



University of Pennsylvania  
**ScholarlyCommons**

---

Publicly Accessible Penn Dissertations


---

2021

## The Immediacy Of Linguistic Computation

Spencer Philip Caplan  
*University of Pennsylvania*

Follow this and additional works at: <https://repository.upenn.edu/edissertations>

 Part of the [Cognitive Psychology Commons](#), [Computer Sciences Commons](#), and the [Linguistics Commons](#)

---

### Recommended Citation

Caplan, Spencer Philip, "The Immediacy Of Linguistic Computation" (2021). *Publicly Accessible Penn Dissertations*. 4993.  
<https://repository.upenn.edu/edissertations/4993>

This paper is posted at ScholarlyCommons. <https://repository.upenn.edu/edissertations/4993>  
For more information, please contact [repository@pobox.upenn.edu](mailto:repository@pobox.upenn.edu).

---

# The Immediacy Of Linguistic Computation

## Abstract

This dissertation investigates the wide-ranging implications of a simple fact: language unfolds over time. Whether as cognitive symbols in our minds, or as their physical realization in the world, if linguistic computations are not made over transient and shifting information as it occurs, they cannot be made at all. This dissertation explores the interaction between the computations, mechanisms, and representations of language acquisition and language processing—with a central theme being the unique study of the temporal restrictions inherent to information processing that I term the immediacy of linguistic computation. This program motivates the study of intermediate representations recruited during online processing and acquisition rather than simply an Input/Output mapping. While ultimately extracted from linguistic input, such intermediate representations may differ significantly from the underlying distributional signal. I demonstrate that, due to the immediacy of linguistic computation, such intermediate representations are necessary, discoverable, and offer an explanatory connection between competence (linguistic representation) and performance (psycholinguistic behavior). The dissertation is comprised of four case studies. First, I present experimental evidence from a perceptual learning paradigm that the intermediate representation of speech consists of probabilistic activation over discrete linguistic categories but includes no direct information about the original acoustic-phonetic signal. Second, I present a computational model of word learning grounded in category formation. Instead of retaining experiential statistics over words and all their potential meanings, my model constructs hypotheses for word meanings as they occur. Uses of the same word are evaluated (and revised) with respect to the learner's intermediate representation rather than to their complete distribution of experience. In the third case study, I probe predictions about the time-course, content, and structure of these intermediate representations of meaning via a new eye-tracking paradigm. Finally, the fourth case study uses large-scale corpus data to explore syntactic choices during language production. I demonstrate how a mechanistic account of production can give rise to highly "efficient" outcomes even without explicit optimization. Taken together these case studies represent a rich analysis of the immediacy of linguistic computation and its system-wide impact on the mental representations and cognitive algorithms of language.

## Degree Type

Dissertation

## Degree Name

Doctor of Philosophy (PhD)

## Graduate Group

Linguistics

## First Advisor

Charles D. Yang

## Second Advisor

John C. Trueswell

## Keywords

Cognitive Modeling, Language Acquisition, Language Processing, Linguistics, Psycholinguistics

---

**Subject Categories**

Cognitive Psychology | Computer Sciences | Linguistics

THE IMMEDIACY OF LINGUISTIC COMPUTATION

Spencer Caplan

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2021

Supervisor of Dissertation

Co-Supervisor of Dissertation

---

Charles Yang, Professor of Linguistics

---

John C. Trueswell, Professor of Psychology

Graduate Group Chairperson

---

Eugene Buckley, Associate Professor of Linguistics

Dissertation Committee

Mitchell P. Marcus, RCA Professor of Artificial Intelligence



THE IMMEDIACY OF LINGUISTIC COMPUTATION

© COPYRIGHT

2021

Spencer Philip Caplan

*For Julia/n*

.....

*I'd say you make a perfect Angel in the snow*

*All crushed out on the way you are*

*Better stop before it goes too far*

*Don't you know that I love you*

## ACKNOWLEDGEMENT

A dissertation is the culmination of a lot of work and a lot of growth. On both fronts I have an immense amount to be thankful for. Yet, as I write this I can't help but feel a bit conflicted. This bookends a major chapter in my life, both personally and academically; and I will always be tremendously grateful for the wonderful opportunities and conversations I've had at Penn over the last six years. But there's also an element of bittersweetness for me (you're only a graduate student once after all!), particularly as I reflect on an environment that will never again be quite as I remember it. Nonetheless, many acknowledgments are in order.

First among those whom I'd like to thank for guiding and supporting me throughout my time at Penn is my exceptional team of advisors: Charles Yang, Mitch Marcus, and John Trueswell—my triumvirate, the computational psycholinguistics *dream team*. I always felt “built up” by you all: fostering great confidence for the things I could go on to study, while always pushing me to improve at my shortcomings. I was always given high expectations but never demands; freedom to pursue whatever path I chose, but always a patient ear to guide me towards the *real* questions, wherever they were hiding. I'm certain that I would have been far worse off had I chosen to study elsewhere.

I am immensely thankful for the time I spent in conversation/debate with Tony Kroch, who taught me as much as anyone else in grad school. I could always express exactly what I was thinking to Tony, and I'd get nothing but the same intensity and earnestness right back. I would also like to thank Lila Gleitman—whose depth of knowledge and infectiously invested attitude were unparalleled in psycholinguistics—for many wonderful comments and discussions during lab meetings.

I have co-authored papers with a number of people during my time here, all of whom deserve mention: Deniz Beser, Kajsa Djärv, Alon Hafri, Jordan Kodner, Mitch Marcus, Katie Schuler, John Trueswell, Hongzhi Xu, and Charles Yang.

My time at Penn was also made much brighter by a host of friends, both near and far. Worthy of particular mention is Jordan Kodner, my Comrade in StarLab and academic double-sibling. The CIS- and Ling-fueled journey was far more fulfilling taken together. Thank you to Doug Guilbeault, with whom I've shared countless laughs and an even greater countless number of stimulating conversations (“*soak to squeeze, cherish to yearn, 8 to 11 jug milk, 7 days*”). A big thank you to Patrick O’Callahan and Billy Shinevar for our continued weekly correspondence over the last six years: the Adorno-inspired, Zohar-curious, Freudo-Marxist reading group was as good an intellectual outlet as it gets, and in many ways represents an instantiation of the academic ideal. I would like to acknowledge all my classmates in the cohort of 2015 (the *Smartbeginners*) and the licorice-fueled DARPA LORELEI team. Additional thanks to Faruk Akkuş, Ryan Budnick, Andrea Ceolin, Victor Gomes, Alex Kalomoiros, Steve O’Neill, Vichet Ou, Zack Wiener, Hongzhi Xu, and many others not listed here. And lastly, thank you to Hannah Brooks, whose radiant and persistent kindness is so rare in this world and always appreciated.

I would like to thank everyone I’ve known through the Ballroom dance community, as dance was a particularly helpful outlet to escape the stresses of graduate life. In particular I’d like to acknowledge my dance partners Maria Peifer and Alexa Gamburg for hundreds of rewarding hours of practice, training, co-teaching, and competition; as well as my coaches Emanuele Pappacena and Francesca Lazzari for never letting my ego get too big.

Finally, I would like to thank my parents, who taught me to learn independently and never offered anything less than their unconditional support. And to Julian, you still have such an impact on me. I only wish you would have been able to read this and tell me it’s beautiful, or tell me it’s shit :)

— Thanks for the ride!

## ABSTRACT

### THE IMMEDIACY OF LINGUISTIC COMPUTATION

Spencer Caplan

Charles Yang

John C. Trueswell

This dissertation investigates the wide-ranging implications of a simple fact: language unfolds over time. Whether as cognitive symbols in our minds, or as their physical realization in the world, if linguistic computations are not made over transient and shifting information as it occurs, they cannot be made at all. This dissertation explores the interaction between the computations, mechanisms, and representations of language acquisition and language processing—with a central theme being the unique study of the temporal restrictions inherent to information processing that I term the *immediacy of linguistic computation*. This program motivates the study of *intermediate representations* recruited during online processing and acquisition rather than simply an Input/Output mapping. While ultimately extracted from linguistic input, such intermediate representations may differ significantly from the underlying distributional signal. I demonstrate that, due to the immediacy of linguistic computation, such intermediate representations are necessary, discoverable, and offer an explanatory connection between competence (linguistic representation) and performance (psycholinguistic behavior). The dissertation is comprised of four case studies. First, I present experimental evidence from a perceptual learning paradigm that the intermediate representation of speech consists of probabilistic activation over discrete linguistic categories but includes no direct information about the original acoustic-phonetic signal. Second, I present a computational model of word learning grounded in category formation. Instead of retaining experiential statistics over words and all their potential meanings, my model constructs hypotheses for word meanings as they occur. Uses of the same word

are evaluated (and revised) with respect to the learner’s intermediate representation rather than to their complete distribution of experience. In the third case study, I probe predictions about the time-course, content, and structure of these intermediate representations of meaning via a new eye-tracking paradigm. Finally, the fourth case study uses large-scale corpus data to explore syntactic choices during language production. I demonstrate how a mechanistic account of production can give rise to highly “efficient” outcomes even without explicit optimization. Taken together these case studies represent a rich analysis of the immediacy of linguistic computation and its system-wide impact on the mental representations and cognitive algorithms of language.

## TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iv
ABSTRACT . . . . .	vi
LIST OF TABLES . . . . .	xv
LIST OF ILLUSTRATIONS . . . . .	xix
CHAPTER 1 : Introduction . . . . .	1
1.1 Outline of the Dissertation . . . . .	2
1.1.1 Study 1: The Intermediate Representation of Speech . . . . .	2
1.1.2 Study 2: Word Learning as Category Formation . . . . .	3
1.1.3 Study 3: A More Direct Probe of Intermediate Representations during Word Learning . . . . .	5
1.1.4 Study 4: Choices in Language Production . . . . .	6
1.2 In Sum . . . . .	7
CHAPTER 2 : The Immediacy of Linguistic Computation and the Representation of Speech . . . . .	9
2.1 Experiment 1 . . . . .	12
2.1.1 Design . . . . .	12
2.1.2 Participants . . . . .	13
2.1.3 Stimuli . . . . .	14
2.1.4 Procedure . . . . .	15
2.1.5 Predictions . . . . .	17
2.1.6 Exclusions . . . . .	18
2.1.7 Analysis . . . . .	18

2.1.8	Results	19
2.1.9	Interim Discussion	23
2.2	Experiment 2	23
2.2.1	Design	24
2.2.2	Participants	24
2.2.3	Stimuli	25
2.2.4	Procedure	25
2.2.5	Results	26
2.3	General Discussion	30
CHAPTER 3 : Word Learning as Category Formation		34
3.0.1	Word Learning and Generalization	35
3.0.2	Algorithms and Rational Behavior	37
3.0.3	Organization of the Chapter	39
3.1	Models, Experiments, and Major Findings in Generalization	40
3.1.1	Word Learning as Bayesian Inference	40
3.1.2	Immediate Generalization Paradigm	41
3.1.3	Experimental Phenomena	44
3.2	Robustness of the Presentation-Style Effect	46
3.2.1	Analyzing data from Lewis and Frank	47
3.2.2	Presentation-Style and Learning in Similar Domains	50
3.3	Naïve Generalization Model	51
3.3.1	Features	52
3.3.2	Learning	55
3.3.3	Computing distances	58
3.4	Modeling Results	60
3.4.1	Scoring	60
3.4.2	Parameter-independent Evaluation	61
3.4.3	Parameter-tuned Evaluation	62



3.5	General Discussion . . . . .	66
CHAPTER 4 : Selective Attention and the Intermediate Representation of Word		
	Meanings . . . . .	69
4.0.1	Design Constraints on Word Learning . . . . .	70
4.0.2	Hypothesis Generation vs. Evaluation . . . . .	73
4.1	Experiment . . . . .	74
4.1.1	Design . . . . .	74
4.1.2	Stimuli . . . . .	77
4.1.3	Procedure . . . . .	80
4.1.4	Measures for Analysis . . . . .	80
4.1.5	Exclusions . . . . .	82
4.1.6	Predictions . . . . .	83
4.2	Results . . . . .	84
4.2.1	Effect of Timing on generalization . . . . .	84
4.2.2	Relationship between eye-gaze and learning outcome . . . . .	85
4.3	General Discussion . . . . .	90
CHAPTER 5 : The Incremental Mechanisms of Functional Design: Language Pro-		
	duction and the Immediacy of Computation . . . . .	92
5.1	Language Production . . . . .	94
5.2	Verb-Particle Construction . . . . .	98
5.2.1	Data Extraction . . . . .	99
5.2.2	Data Quality . . . . .	101
5.3	IG Predictions on the Verb-Particle Construction . . . . .	102
5.3.1	Frequency . . . . .	104
5.3.2	Predictability . . . . .	105
5.3.3	Definiteness . . . . .	106
5.3.4	Object Length . . . . .	107

5.3.5	Prior Mention . . . . .	107
5.4	Primary Model . . . . .	108
5.5	Efficiency, Optimization and Uniform Information Density . . . . .	109
5.5.1	UID and Levels of Analysis . . . . .	112
5.5.2	UIDA . . . . .	114
5.6	Object Length Experiment between IG and UIDA . . . . .	115
5.7	General Discussion . . . . .	118
CHAPTER 6 : Conclusions . . . . .		121
APPENDIX A: Supplemental Material for Chapter 2 . . . . .		124
A.1	Stimulus Lists . . . . .	124
A.2	Full Null Model Structures for Mixed Effects Regression Analyses . . . . .	125
A.3	Full Regression Outputs for Best Fitting Models . . . . .	127
A.4	Bayes Factor Calculation . . . . .	130
A.5	Secondary Analyses . . . . .	133
A.6	Effect of Pre-Registered Exclusion Criteria for All Experiments . . . . .	137
A.7	Distribution of Participant Exclusions . . . . .	140
A.8	Visualizing three-way interactions in main experiments . . . . .	141
A.9	Norming study . . . . .	143
A.10	Experiment S1 . . . . .	145
APPENDIX B: Supplemental Material for Chapter 3 . . . . .		151
B.1	Parameter Sensitivity . . . . .	151
B.2	Gradient analysis of PSE in Lewis and Frank (2018) . . . . .	151
APPENDIX C: Supplemental Material for Chapter 4 . . . . .		153
C.1	Nonce word labels . . . . .	153
C.2	Possible Feature Alternations . . . . .	153
C.3	Gaze Heatmaps . . . . .	155

C.4 Timecourse Plots . . . . .	157
BIBLIOGRAPHY . . . . .	157

## LIST OF TABLES

TABLE 1 :	Output of the best fitting model predicting /t/ responses on the first half of test trials for Experiment 1. Bracketed values are 95% confidence intervals. . . . .	21
TABLE 2 :	Output of the best fitting model predicting /t/ responses on the first half of test trials for Experiment 2. Bracketed values are 95% confidence intervals. . . . .	29
TABLE 3 :	Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization on all trials. Mixed-effects logistic regression predicting generalization based on listed effects as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects along with a three-way interaction between Presentation-Style, Training-Number, and Block-Order. . . .	48
TABLE 4 :	Mixed-effects logistic regression predicting generalization outcome on second-block trials (data from Lewis and Frank (2018)) based on listed effects (Presentation-Style, Training-Number, Number-Timing Interaction) as well as random slopes for each subject and stimulus class. Neither SCE nor PSE manifest on second-block trials. . . . .	49
TABLE 5 :	Mixed-effects logistic regression predicting generalization outcome on first-block trials (data from Lewis and Frank (2018)) based on listed effects (Presentation-Style, Training-Number, Number-Timing Interaction) as well as random slopes for each subject and stimulus class. PSE and SCE emerge as significant main effects. . . . .	49
TABLE 6 :	In this toy example, the initial training set contains a single instance of a dalmatian with features (A:1, B:1, C:1, D:0). From this, the learner extracts a mental representation of (A:0.3, B:0.8, C:0.3, D:0). During testing, a few potential items are all compared against mental representation in order to select category members. Only values present in mental representation but missing from the evaluated items incur a penalty. If the maximum category cutoff were 1.0, then both the dalmatian and the poodle (shown with shaded background) would be selected in this case. . . . .	60
TABLE 7 :	Major patterns to be captured by models of word learning and generalization. Both the size of the training set (SCE) as well as the temporal manner of presentation (PSE) have reliable effects on the meanings posited by learners. “0.15” represents the typical standard deviation from results in Spencer et al. (2011) . . . . .	62
TABLE 8 :	Output of the best fitting model predicting Narrow generalization. Bracketed values are 95% confidence intervals. . . . .	84
TABLE 9 :	Output of primary logistic regression model where the dependent variable was particle-first order. . . . .	109

TABLE 10 :	Evaluating cases of N=2 or more words . . . . .	118
TABLE 11 :	Evaluating cases of N=4 or more words. The effect of conditional probability is absent, while the effects of frequency, object length, and definiteness remain. . . . .	118
TABLE 12 :	Target stimulus pairs used in Experiments 1, 2, and S1. . . . .	124
TABLE 13 :	Filler stimuli used in Experiments 1, 2, and S1. . . . .	125
TABLE 14 :	Output of the best fitting model on all trials for Experiment 1 . . . . .	127
TABLE 15 :	Output of the best fitting model on the first half of test trials for Experiment 1 . . . . .	128
TABLE 16 :	Output of the best fitting model on the last half of test trials for Experiment 1 . . . . .	128
TABLE 17 :	Output of the best fitting model on the first half of test trials, text-before condition for Experiment 1 . . . . .	128
TABLE 18 :	Output of the best fitting model on the first half of test trials, text-after condition for Experiment 1 . . . . .	128
TABLE 19 :	Output of the best fitting model on all trials for Experiment 2 . . . . .	128
TABLE 20 :	Output of the best fitting model on the last half of test trials for Experiment 2 . . . . .	129
TABLE 21 :	Output of the best fitting model on the first half of test trials, text-before condition for Experiment 2 . . . . .	129
TABLE 22 :	Output of the best fitting model on the first half of test trials, text-after condition for Experiment 2 . . . . .	129
TABLE 23 :	Exclusions with ceiling/floor cutoff for each experiment (by condition).137	
TABLE 24 :	Tuned parameter values from Section 3.4.3 . . . . .	151
TABLE 25 :	Data from Lewis and Frank (2018). Dependent variable is the generalization-level outcome on all trials. Linear mixed model predicting generalization based on listed effects as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects along with a three-way interaction between Presentation-Style, Training-Number, and Block-Order. . . . .	152
TABLE 26 :	Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization proportion on second-block trials. Linear mixed model predicting generalization based on presentation-style, training-number, the presentation-number interaction, as well as random slopes for subject and stimulus class. Neither SCE nor PSE manifest on second-block trials. . . . .	152
TABLE 27 :	Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization on first-block trials. Linear mixed model predicting generalization based on presentation-style, training-number, the presentation-number interaction, as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects. . . . .	152
TABLE 28 :	Disyllabic nonce word labels used in Experiment 1 (Chapter 4) . . . . .	153

