools for {Python}. URL `http://www.scipy.org/`.

Kako, Edward. 1997. Subcategorization Semantics and the Naturalness of Verb-Frame Pairings. *University of Pennsylvania Working Papers in Linguistics* 4:11.

Kako, Edward. 2006. Thematic role properties of subjects and objects. *Cognition* 101:1–42.

Karttunen, Lauri. 1971. Some observations on factivity. *Paper in Linguistics* 4:55–69. URL http://www.tandfonline.com/doi/abs/10.1080/08351817109370248.

Karttunen, Lauri, and Stanley Peters. 1979. Conventional implicature. *Syntax and Semantics* 11:1–56.

Katz, Jerrold J., and Jerry A. Fodor. 1963. The Structure of a Semantic Theory. *Language* 39:170–210. URL http://www.jstor.org/stable/411200.

Kiparsky, Paul, and Carol Kiparsky. 1970. Fact. In *Progress in Linguistics: A Collection of Papers*, ed. Manfred Bierwisch and Karl Erich Heidolph, 143–173. The Hague: Mouton.

Kondor, Risi Imre, and John Lafferty. 2002. Diffusion kernels on graphs and other discrete input spaces. In *ICML*, volume 2, 315–322.

Korhonen, Anna. 2002. Subcategorization Acquisition. Doctoral Dissertation, University of Cambridge.

Korhonen, Anna, Yuval Krymolowski, and Ted Briscoe. 2006. A large subcategorization lexicon for natural language processing applications. In *Proceedings of LREC*, volume 6.

Krifka, Manfred. 2001. Quantifying into question acts. *Natural language semantics* 9:1–40.

Kripke, Saul A. 1982. *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press.

Kruskal, Joseph B. 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29:1–27.

Kruskal, Joseph B. 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29:115–129.

Lahiri, Utpal. 2002. *Questions and answers in embedded contexts*. Oxford University Press.

Landau, Barbara, and Lila R. Gleitman. 1985. *Language and experience: Evidence from the blind child*, volume 8. Harvard University Press.

Landauer, Thomas K., and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review* 104:211.

Larochelle, Hugo, and Yoshua Bengio. 2008. Classification using discriminative restricted Boltzmann machines. In *Proceedings of the 25th international conference on Machine learning*, 536–543. ACM.

Lasnik, Howard. 1989. On certain substitutes for negative data. In *Learnability and linguistic theory*, 89–105. Springer.

Lederer, Anne, Henry Gleitman, and Lila Gleitman. 1995. Verbs of a feather flock

together: Semantic information in the structure of maternal speech. *Beyond names for things: Young children's acquisition of verbs* 277.

Lee, Daniel D., and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791.

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

Levin, Beth, and Malka Rappaport Hovav. 2005. *Argument realization*. Cambridge University Press.

Levy, Omer, and Yoav Goldberg. 2014a. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, 302–308.

Levy, Omer, and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2177–2185.

Lewis, Mike, and Mark Steedman. 2013. Combining Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics* 1:179–192.

Lidz, Jeffrey, Henry Gleitman, and Lila Gleitman. 2004. Kidz in the 'hood: Syntactic bootstrapping and the mental lexicon. In *Weaving a Lexicon*, ed. D.G. Hall and S.R. Waxman, 603–636. Cambridge, MA: MIT Press.

MacWhinney, Brian. 2014a. *The CHILDES project: Tools for analyzing talk, Volume I: Transcription format and programs*. Psychology Press.

MacWhinney, Brian. 2014b. *The CHILDES project: Tools for analyzing talk, Volume II: The database*. Psychology Press.

Markman, Ellen M. 1990. Constraints children place on word meanings. *Cognitive Science* 14:57–77.

Markman, Ellen M., and Jean E. Hutchinson. 1984. Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations. *Cognitive psychology* 16:1–27.

Markman, Ellen M., and Gwyn F. Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology* 20:121–157.

Marr, David. 1982. Vision: a computational investigation into the human representation and processing of visual information. *Henry Holt and Co.* .

Masur, Elise F., and Jean B. Gleason. 1980. Parent–child interaction and the acquisition of lexical information during play. *Developmental Psychology* 16:404.