TABLE 29 : Potential feature alternations for each domain. . . . . 154

## LIST OF ILLUSTRATIONS

FIGURE 1 :	Timeline of the main experimental manipulation. Participants were provided with disambiguating text either before (a) or after (b) hearing the corresponding audio. . . . .	12
FIGURE 2 :	Design of the exposure and test phases in both experiments. Each participant was assigned to one of four possible conditions during the exposure phase (a), which had a 2 x 2 design: shifted phone (/d/ or /t/) and audio–text order (text before or text after). All participants then completed the same task at test (b), categorizing audio on a continuum of voice-onset time (VOT) as either “ta” or “da.” The graph illustrates predicted categorization patterns (separately for each shifted-phone condition) in cases in which adaptation occurs.	14
FIGURE 3 :	Pairing of text and audio used in Experiments 1 and 2 in the shifted-/d/ and shifted-/t/ conditions. Although all participants were exposed to the same text, participants in the shifted-/d/ condition heard audio with ambiguous voice-onset times (VOTs) paired with “d” text, whereas participants in the shifted-/t/ condition heard audio with ambiguous VOTs paired with “t” text. . . . .	17
FIGURE 4 :	Psychometric functions for Experiment 1: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals. Adaptation occurred in the text-before condition, but did not occur in the text-after condition. . . . .	20
FIGURE 5 :	Psychometric functions for Experiment 2: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals. Adaptation occurred in the text-before condition, but did not occur in the text-after condition. . . . .	28
FIGURE 6 :	Example word learning trial with test-grid shown to participants in Xu and Tenenbaum (2007b); Spencer et al. (2011); Lewis and Frank (2018). Figure adapted from Spencer et al. (2011). . . . .	43
FIGURE 7 :	Proportion of broad property-projections based on sample size and presentation-style (figure reproduced from Lawson (2014b) with permission). When presented simultaneously, the size of training has no effect on projection. When presented in sequence, the rates of broad property-projection quickly approach a ceiling condition as the size of training increases. . . . .	51

FIGURE 8 :	Computation of mental representation from single training example and subsequent comparison to test objects. Values are schematic and for illustration only. . . . .	55
FIGURE 9 :	Algorithmic flow charts highlighting some possible paths of NGM behavior. This illustrates the common difference in experimental outcome under parallel (left) and sequential (right) presentation of stimuli. . . . .	57
FIGURE 10 :	Implementation of distance computation between an object and mental representation under the NGM . . . . .	59
FIGURE 11 :	Chart of all seven training configurations. Conditions used for parameter tuning shown in light red. Time during training is indicated within each block vertically; the objects in the parallel condition are co-present at the same time, while the “sequential” trials training objects are never co-present. . . . .	63
FIGURE 12 :	Training on a single item. Experimental results from Spencer et al. (2011) are shown in gold. Output of NGM in grey. Bars indicate standard deviations. . . . .	64
FIGURE 13 :	Training items presented in sequence. Experimental results in gold. Output of NGM in grey. Bars indicate standard deviations. . . . .	64
FIGURE 14 :	Training items presented simultaneously. Experimental results in gold. Output of NGM in grey. Bars indicate standard deviations. . . . .	64
FIGURE 15 :	Visualization of parallel vs. sequential training conditions. Word <sub>1</sub> in red and Word <sub>2</sub> in blue. The total number of exemplars and display time remained constant across conditions. . . . .	76
FIGURE 16 :	Sample pairs of maximally divergent stimuli (differing on all five features) for each domain. . . . .	79
FIGURE 17 :	Example AOI calculation. The RF might be the front and the bottom in blue. While the NF are the tail and the top in red. . . . .	82
FIGURE 18 :	Bar graph of proportion of learning outcomes as a function of training condition (parallel vs. sequential) . . . . .	85
FIGURE 19 :	Violin plot of the proportion of gaze-time allocated to RFs vs. NFs as a function of Learning Outcome (learned vs. mislearned) . . . . .	87
FIGURE 20 :	Violin plot showing RF-Skew as a function of learning outcome (Narrow vs. Broad) . . . . .	88
FIGURE 21 :	Gaze to posited features is not affected by learning outcome (learned vs mislearned). . . . .	89
FIGURE 22 :	Gaze to RF-set during each training exposure. Participants, in aggregate, are likely to converge on their initial hypothesis. This figure includes trials from both parallel- and sequential-participants—however, for parallel-participants the first three “trials” were actually objects on the screen simultaneously. . . . .	90
FIGURE 23 :	Basic outline of the language production architecture—adapted from Bock and Levelt (2002). Time is indicated from left-to-right. . . . .	95



FIGURE 24 :	Example illustration of IG output of verb-particle construction. Variations in lemmas access or constituent assembly speed manifest as comparable variations in linear order (when permitted by the grammar): whichever element is retrieved and constructed first is sent off to positional processing first. . . . .	103
FIGURE 25 :	Distribution of object length (in words) within the present sample of verb-particle data (67,905 sentences) . . . . .	117
FIGURE 26 :	Distributions of the 50% categorization thresholds in Experiment 1 from the main manuscript. No evidence for bimodality was observed in any condition. . . . .	133
FIGURE 27 :	Distributions of the 50% categorization thresholds in Experiment 2 from the main manuscript. No evidence for bimodality was observed in any condition. . . . .	134
FIGURE 28 :	There is no significant relationship between RTs and categorization thresholds at test . . . . .	135
FIGURE 29 :	There is no significant relationship between RTs and categorization thresholds at test . . . . .	136
FIGURE 30 :	Psychometric functions for phoneme categorization during testing in Experiment 1. Output split by Shifted Phone (/t/ or /d/), Timing condition (text-before or text-after), and test-phase half (first four test blocks or remaining five). Adaptation occurred in the text-before but not text-after condition and faded over the course of the test phase (first vs. last half). Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008).	141
FIGURE 31 :	Psychometric functions for phoneme categorization during testing in Experiment 2. Output split by Shifted Phone (/t/ or /d/), Timing condition (text-before or text-after), and test-phase half (first four test blocks or remaining five). Adaptation occurred in the text-before but not text-after condition and faded over the course of the test phase (first vs. last half). Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008).	142
FIGURE 32 :	Violin plot of the median 50% threshold for “t” / “d” categorization for each continuum in the norming study. Red line shows the overall median 50% threshold at 46.9ms. “_SHIFTED” and “_ORIG” correspond to pitch-edited and original-pitch CV continua respectively. . . . .	145
FIGURE 33 :	Assignment of audio to text in Experiment S1. There is a confound between “edited speech” and the particular phonological category under manipulation. . . . .	146
FIGURE 34 :	Main results for Experiment S1. Psychometric functions for phoneme discrimination during testing. Output split by Shifted Phone (/t/ or /d/) and Timing condition (text-before or text-after). Unlike in Experiments 1 and 2, adaptation occurred in both the text-before and text-after conditions. Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008). . .	148

FIGURE 35 : Heatmap of gaze (within stimulus bounding box) throughout all training trials and all stimulus domains. . . . .	155
FIGURE 36 : Heatmaps of overall gaze split by domain and overlaid on example stimulus . . . . .	156
FIGURE 37 : Plots showing timecourse of gaze-time to the RF-set as a function of learning outcome . . . . .	157

## CHAPTER 1 : Introduction

One the fundamental question of linguistics asks “what does a person *know* when they know a language?” What are the mental representations that underlie our cognitive system of linguistic meaning, how do we learn them, and how are they processed in real time? To address these question, this dissertation investigates the wide ranging implications of a simple fact: language unfolds over time.

Whether as cognitive symbols in our minds, or as their physical realization in sound waves and ever-changing referents in the world, if linguistic computations are not made over transient and shifting information as it occurs, they cannot be made at all. This dissertation explores the interaction between the computations, mechanisms, and representations of language acquisition and language processing — with a central theme being the unique study of the temporal restrictions inherent to information processing that I term the *immediacy of linguistic computation*. This program motivates the study of *intermediate representations* recruited during online processing and acquisition rather than simply an Input/Output mapping. While ultimately extracted from linguistic input, such intermediate representations may differ significantly from the underlying distributional signal. I demonstrate in several lines of work that, due to the immediacy of linguistic computation, such intermediate representations are necessary, discoverable, and offer an explanatory connection between competence (linguistic representation) and performance (psycholinguistic behavior).

Given the system-wide impact that temporal restrictions have on language processing and acquisition, this dissertation examines the immediacy of computation and the impact of intermediate representations through a diverse set of projects involving computational modeling, quantitative corpus analysis of language use, and psycholinguistic experimentation. In particular, I address:

1. How are linguistic hypotheses (i.e. “what phone did I just hear”, or “what does that word mean”) formed in real-time and what contents do these hypotheses comprise?

How does the process of generating hypotheses shape language compared with the statistical evaluation of those ideas? A widespread intuition is that linguistic knowledge and behavior are somehow governed by the raw computing power available to the brain, but instead I argue that temporal restrictions have a far greater impact by shaping the structure and content of hypotheses themselves<sup>1</sup>.

2. How do the algorithms implemented in our minds use simple tools to actually compute the often complicated Input/Output relationships we see in language processing and acquisition? A set of outputs is often largely consistent with many *possible* algorithms—this work attempts to identify which algorithms are most likely at play and why, disentangling causes from effects.

### 1.1. Outline of the Dissertation

The dissertation is comprised of four distinct studies. This includes a set of experiments probing the intermediate representation of speech during online processing (Chapter 2), a computational model and corresponding eye-tracking study of how learners handle semantic ambiguity during word learning (Chapters 3 and 4 respectively), and a statistical analysis of how speakers make real-time choices during language production (Chapter 5). I discuss each of these projects and the methods used to address these questions below in the remainder of Chapter 1. While the individual findings stand on their own, when taken together they represent a rich analysis of the immediacy of linguistic computation and its system-wide impact on the mental representations and cognitive algorithms of language.

#### *1.1.1. Study 1: The Intermediate Representation of Speech*

Chapter 2 answers a question in speech processing: what happens to the acoustic-phonetic signal after it enters the mind of a listener. Previous work (Connine et al., 1991, *inter alia*) demonstrates that listeners maintain intermediate speech representations over time. Successful parsing necessitates the maintenance of some sort of intermediate representation in order for listeners to use subsequent context to aide in the interpretation of prior phonetic

---

<sup>1</sup>Thus an alternative title of the dissertation might have been: “How to Get (Linguistically) Rich when you’re (Computationally) Poor”

input. Consider for instance how one would decide between the interpretation of a potentially ambiguous word pair like “[t/d]ent” in a sentence such as “That was the [t/d]ent that we saw in the forest/fender.” However, the internal structure of such representations—be they the acoustic-phonetic signal or more general information about the probability of possible categories—has remained underspecified. I present experimental evidence from a novel perceptual learning (“accent adaptation”) paradigm which supports the view that information about the acoustic-phonetic signal is not maintained over time. In particular, I exposed listeners to a speaker whose utterances contained acoustically ambiguous information concerning phones/words and manipulated the temporal availability of disambiguating cues via visually presented text (i.e., presentation before or after each utterance). Results show that listeners adapt to the modified acoustic distribution only when disambiguating text is provided *before* the auditory information, but not *after*. This finding supports the position that intermediate representations of speech consist of probabilistic activation over discrete linguistic categories (an account I call “AOC”) but not a direct record of the acoustic-phonetic signal. Such results have impactful ramifications far beyond speech processing: limits to the storage of sensory input place real limits on mental representations. This may inform longstanding debates in other areas of linguistics regarding the exemplar vs. abstract/discrete representation of phones, morphemes, syntactic units and general mental categories.

### *1.1.2. Study 2: Word Learning as Category Formation*

Children famously face ambiguity during of morphological and syntactic acquisition (Yang, 2002, 2016; Tyler and Nagy, 1989; Pinker, 1989; Rumelhart and McClelland, 1985): how does the learner deal with such ambiguity when multiple grammars are, in principle, consistent with the words and sentences they have heard (Gold, 1967)? While words, unlike syntactic units, are often thought of as atomic, a fundamental question in word learning is strikingly similar: how, given only evidence about what objects a word has previously referred to, are children able to generalize to the total class? How does a child end up knowing that “poodle” picks out a specific subset of dogs despite their overlapping extensions? Chapter 3 presents a model of word learning grounded in category formation (the

Naïve Generalization Model or “NGM”). While learners have been argued to display optimal behavior by performing statistical inference over the input distribution of their experience (e.g. via Bayesian inference—Xu and Tenenbaum (2007b)), they are also sensitive to input conditions that are orthogonal to purely statistical reasoning (Spencer et al., 2011), like the timing with which referents are encountered (for instance, whether stimuli are co-present on the screen or viewed in sequence one second apart).

I contrast the NGM with the popular Bayesian inference theory of generalization (Xu and Tenenbaum, 2007b). On the Bayesian account, learners have some representation of many potential meanings for a word, and engage in statistically sensitive calculations to select the hypothesis that is most probable given a distribution of attested exemplars. The “heavy-lifting” and explanatory power resides in *evaluation* (via statistical inference) of many hypotheses without specifying the process which generates them. In contrast with previous Bayesian (Xu and Tenenbaum, 2007b) or associative (Regier, 2005) accounts, computation in the NGM is local and lacks any global optimization over an evaluation metric. On my view, word learning is an incremental and mechanistic process. Instead of retaining experiential statistics over words and all their potential meanings, the NGM constructs hypotheses for word meanings as they occur. Uses of the same word are evaluated (and revised) with respect to the learner’s intermediate representation (e.g. their current working conception) rather than to their complete distribution of experience. While in some cases this “working conception” ends up being extremely similar to the distribution of experience, other cases lead to divergent and highly-informative outcomes, in particular when stimuli are presented sequentially rather than simultaneously. What you see (during learning) is not necessarily what you get (in subsequent mental representation).

I evaluate, and find support for, the NGM on a range of experimental data—varying the number and presentation-timing of stimuli, among other factors—on semantic generalization in word learning (Xu and Tenenbaum, 2007b; Spencer et al., 2011). Learning behavior is shaped by the immediacy of linguistic computation: learners are limited to locally eval-

uating only the fit of whatever structures they posit. Through this temporally constrained process, one hypothesis will end up winning out because it offers a satisfactory fit to the data, but this does not mean that the final meaning or grammar is provably optimal (as often assumed by alternative accounts). Learners do the best job they can, not the best job possible.

### 1.1.3. Study 3: A More Direct Probe of Intermediate Representations during Word Learning

The experiments modeled in Chapter 3 are informative as to the word representations that result from successful learning, and the NGM makes *predictions* about intermediate states of acquisition, but this does not provide direct evidence as to the fine-grained time-course over which the relevant semantic generalizations emerge. Just like the *generation/evaluation* duality, it is important for work in cognitive science to distinguish between an underlying function (intension) and measuring its output (extension) as this is frequently a many-to-one mapping. Chapter 4 introduces and presents results from a new eye-tracking paradigm (inspired by Rehder and Hoffman (2005a)) designed to test the predictions of broad classes of word learning theories: accounts grounded in hypothesis *generation* like the NGM in contrast with accounts based on the statistical *accumulation* and *evaluation* of evidence. The paradigm uses artificially created stimuli with spatially distributed features—each region uniquely corresponding to a particular semantic dimension. By using eye-gaze as a measure of selective attention to these individual features, we are able to study the content and time-course of intermediate representations as they emerge throughout learning. A statistical accumulation theory predicts that learners should initially attend to all the dimensions that they can in order to extract a representative sample before applying any evaluative filter for the most likely meaning. Conversely the NGM predicts that learners should extract an intermediate hypothesis on the basis of initial exposure; given the immediacy of linguistic computation subsequent trials are evaluated only with respect to the hypothesized meaning<sup>2</sup>.

I find that, consistent with the NGM, learners’ attention is limited only to the features

---

<sup>2</sup>Perhaps the modernist movement had it right all along! “*Nothing is less real than realism. Details are confusing. It is only by selection, by elimination, by emphasis, that we get at the real meaning of things*” (attributed to Georgia O’Keefe—as quoted in Stuhlman (2007))

present in the intermediate mental representations they have posited up to a given moment in time. This provides evidence against a framework under which learners evaluate stimuli holistically and perform an explicit optimization for the most probable meaning. Taken together, Chapters 3 and 4 highlight the utility of studying algorithm-level causal mechanisms operating within the learning process rather than high-level computational descriptions of the input and output. Effective models generate novel predictions which, when rigorously evaluated experimentally, provide much deeper insights than would be possible from either methodology alone.

#### *1.1.4. Study 4: Choices in Language Production*

Chapter 5 investigates the language production system: what mental architecture and algorithms govern the process that translates from abstract semantic thoughts into articulated utterances? A useful window into language production is the study of “syntactic optionality”; i.e. given multiple potential syntactic encodings for equivalent semantic content, what factors govern the use of one form rather than another (e.g. “Julian picked up the book” vs. “Julian picked the book up”)? According to an influential set of accounts (e.g. the Uniform Information Density Hypothesis, Jaeger (2010)), speakers’ choices are governed by a preference to distribute information evenly and efficiently over the signal. Such a notion of efficiency is, however, fundamentally a description of the output and does not specify a mechanism that underlies behavior. Adopting the classic terminology from Marr (1982), every computational-level theory necessarily requires some algorithm underlying it. While it is tempting to assume (perhaps implicitly) that some computational description was generated by the optimal, simplest, or most straightforward algorithm for generating such output, this is not a safe assumption. There is no invisible hand that makes it so. In many cases there may be a large set of naïve, local, or greedy algorithms which produce output that is very close to the optimal solution without reference to or explicit optimization over the function we describe at a computational-level.

I describe an alternative mechanism-level framework of language production (Incremental Generation or “IG”), grounded in classic findings from psycholinguistics (Bock and Levelt,



1994). Under IG, behavior is rapid and incremental: the system outputs lexical items as soon as they are retrieved (within the bounds of the syntax). Factors affecting the speed of lexical retrieval (e.g. frequency, predictability, discourse status, length, etc.) are thus predicted to have an impact of output linear order. I evaluate the predictions of this incremental framework in comparison to Uniform Information Density by performing statistical analyses over large-scale corpus data (approximately 70,000 unique test sentences) on the English verb-particle alternation. On my view, the output of the language production system is best understood as the by-product of mechanical psycholinguistic factors: whether the object (“the book up”) or the particle (“up the book”) is likely to be ordered first depends on the *intermediate* state of the system right after the output of the verb. If lexical retrieval for the object has already occurred at this point, but not retrieval of the particle, then it is the object that gets linearized first (and vice versa).

Results are consistent with the Incremental Generation framework, which I argue is the underlying causal mechanism responsible for giving rise to patterns previously seen as supporting Uniform Information Density. While aggregate language use appears to be optimizing for an information theoretic measure, directly doing so would require speakers to calculate per word entropy on the fly (which is intractable at scale). Global behavior is, in fact, the accidental by-product of a local, incremental process: the immediacy of computation reduces demands on the language producer by placing a grammatical structure in the output as quickly as possible. To whatever degree we can characterize the output of the language production system as “efficient” in information ordering (Jaeger, 2010), this is an emergent property of an incremental generation system rather than an explicit optimization performed by individual speakers. This framework serves both as an instantiation of the immediacy of linguistic computation and highlights how a simple algorithm is able to account for the sort of complicated Input/Output mapping we see in production and other systems.

## 1.2. In Sum

Taken together, these chapters represent a diverse body of work — both in content and methods — interfacing with the immediacy of linguistic computation, and addressing questions

of how language learning and language processing unfold over time. I hope this dissertation is able to serve as a general research program and philosophy which future work can build on. Cognitive science (of language and otherwise) must stay true to a tradition in which we study the computational systems responsible for individual human behavior. The mathematical tools of information theory, statistical inference, etc. have their place and, no doubt, have had utility in our development of theories of acquisition and psycholinguistics, but the identification of statistical phenomena is not an explanation of them. Above all, I emphasize that we cannot achieve a thorough understanding of language without a commitment to computational *mechanisms* and simple explanations.

## CHAPTER 2 : The Immediacy of Linguistic Computation and the Representation of Speech

My point will be that not only considered action, but also learning and perception, must surely be viewed as based upon computational processes; and, once again, no computation without representation.