Medina, Tamara Nicol, Jesse Snedeker, John C. Trueswell, and Lila R. Gleitman. 2011. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences* 108:9014–9019.

Merlo, Paola, and Suzanne Stevenson. 2001. Automatic verb classification based on

statistical distributions of argument structure. *Computational Linguistics* 27:373–408.

Merriman, William E., and Laura L. Bowman. 1989. *The mutual exclusivity bias in children's word learning*. Number 220 in Monographs of the Society for Research in Child Development.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* .

Moulton, Keir. 2009a. Clausal Complementation and the Wager-class. In *Proceedings of the 38th annual meeting of the North East Linguistic Society*, ed. Anisa Schardl, Martin Walkow, and Muhammad Abdurrahman, 165–178. Amherst, MA: GLSA.

Moulton, Keir. 2009b. Natural selection and the syntax of clausal complementation. Doctoral Dissertation, University of Massachusetts, Amherst.

Murphy, Brian, Partha Pratim Talukdar, and Tom M. Mitchell. 2012. Learning Effective and Interpretable Semantic Models using Non-Negative Sparse Embedding. In *COLING*, 1933–1950.

Navarro, Daniel J., and Thomas L. Griffiths. 2008. Latent features in similarity judgments: A nonparametric Bayesian approach. *Neural computation* 20:2597–2628.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi. 2007. MaltParser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13:95–135.

Papafragou, Anna, Kimberly Cassidy, and Lila Gleitman. 2007. When we think about thinking: The acquisition of belief verbs. *Cognition* 105:125–165. URL `http://www.sciencedirect.com/science/article/pii/S0010027706002009`.

Patil, Anand, David Huard, and Christopher J. Fonnesbeck. 2010. PyMC: Bayesian stochastic modelling in Python. *Journal of statistical software* 35:1.

Pesetsky, David. 1991. Zero syntax: vol. 2: Infinitives.

Pinker, Steven. 1989. *Learnability and cognition: The acquisition of argument structure*. The MIT press.

Pinker, Steven. 1994. How could a child use verb syntax to learn verb semantics? *Lingua* 92:377–410.

Portner, Paul, and Aynat Rubinstein. 2013. Mood and contextual commitment. In *Proceedings of SALT*, volume 22, 461–487.

Portner, Paul Howard. 1992. Situation theory and the semantics of propositional expressions .

Postal, Paul M. 1993. Some defective paradigms. *Linguistic Inquiry* 347–364.

Postal, Paul Martin. 1974. *On raising: one rule of English grammar and its theoretical implications*. Current Studies in Linguistics. Cambridge, MA: MIT Press.

Prince, Ellen F. 1976. The syntax and semantics of neg-raising, with evidence from French. *Language* 404–426.

Quer, Josep. 1998. Mood at the Interface. Doctoral Dissertation, Utrecht Institute of Linguistics, OTS.

Quine, Willard Van Orman. 1960. *Word and object*. MIT press.

Rawlins, Kyle. 2013. About 'about'. In *Proceedings of the 23rd Semantics and Linguistic Theory Conference*, 336–357.

Reinhart, Tanya. 1983. Point of view in language–The use of parentheticals. In *Essays on Deixis*, ed. Gisa Rauh, volume 188, 169–194. Tübingen: Narr.

Resnik, Philip. 1996. Selectional constraints: An information-theoretic model and its computational realization. *Cognition* 61:127–159.

Ritter, Alan, and Oren Etzioni. 2010. A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 424–434. Association for Computational Linguistics.

Rizzi, Luigi. 1997. The fine structure of the left periphery. In *Elements of grammar*, 281–337. Springer.

Romoli, Jacopo. 2011. The presuppositions of soft triggers aren't presuppositions. In *Proceedings of SALT*, volume 21, 236–256.

Rooryck, Johan. 2001. Evidentiality, part I. *Glot International* 5:125–133.

Rooth, Mats. 1995. Two-dimensional clusters in grammatical relations. In *AAAI Symposium on representation and acquisition of lexical knowledge*.

Ross, John R. 1970. On declarative sentences. *Readings in English transformational grammar* 222:272.

Ross, John Robert. 1973. Slifting. In *The formal analysis of natural languages*, ed. Maurice Gross, Morris Halle, and Marcel-Paul Schützenberger, 133–169. The Hague: Mouton de Gruyter.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Cognitive modeling* 5.