---

Jerry Fodor. *The Language of Thought*

Variability is a constant in the world.<sup>1</sup> How cognitive systems represent and process input signals to adapt to such a gradient and shifting landscape is a classic problem in cognitive science ranging from learning and decision making (e.g. Erev and Barron, 2005; Gallistel, 1990) to plasticity in visual processing (e.g. Postle, 2015; Sagi, 2011). In this regard, language represents an ideal domain to study the *structure* of mental representations built up in real-time, and what type of information is thus available for learning. This chapter investigates these questions through the lens of speech processing, asking: How do listeners convert a gradient and variable acoustic signal into cognitive units like phones and words in order to reconstruct the underlying meaning? What happens to the acoustic-phonetic signal after it enters the mind of a listener?

Language unfolds over time. Unlike in reading or visual search, an acoustic signal is inherently ephemeral: if cognitive computations are not made over transient and shifting information as it occurs, they cannot be made at all. This inherent constraint, which I term the *Immediacy of Linguistic Computation*, means that listeners cannot and do not wait until the end of an utterance to begin building a representation of speech (Christiansen and Chater, 2016; *see*. Marslen-Wilson and Tyler, 1980). Thus a design feature of all models of speech perception (Marslen-Wilson and Welsh, 1978; McClelland and Elman, 1986, *inter alia*) is the real-time construction of intermediate representations — that is, representations

---

<sup>1</sup>This chapter represents work that was originally co-authored with Alon Hafri and John Trueswell; published as Caplan et al. (2021)

held in memory (irrespective of content) that outlast the stimulus itself but which may be integrated with additional information over time. This intermediate structure serves as a listener’s working hypothesis for recognition, but given the Immediacy of Computation, the form of these representations is a bottleneck: computation can occur only over the material that is constructed rather than the original, ephemeral signal.

Real-time processing involves extracting and integrating linguistic evidence from varied sources, with disambiguating information arriving in the form of multiple, temporally disjoint cues, e.g. visual articulatory cues (McGurk and MacDonald, 1976), prior lexical knowledge (Ganong, 1980), etc. Speech processing is thus a problem of handling and representing uncertainty. Experimental evidence shows that listeners maintain and update intermediate representations over time, both locally (Galle et al., 2019) and over long-distances (Bushong and Jaeger, 2017; Connine et al., 1991; Zellou and Dahan, 2019). However, while claims in the literature (e.g. Bicknell et al., 2016; Darwin and Baddeley, 1974; Galle et al., 2019) are varied, such work on long-distance cue integration has not directly addressed the *structure* of information included in these intermediate representations and how this is recruited for adapting to variability. I contrast two classes of theories.

Under a “*signal retention*” account, listeners maintain acoustic-phonetic detail (e.g. Bicknell et al., 2016; Goldinger, 1998; McMurray et al., 2009). This would include information like acoustic cues, among other properties, for example: “*acoustic memory would thus be viewed as an initially tape-recording-like representation of the stimulus*” (Darwin and Baddeley, 1974), and more recently “*data from speech perception and sentence processing, however, demonstrate that comprehenders can maintain fine-grained lower-level perception information for substantial durations*” (Bicknell et al., 2016). A second family of accounts—which I develop here—I term the “*Activation over Categories*” (AOC) theory. Under AOC, listeners maintain a graded activation pattern over some set of cognitive/linguistic categories (phones, words, etc.). Crucially, this is a *Markovian* process: listeners encode a state of activation but do not retain the precise sensory evidence which led to that belief. These states of ac-

tivation can be understood as “predictions” which are updated by later linguistic input and thus support learning variation. Such a scenario is analogous to the mechanisms underlying visual object detection (Marr, 1982): an observer of a scene constructs a number of stages of visual representation but does not have access to the raw primitives (e.g. zero-crossings) which were used to construct these visual sketches. Phonetic information is recruited for identifying higher-level categories but is not stored or isolable within the speech processing system. Past work interpreted as evidence for maintenance of acoustic detail (Bushong and Jaeger, 2017; Connine et al., 1991; Crowder and Morton, 1969; Frankish, 2008; McMurray et al., 2009) is also compatible with AOC because AOC maintains gradience through probabilistic information about linguistic categories, not the acoustic details that gave rise to those probabilities. Such debate between these general accounts of mental representations are pervasive across psychology; for example, in the exemplar vs. abstract representations of concepts and categories (Schuler et al., 2020; Smith and Medin, 2013, *inter alia*).

To investigate the contents of intermediate speech representations and evaluate the predictions of signal-retention against AOC, I looked at how people adapt to shifts in speech when disambiguating information appears *after* the original signal rather than *before*. I present findings from two experiments using a novel variant of the “accent-adaptation” paradigm (more specifically a type of “Lexically-Guided Perceptual Learning”) (Norris et al., 2003; Samuel and Kraljic, 2009): in this paradigm, after encountering a series of target words with a manipulated distribution over an acoustic cue to some phone, participants subsequently exhibit shifted criteria for categorizing phones, e.g. /t/ vs. /d/ (Bertelson et al., 2003; Clayards et al., 2008; Jesse and McQueen, 2011; Kraljic and Samuel, 2006; Reinisch and Holt, 2014). In the current study, I exposed participants to acoustically ambiguous audio via minimal pairs (e.g. “time/dime”). Disambiguation was provided by a text subtitle that appeared either briefly before or after the audio and systematically biased the ambiguous audio to be interpreted either as /t/ or /d/ (Figure 1; see Section 2.1.1 for complete details).

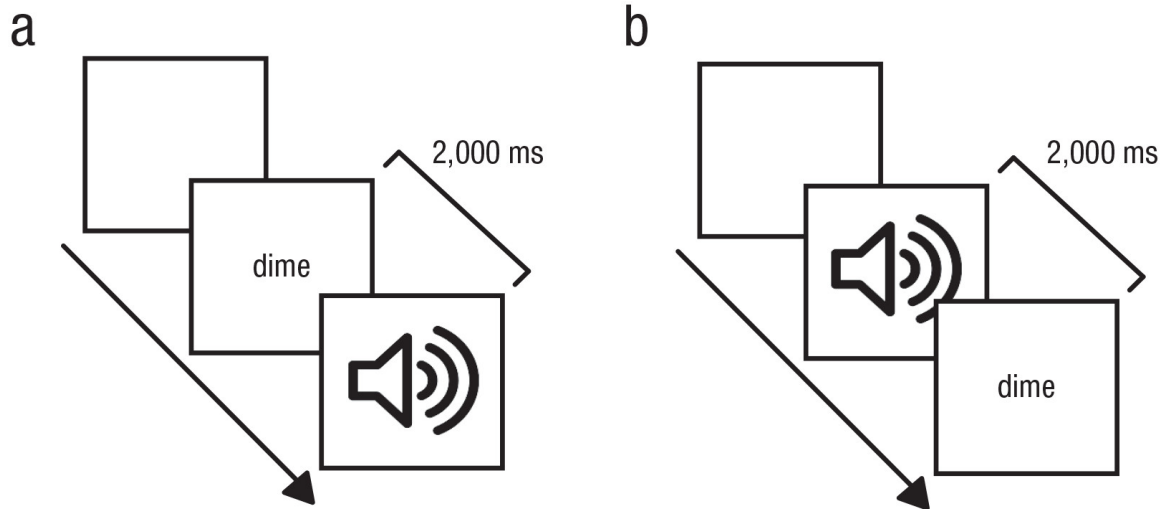


Figure 1: Timeline of the main experimental manipulation. Participants were provided with disambiguating text either before (a) or after (b) hearing the corresponding audio.

When the disambiguating text is provided before, both signal-retention and AOC predict that participants should adapt to the shifted phonetic distribution. When reading the word first, participants know the intended phones ahead of time and can evaluate the upcoming ambiguous audio accordingly: the signal can be evaluated given the prior hypothesis. When the text is provided *after* the audio, then only the signal-retention account predicts adaptation to occur (maintenance of the phonetic-detail is the central tenet of the theory). AOC conversely predicts no adaptation, since while the graded activation over /t/ and /d/ allows for the proper lexical interpretation once text arrives, the reason for that particular activation state is lost and so there is no pattern to generalize.

## 2.1. Experiment 1

### 2.1.1. Design

The experiment had a 2 (shift direction: shifted-/d/ vs. shifted-/t/) x 2 (timing: text-before vs. text-after) between-subjects design. During the exposure phase, participants heard and saw a sequence of 142 words presented once in a random order. Exposure words were divided between 44 Target items (22 “t”-onset and 22 “d”-onset) and 98 Fillers. Each Target word was paired with corresponding audio that had an ambiguous (60ms) or unambiguous (10ms for “d”-words and 100ms for “t”-words) onset Voice Onset Time (VOT)—the time delay

between the release of a stop consonant and the onset of glottal pulses from the closed vocal folds. VOT is the primary acoustic cue for distinguishing voiced stops (e.g. /b/, /d/, and /g/) from their voiceless counterparts (/p/, /t/, and /k/). The ambiguous vs. unambiguous mapping was controlled by the shift-direction condition: “t”-words paired with ambiguous VOT for the shifted-/t/ group and “d”-words paired with ambiguous VOT for the shifted-/d/ group. Since I used a fully crossed design, each shift-direction occurred with a timing manipulation of getting the subtitle two seconds before the audio (text-before) or two seconds after the audio (text-after).

In previous studies (Jesse and McQueen, 2011; Kraljic and Samuel, 2006) the interpretation of manipulated audio under an accent-adaptation paradigm was provided by local lexical context (e.g. only one interpretation of “croco[t/d]ile” results in a real word). However, adaptation induced by lexical context is not informative to the structure of intermediate representations as listeners can resolve the [t/d] ambiguity locally, regardless of their ability to store phonetic detail. I explicitly removed information needed to disambiguate words internally by using minimal pairs: words which differ in exactly one phoneme. This is similar to distributional approaches to adaptation (Clayards et al., 2008; Munson, 2011), except that my method does not require hearing a large number of repeated tokens and allows for the direct manipulation of disambiguation timing.

The test-phase was identical for all participants. On each test trial, participants heard a syllable beginning with an alveolar stop consonant with a particular VOT (ranging from 20 to 80ms, order randomized) and they were asked to judge whether they heard a /t/ or a /d/. The design for the exposure and test phases is is schematized in Figure 2.

### *2.1.2. Participants*

I recruited 132 University of Pennsylvania undergraduates who received course credit for their participation. All participants were native English speakers with no reported hearing or visual impairments. As was planned in the preregistration of this experiment, the final sample consisted of 128 participants after exclusion (see criteria below) and is in line with

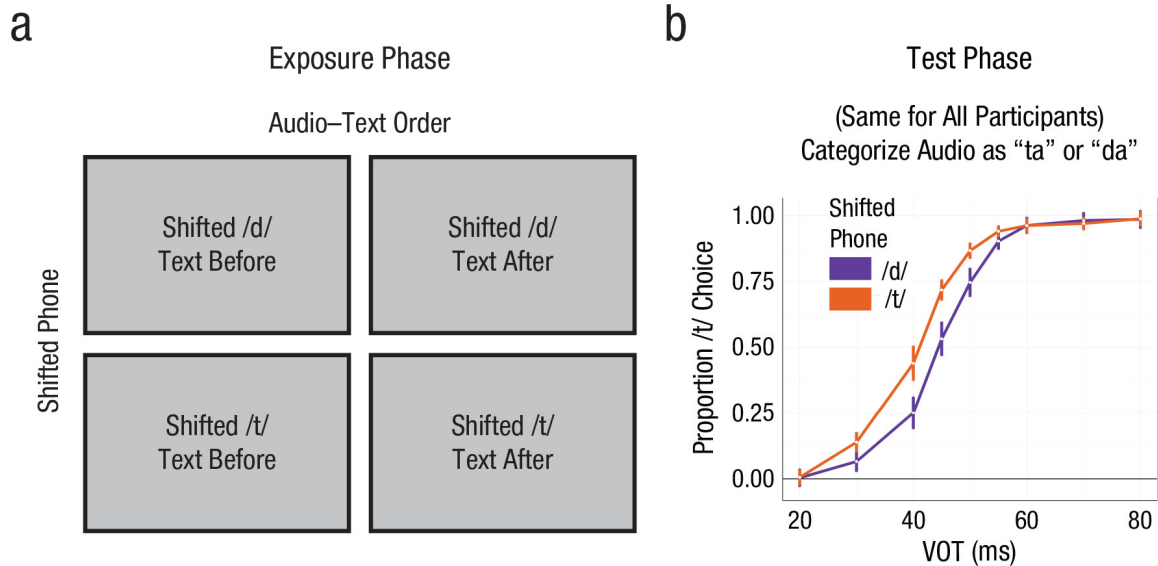


Figure 2: Design of the exposure and test phases in both experiments. Each participant was assigned to one of four possible conditions during the exposure phase (a), which had a 2 x 2 design: shifted phone (/d/ or /t/) and audio–text order (text before or text after). All participants then completed the same task at test (b), categorizing audio on a continuum of voice-onset time (VOT) as either “ta” or “da.” The graph illustrates predicted categorization patterns (separately for each shifted-phone condition) in cases in which adaptation occurs.

previous studies measuring similar effects (Kraljic and Samuel, 2006). Participants were approximately evenly divided between the four different exposure conditions, with test stimuli held constant across all participant groups.

### 2.1.3. Stimuli

Target words for the exposure phase were selected by identifying minimal pairs in CELEX (Baayen et al., 1995) which are differentiated solely by an onset position /t/ vs. /d/. This resulted in a list of 82 such minimal pairs, from which I manually selected 44 words (22 pairs) based on part of speech category and approximate match of overall corpus frequency. The 98 filler words were randomly selected from CELEX based on the following constraints: fillers did not contain the phonemes /t/ or /d/; did not contain the orthographic letter strings “t” or “d”; did not begin with a capital letter (to exclude proper nouns) or include apostrophes or hyphenation; were not longer than four syllables; were a minimum of four letters long; and had CELEX frequency of at least 150. The full lists of both target and



filler words are provided in the Appendix A.1.

Audio versions of each word were recorded by a 20-year-old female native speaker of American English from the Pacific Northwest. The VOT for target items was edited by splicing the onset of each “t”-word onto the rime of the corresponding “d”-word. The “t”-onsets were trimmed in order to impose the specified VOT level (10, 60, or 100ms) within an acceptable range of several milliseconds. Minor deviation from goal VOTs was caused by gluing onsets to rimes at zero-crossing points in order to minimize noticeable acoustic distortions. This editing procedure is consistent and generalizable but retains secondary acoustic (non-VOT) cues to voicing from the “d”-rimes, and thus an overall bias towards /d/ responses, explaining the higher than normal VOT value (60ms) for ambiguous tokens.

Test phase stimuli were CV syllables (a consonant followed by a vowel) of the following form: a t/d onset edited along the VOT continuum followed by the vowel /ɑ/ (pronounced as in the word “spa”). Recordings for the test items were taken from the same speaker as for the exposure stimuli and audio manipulation was performed using the same procedure as was applied to target exposure items. As with the exposure stimuli, specified VOT levels imposed over test items varied within an acceptable range of several milliseconds.

#### *2.1.4. Procedure*

Participants completed the experiment in-lab with headphones. The experiment was implemented using custom javascript code interfaced with psiTurk (version 2.2.3), a toolbox for conducting psychology experiments on MTurk (Gureckis et al., 2016). This was done to ease replication and extension using the same scripts with online participants (which was done in Experiment 2). After consenting, participants completed several questionnaires (demographics, language, attention check) before beginning the experiment. An audio captcha requiring subjects to correctly identify numbers embedded in static noise ensured that audio was at sufficient volume.

Instructions prior to the exposure phase informed participants that they were completing a word comprehension/memory experiment. Part of the instructions encouraged partici-

pants to confirm even slightly “unnatural” sounding words: *“Some of the audio may sound somewhat unnatural but try to ignore this. This is designed to distract you from comparing the audio to the text.”* This was to encourage participants to confirm the ambiguous target items as conforming to the word displayed in the subtitle.

All items in the exposure phase were played along with an accompanying text subtitle and participants were asked to push a button to confirm whether the text and audio “matched” and were provided explicit feedback after each trial. All target words—regardless of audio ambiguity status—were paired with an accurate subtitle. Seventy-eight of the ninety-eight filler items were similarly paired with accurate accompanying text. In order for participants not to be distracted by some proportion of potentially “unnatural” sounding audio (for the manipulated targets) and to conceal the manipulation of interest in the experiment, the remaining 20 filler words were randomly assigned an unrelated text subtitle (e.g. audio is “coffee” but text is “green”) to which the participant was expected to press the NO button. The order of word trials during exposure was randomized for each participant. The use of subtitles ensured that the intended lexical (and hence phonemic) interpretation for the manipulated targets was upheld while also affording direct control of the temporal availability of disambiguating cues for integration.

For those participants in the shifted-/t/ condition, visually presented “t”-words were paired with ambiguous audio (60ms VOT) whereas visually presented “d”-words were paired with unambiguous audio (10ms VOT). For those in the shifted-/d/ condition, the opposite was true: “d”-words were paired with ambiguous audio (60ms VOT) whereas “t”-words were paired to unambiguous audio (100ms VOT). This pattern is illustrated in Figure 3.

After completing the exposure phase in their assigned condition, each participant undertook the same test phase—a classic phoneme categorization task (Liberman et al., 1957)—consisting of 162 trials. Participants received new instructions telling them to press a button to decide whether the audio they heard was “ta” or “da”. The side of the screen on which the “ta” and “da” choices appeared was consistent within each participant but randomized

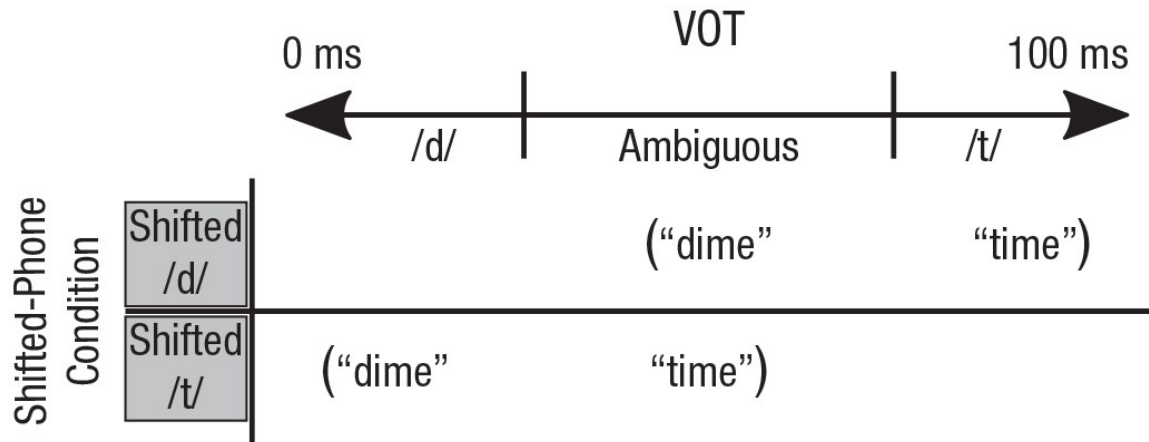


Figure 3: Pairing of text and audio used in Experiments 1 and 2 in the shifted-/d/ and shifted-/t/ conditions. Although all participants were exposed to the same text, participants in the shifted-/d/ condition heard audio with ambiguous voice-onset times (VOTs) paired with “d” text, whereas participants in the shifted-/t/ condition heard audio with ambiguous VOTs paired with “t” text.

between participants. On each trial, participants were exposed to audio of a CV syllable edited along a continuum between “ta” and “da”. After listening to the audio, participants were asked to judge whether the syllable contained “t” or “d”. The 162 test trials were divided between two exemplar “ta/da” tokens and nine VOT levels [20, 30, 40, 45, 50, 55, 60, 70, 80ms], with nine repetitions for each exemplar and level (2x9x9). Test items were randomized within a set of nine blocks, so every stimulus was heard once before it was repeated.