Salakhutdinov, Ruslan, Andriy Mnih, and Geoffrey Hinton. 2007. Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, 791–798. ACM.

Sarle, W.S. 1979. Numerical methods for fitting an individual differences additive clustering model for similarity data. University of North Carolina, Chapel Hill.

Scheffler, Tatjana. 2009. Evidentiality and German attitude verbs. *University of Pennsylvania Working Papers in Linguistics* 15.

Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, 44–49. Citeseer.

Schütze, Carson T., and Jon Sprouse. 2014. Judgment data. In *Research Methods in Linguistics*, ed. Robert J. Podesva and Devyani Sharma, 27–50. Cambridge University Press.

Shepard, Roger N. 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika* 27:125–140.

Shepard, Roger N. 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* 27:219–246.

Shepard, Roger N. 1987. Toward a universal law of generalization for psychological science. *Science* 237:1317–1323.

Shepard, Roger N., and Phipps Arabie. 1979. Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review* 86:87.

Simons, Mandy. 2001. On the conversational basis of some presuppositions. In *Proceedings of Semantics and Linguistic Theory 11*, ed. R. Hasting, B. Jackson, and Z. Zvolensky, 431–448. Ithaca, NY: Cornell University.

Simons, Mandy. 2007. Observations on embedding verbs, evidentiality, and presupposition. *Lingua* 117:1034–1056.

Smolensky, Paul. 1986. Information processing in dynamical systems: Foundations of harmony theory .

Snedeker, Jesse. 2000. Cross-situational observation and the semantic bootstrapping hypothesis. In *Proceedings of the thirtieth annual child language research forum. Stanford, CA: Center for the Study of Language and Information.*

Snedeker, Jesse, and Lila Gleitman. 2004. Why it is hard to label our concepts. *Weaving a lexicon* 257–294.

Snedeker, Jesse, Lila Gleitman, and Michael Brent. 1999. The successes and failures of word-to-world mapping. In *Proceedings of the Twenty-third Boston University Conference on Language Development*, ed. A. Greenhill, M. Hughs, and H. Walsh. Citeseer.

Snedeker, Jesse, and John C. Trueswell. 2004. The developing constraints on parsing decisions: The role of lexical-biases and referential scenes in child and adult sentence processing. *Cognitive Psychology* 49:238–299.

Speas, Margaret. 2004. Evidentiality, logophoricity and the syntactic representation of pragmatic features. *Lingua* 114:255–276.

Speas, Peggy, and Carol Tenny. 2004. Configurational properties of point of view roles. *Asymmetry in grammar* 1:315–345.

Spector, Benjamin, and Paul Egré. 2014. A uniform semantics for embedded interrogatives: An answer, not necessarily the answer.

Stalnaker, Robert. 1973. Presuppositions. *Journal of philosophical logic* 2:447–457.

Stalnaker, Robert. 1984. *Inquiry*. Cambridge University Press.

Stevens, Stanley Smith. 1946. On the theory of scales of measurement. *Science* 103:677–680.

Stevenson, Suzanne, and Eric Joanis. 2003. Semi-supervised verb class discovery using noisy features. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, 71–78. Association for Computational Linguistics.

Stevenson, Suzanne, and Paola Merlo. 1999. Automatic verb classification using distributions of grammatical features. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, 45–52. Association for Computational Linguistics.

Sæbø, Kjell Johan. 2007. A whether forecast. In *Logic, Language, and Computation*, ed. B.D. ten Cate and H.W. Zeevat, 189–199. Verlag, Berlin, Heidelberg: Springer.

Teh, Yee W., Dilan Görür, and Zoubin Ghahramani. 2007. Stick-breaking construction for the Indian buffet process. In *International Conference on Artificial Intelligence and Statistics*, 556–563.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2006. Hierarchical dirichlet processes. *Journal of the american statistical association* 101.

Tenenbaum, Joshua B. 1996. Learning the structure of similarity. *Advances in neural information processing systems* 3–9.

Tenenbaum, Joshua B., and Thomas L. Griffiths. 2001. Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences* 24:629–640.

Truckenbrodt, Hubert. 2006. On the semantic motivation of syntactic verb movement to C in German. *Theoretical Linguistics* 32:257–306.

Trueswell, John C., Tamara Nicol Medina, Alon Hafri, and Lila R. Gleitman. 2013. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology* 66:126–156.

Trueswell, John C., Michael K. Tanenhaus, and Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language* 33:285–318.

Trueswell, John C., Michael K. Tanenhaus, and Christopher Kello. 1993. Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 19:528.

Tversky, Amos. 1977. Features of similarity. *Psychological Review* 84:327–352. URL `http://search.ebscohost.com/login.aspx?direct=true&db=pdh&AN=1978-09287-001&site=ehost-live`.

Uegaki, Wataru. 2012. Content nouns and the semantics of question-embedding predicates. *Proceedings of Sinn und Bedeutung 16* .

Urmson, James O. 1952. Parenthetical verbs. *Mind* 61:480–496.

Villalta, Elisabeth. 2000. Spanish subjunctive clauses require ordered alternatives. In *Proceedings of SALT*, volume 10, 239–256.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31:467–522.

de Villiers, Jill G. 2005. Can Language Acquisition Give Children a Point of View? In *Why language matters for theory of mind*, ed. Janet W. Astington and Jodie A. Baird, 186–219.

Vlachos, Andreas, Zoubin Ghahramani, and Anna Korhonen. 2008. Dirichlet process mixture models for verb clustering. In *Proceedings of the ICML workshop on Prior Knowledge for Text and Language*.

Vlachos, Andreas, Anna Korhonen, and Zoubin Ghahramani. 2009. Unsupervised and constrained Dirichlet process mixture models for verb clustering. In *Proceedings of the workshop on geometrical models of natural language semantics*, 74–82. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2000. Clustering verbs semantically according to their alternation behaviour. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, 747–753. Association for Computational Linguistics.

Schulte im Walde, Sabine. 2003. Experiments on the Automatic Induction of German Semantic Verb Classes. Doctoral Dissertation, Universität Stuttgart.

Schulte im Walde, Sabine. 2006. Experiments on the automatic induction of German semantic verb classes. *Computational Linguistics* 32:159–194.

Schulte im Walde, Sabine, and Chris Brew. 2002. Inducing German semantic verb classes from purely syntactic subcategorisation information. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 223–230. Association for Computational Linguistics.

Watanabe, Sumio. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research* 11:3571–3594.

Watanabe, Sumio. 2013. A widely applicable Bayesian information criterion. *The Journal of Machine Learning Research* 14:867–897.

Wexler, Kenneth N. 1970. Embedding structures for semantics. University of California, Irvine.

Wexler, Kenneth N., and Henry Hamburger. 1973. On the insufficiency of surface data for the learning of transformational languages. In *Approaches to natural language*, 167–179. Springer.

White, Aaron Steven, Rachel Dudley, Valentine Hacquard, and Jeffrey Lidz. 2014. Discovering classes of attitude verbs using subcategorization frame distributions. In *Proceedings of the 42nd annual meeting of the North East Linguistic Society*, ed. Hsin-Lun Huang, Ethan Poole, and Amanda Rysling.

Williams, Alexander. 2012. Null Complement Anaphors as definite descriptions. In *Proceedings of SALT*, volume 22, 125–145.

Williams, Alexander. 2015. *Arguments in Syntax and Semantics*. Cambridge University Press.

Wurmbrand, Susi. 2014. Tense and aspect in English infinitives. *Linguistic Inquiry* 45:403–447.

Xu, Fei, and Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological review* 114:245.

Xu, Wei, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 267–273. ACM.

Yu, Chen, and Linda B. Smith. 2012. Modeling cross-situational word–referent learning: Prior questions. *Psychological review* 119:21.

Zuber, Richard. 1983. Semantic restrictions on certain complementizers. In *Proc. of the 12th International Congress of Linguists, Tokyo*, 434–436.

Zwicky, Arnold M. 1971. In a manner of speaking. *Linguistic Inquiry* 2:223–233.

Ó Séaghdha, Diarmuid. 2010. Latent variable models of selectional preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 435–444. Association for Computational Linguistics.

Ó Séaghdha, Diarmuid, and Anna Korhonen. 2014. Probabilistic distributional semantics with latent variable models. *Computational Linguistics* 40:587–631.