#### 2.1.5. Predictions

In the text-before condition, the subtitles appeared on screen during each trial of the exposure phase 2000ms prior to the start of audio (shown in part (a) of Figure 1). Participants in the text-before condition could thus activate the correct lexical hypothesis before hearing the manipulated targets and would thus be able to map the acoustic-signal to the proper interpretation independent of signal-retention. Therefore, adaptation would be expected under either signal-retention or AOC.

In the text-after condition, the subtitles appeared on screen two seconds after the audio had begun playing. Since the 2000ms gap was measured from the onset of audio, the actual gap

from the end of audio to the display of text was somewhat less than 2000ms (between 1000ms and 1500ms depending on the duration of the spoken word). This is illustrated in part (b) of Figure 1. If the text-after group shows the same adaptation as the text-before group, this would be in line with a signal-retention account that intermediate speech representations include information about phonetic-cues. Conversely, AOC predicts no adaptation for the text-after group. On this view, participants are able to update their representations of the correct lexical item, and thus properly perform the match-mismatch task during exposure, but they are unable to generalize the shifted audio, since they have not stored the underlying acoustic-phonetic information required to do so.

#### *2.1.6. Exclusions*

I excluded participants whose match/mismatch response accuracy during exposure was less than 80% and participants whose exposure response times were less than 150ms on more than 25% of all responses (indicating a misunderstanding or noncompliance with the task). I further excluded participants whose “da” confirmation rates during test were lower for low-VOT trials than high-VOT trials (indicating either random responses or having accidentally flipped the scale). This resulted in 128 remaining participants for analysis (exclusion rate of 3%), divided among the conditions in the following way: 33 in text-before shifted-/t/, 30 in text-before shifted-/d/, 36 in text-after shifted-/t/, and 29 in text-after shifted-/d/.

#### *2.1.7. Analysis*

A mixed effects logistic regression analysis was conducted on trial-level data. The main dependent variable was “t”-responses: whether participants chose the t- or d-item on each trial of the categorization task. The independent variables were experimental condition: Shifted Phone (shifted t vs. shifted d, sum-coded), and Timing (text-before and text-after, sum-coded), as well as their interaction. VOT (continuous variable, scaled and centered) and Test Half (first vs. last, sum-coded) were included as main effects and interaction terms with experimental conditions to test whether the effects of interest changed over the course of the test phase; this follows previous observations (e.g. Liu and Jaeger, 2018) that perceptual adaptations may be unlearned, to some degree, throughout testing. I attempted to include

block number (1 to 9, centered) as a factor, but no models with this factor converged, so I used Test Half (first four blocks vs. last five blocks) instead. I used the maximal random effects structure that converged; this structure included random intercepts for participants and test exemplars, VOT as random slopes for participant and test exemplars, and condition (Shifted Phone and Timing) as random slopes for exemplar (Barr, 2013). Full model structures are available in the Supplemental Material. I tested for significance of factors in models by using likelihood ratio tests on the  $\chi^2$  values from nested model comparisons with the same random effect structure (Matuschek et al., 2017). I computed Bayes Factors where appropriate to quantify the degree of support in favor of accepting or rejecting null hypotheses. All Bayes Factors were computed in R using the `brms` package (Bürkner, 2017) with default parameters, except where required for accurate estimation of posterior probabilities (see Appendix A.4.)

#### *2.1.8. Results*

In the exposure phase, performance of the included participants was high and was comparable across conditions: accuracy in confirming the audio/subtitle match on unambiguous target items was above 99%, on ambiguous targets was above 96%, and on fillers was above 97%. This suggests that for the included participants, the matching task at exposure was not notably more difficult within one set of exposure conditions over another. Indeed, a mixed-model with a main effect of Timing is not a better fit to exposure-accuracy on ambiguous targets than one which includes only random effects,  $\chi^2(1) = 0.42$ ,  $p = .519$  (Bayes Factor = 0.32). This high accuracy (above 96%) on ambiguous targets in the text-after condition suggests that participants held an intermediate representation over time between hearing the word and seeing the text. What type of representation this was can only be revealed by examining the adaptation patterns from the test phase.

Results from the test phase appear in Figure 4 (split by Shifted Phone and Timing). As can be observed, adaptation was successful: the psychometric functions are different between shifted-/t/ and -/d/ ranges. Remarkably and as predicted under AOC, such an effect was only observed in the text-before condition; the categorization functions are not reliably

different in the text-after condition as a function of shift-direction, i.e. adaptation did not occur in the text-after condition. The adaptation additionally began to fade over time: the magnitude of adaptation (in conditions where it was present) was larger in the first half of testing than in the second half of testing (see Supplemental Material for additional visualizations). This reduction in the adaptation effect over time is in line with previous findings (Liu and Jaeger, 2018) perhaps not surprising given the remarkably limited sample during exposure (only 22 edited tokens out of 142 total) and comparatively long testing phase.

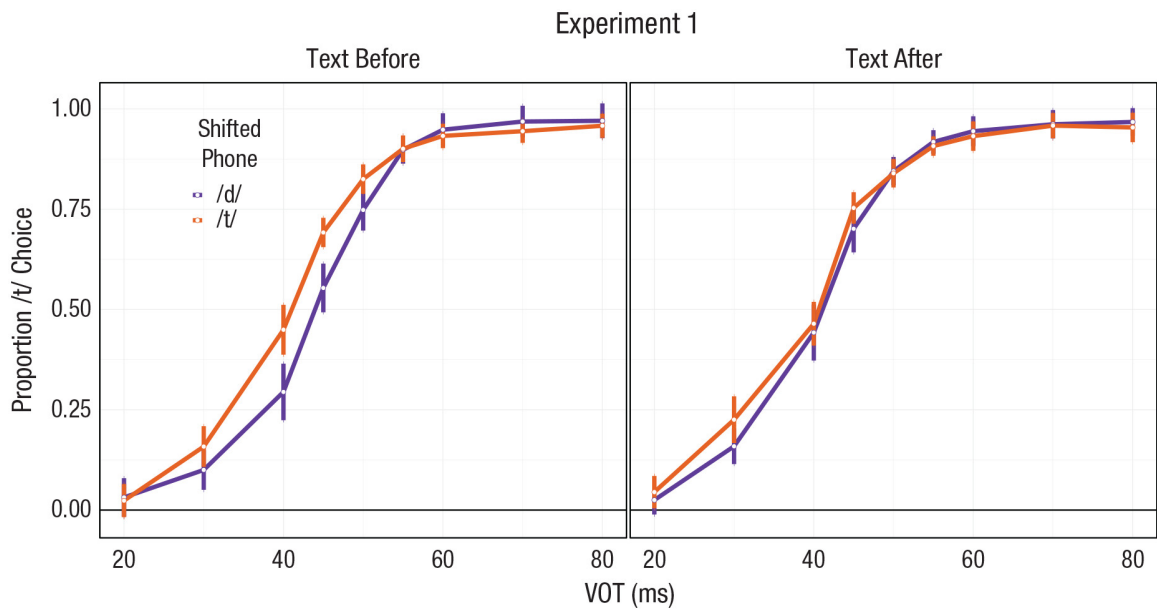


Figure 4: Psychometric functions for Experiment 1: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals. Adaptation occurred in the text-before condition, but did not occur in the text-after condition.

These results were confirmed in mixed-effects model comparisons. First, I compared models over all of the data. The best-fitting model was one including a main effect of VOT and a main effect of Test Half, with main effects and interactions of Shifted Phone, Timing, and Test Half. This model was a better fit than one that did not include the interaction of Shifted Phone and Timing and the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(2)$

= 6.42,  $p = .040$ , and better than a model without the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(1) = 5.66$ ,  $p = .017$ . These modeling results demonstrate that adaptation was higher in the text-before condition than text-after, and that the adaptation effect faded over time during the test phase.

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.84 [0.45, 3.23]	2.6	.009	6.29 [1.57, 25.17]
VOT	3.47 [3.31, 3.63]	43.13	< .001	32.2 [27.5, 37.7]
Shifted Phone	-0.12 [-0.34, 0.1]	-1.08	.282	0.89 [0.71, 1.1]
Timing	-0.23 [-0.45, -0.01]	-2.08	.038	0.79 [0.64, 0.99]
VOT x Shifted Phone	0.24 [0.1, 0.37]	3.41	< .001	1.27 [1.11, 1.45]
VOT x Timing	0.02 [-0.12, 0.15]	0.24	.811	1.02 [0.89, 1.17]
Shifted Phone x Timing	-0.26 [-0.48, -0.04]	-2.33	.02	0.77 [0.62, 0.96]
VOT x Shifted Phone x Timing	-0.23 [-0.37, -0.1]	-3.33	< .001	0.79 [0.69, 0.91]

Table 1: Output of the best fitting model predicting /t/ responses on the first half of test trials for Experiment 1. Bracketed values are 95% confidence intervals.

Given the significant triple interaction of Shifted Phone, Timing, and Test Half, I next tested for the effects of interest (Shifted Phone and Timing) in each test half separately. In the First Half, the best-fitting model was one that included main effects and interactions of VOT, Shifted Phone and Timing (Table 1). This model was a better fit than one that did not include the interaction of Shifted Phone and Timing or their interaction with VOT,  $\chi^2(2) = 13.79$ ,  $p = .001$ , and better than one that did not include the triple interaction of VOT, Shifted Phone, and Timing,  $\chi^2(1) = 11.28$ ,  $p < .001$ . In contrast, in the Last Half, the best-fitting model was one that included main effects of VOT, Shifted Phone, and Timing, and interactions of VOT and Shifted Phone, and VOT and Timing, but no interaction of Shifted Phone and Timing. A model with the additional interaction of Shifted Phone and Timing was not a significant improvement,  $\chi^2(1) = 0.26$ ,  $p = .613$ , nor was one with the additional triple interaction of VOT, Shifted Phone, and Timing,  $\chi^2(2) = 1.37$ ,  $p = .503$ . These modeling results confirm that the timing-specific adaptation effect was only present in the first half of the test phase while fading in the second half. Additionally the interaction between VOT and other fixed effects is expected since adaptation is understood to represent a change in participants' criteria for t/d-categorization. This shift manifests

most strongly for otherwise ambiguous stimuli rather than remaining consistent throughout the VOT continuum (as might occur if instead participants had learned a general bias towards one phone or the other).

Next, I directly compared the effect of Shifted Phone separately in the two Timing conditions, text-before and text-after, to confirm that the effect was indeed only present in the text-before condition, and not in the text-after condition (First Half of test phase only). For text-before, the best-fitting model was one that included main effects of VOT and Shifted Phone. This model was a better fit than one that did not include the effect of Shifted Phone,  $\chi^2(1) = 6.92$ ,  $p = .008$  (Bayes Factor = 19.13). In contrast, in the text-after condition, the best-fitting model was one that included only the main effect of VOT. A model with the additional main effect of Shifted Phone was not a better fit,  $\chi^2(1) = 0.01$ ,  $p = .906$  (Bayes Factor = 0.32). These modeling results demonstrate that the adaptation effect was not simply greater in the text-before condition than in the text-after condition, but that no adaptation effect was statistically detectable in the text-after condition.

I additionally performed several secondary analyses to investigate factors that could instead contribute to the lack of adaptation in the text-after condition. Overall I found no notable differences in participants' behavior during the exposure task: Accuracy on target items during the exposure phase was consistently high across conditions (see above), and analyses showed no relationship between exposure-trial response times and test-behavior, nor any evidence of bimodality in participant categorization performance within exposure-condition (see Appendix A.5). Indeed, these kinds of lexically-guided adaptation effects are surprisingly easy to induce in a range of tasks with different demands, including word counting, syntactic judgements, or loudness judgements (Drouin and Theodore, 2018; McQueen et al., 2006) provided that listeners properly resolve ambiguous audio to the right phonological categories.

Lastly, the adaptation attested in text-before participants was mainly driven by the shifted-/d/ condition and not the shifted-/t/ condition. In model comparisons using data from



each Shifted Phone condition separately (First Half of test phase only), a model with a main effect of Timing was significant for the shifted-/d/ group,  $\chi^2(1) = 8.69$ ,  $p = .003$ , but not the shifted-/t/ group,  $\chi^2(1) < 0.001$ ,  $p = .997$ . Perhaps this was due to interference from secondary acoustic cues to voicing such as pitch or vowel length. Indeed an examination of exposure “accuracy” (i.e. confirming the subtitled as a match to the audio) on ambiguous target items across /d/ and /t/ conditions is consistent with such an interpretation: prior to participant exclusions, the mean accuracy in the shifted-/t/ groups (both text-before and text-after) was 93% while for shifted-/d/ groups it was 98%. Nevertheless this /t/ vs. /d/ asymmetry does not impact the main theoretical interpretation with respect to signal retention or AOC, and I took steps to address this in Experiment 2 which I discuss in turn below.

#### *2.1.9. Interim Discussion*

Overall, successful adaptation effects were observed: the condition of Shifted Phone (/t/ vs. /d/) during exposure was successful at modulating participants’ psychometric functions in a phoneme categorization task. Crucially, this adaptation to the exposure phase only occurred when participants received disambiguating information before the acoustic input (text-before condition). Such adaptation did not occur in the text-after condition, when the acoustic stimulus ended before the disambiguating information was viewed. These results support AOC and are inconsistent with a signal-retention account.

## 2.2. Experiment 2

Experiment 2 aimed to replicate the main findings from Experiment 1 while confirming that the effects of interest are robust to minor experimental modifications.<sup>2</sup> The design was the same except that I additionally manipulated pitch to remove the main secondary acoustic cue to voicing, utilized a norming study to select the maximally ambiguous VOT-level for target items, made minor adjustments to display timing to better equate conditions, and sampled participants from an online subject pool.

---

<sup>2</sup>An initial version of this study introduced a confound between stimulus editing and phonological category (reported as Experiment S1 in Appendix A.10)

### *2.2.1. Design*

Experiment 2 matched the design from Experiment 1, but with a change to the display timing. The timing for the exposure phase in Experiment 1 was as follows. In the text-before condition participants saw text for 2000ms before the corresponding audio was played. However, the text remained on screen throughout the presentation of the audio up until the participant had responded with a match-mismatch judgement. In the text-after condition for Experiment 1, the audio was played first, and then after a gap of 2000ms (counting from the onset of audio) the text subtitle appeared and remained on screen until a match-mismatch judgement was provided. There was thus an asymmetry in the duration of text availability between conditions: text-before participants in Experiment 1 saw the subtitles for longer than the text-after participants. To address this, the display timing for Experiment 2 was adjusted. For text-before participants in Experiment 2, the subtitle appeared on screen for a fixed duration of 875ms. Then there was a gap of 1125ms during which a blank screen was displayed prior to the audio. Audio was then played with nothing on screen. Immediately following the end of the audio, instructions were shown prompting participants for a match-mismatch judgement (which did not include the original subtitle). In the text-after condition for Experiment 2, participants first heard the audio (with a blank screen). After a gap of 2000ms from audio-onset, the subtitle appeared for a fixed duration of 875ms. Following that participants saw instructions to provide a match-mismatch judgement which, like the text-before condition, did not include the original subtitle.

### *2.2.2. Participants*

Power analyses of the results from Experiment 1 suggested that I would have 90% power to detect the effect with approximately 37 subjects in each condition, or 148 subjects. Given additional expected dropout from running the study online rather than in-lab, I recruited 194 participants using Amazon Mechanical Turk, divided between the same four exposure conditions as in Experiment 1 (text-before with shifted-/d/, text-before with shifted-/t/, text-after with shifted-/d/, and text-after with shifted-/t/). Somewhat more participants were assigned to the text-before condition overall (106) than the text-after condition (88) due

to an initial glitch in the online platform. Subjects were paid \$2.41 for their participation.

### *2.2.3. Stimuli*

The materials were the same as in Experiment 1, except that target items in the exposure phase were pitch-corrected according to the following procedure. The audio for target stimuli in Experiment 1 was created by gluing different portions of “t”-word onsets onto the rime of the “d”-words. Since pitch contour ( $F_0$ ), which is a secondary cue to voicing (Dmitrieva et al., 2015), is realized on the following vowel, this means that while the VOT values were edited, all the target stimuli retained secondary information consistent with voicing (i.e. the “d” interpretation). To correct for this, I edited new versions of the target audio which were corrected for pitch ( $F_0$ ). I manually extracted the pitch contours for each word-pair and selected a new  $F_0$  onset value at 2/3rds of the gap between the d-onset and t-onset words. I resynthesized the pitch-contours of the d-onset words with a new contour which began at the designated 2/3rds boosted  $F_0$  value and followed a smooth cline (using pseudo-linear interpolation with a step-size of 10ms) down to the original d-word pitch at 160ms into the vocoid.

I conducted a norming study on a separate group of 44 participants (reported in Appendix A.9) to identify the ideal ambiguity point for VOT. For the new pitch-corrected target stimuli I identified the median VOT at which items were classified equally often as the corresponding word beginning with /t/ or /d/ (46.9ms) and used the VOT from our tested range closest to this (45ms) as the cutoff for ambiguous targets in Experiment 2. The test stimuli remained unchanged from Experiment 1 (without pitch-correction) in order to minimize cross-experiment differences.

### *2.2.4. Procedure*

Participants completed the experiment in a web browser using the same interface as in Experiment 1. The only change to the procedure was that I enforced headphone use through a more stringent audio captcha (Woods et al., 2017). In particular, participants were asked to provide loudness judgements on a sequence of tones which were either in matching- or anti-phase between the stereo channels. Since phase differences are greatly attenuated over

loudspeakers, accurate performance on the captcha task was only possible with headphone use. The remainder of the procedure was identical to Experiment 1. The changes were only to the audio stimuli used for target items during the exposure phase and the differences imposed to better equate the display duration of text in the -before and -after conditions. Exclusions and analyses were identical to those in Experiment 1. This resulted in 169 remaining participants for analysis (exclusion rate of 13%), divided among the conditions in the following way: 50 in text-before shifted-/t/, 44 in text-before shifted-/d/, 36 in text-after shifted-/t/, 39 in text-after shifted-/d/. The gap in the final distribution of participants across conditions was due to an initial difference in assignment, with exclusion rates remaining similar (11.3% for text-before participants, and 14.8% for text after participants). The increased exclusion rates in Experiment 2 were primarily driven by participants whose exposure response times were less than 150ms on more than 25% of all responses. Exclusion rates for match-mismatch inaccuracy were about 3% and were comparable to Experiment 1 across both Timing conditions.

The pre-registered analysis plan for Experiment 2 had an additional criterion to exclude those participants whose performance at the extrema of the VOT distributions (20ms and 80ms) was more than 0.15 away from ceiling or floor. This additional exclusion was added to the pre-registration after observing that some participants' psychometric functions in Experiment 1 did not conform to the usual 'S' shape, due to deviance from floor/ceiling performance at the extrema. However, I ultimately decided to diverge from this pre-registered criterion because there was no theoretical reason to expect categorizations at our chosen extrema (e.g. 20ms VOT) to necessarily be at floor or ceiling. I note that excluding these participants did not qualitatively change the reported results in any of the experiments, and results with this exclusion criterion are reported in Appendix A.6.

### *2.2.5. Results*

In the exposure phase, performance of the included participants was high and was comparable across conditions: accuracy in confirming the audio/subtitle match on unambiguous target items was above 97%, on ambiguous targets was above 95%, and on fillers was above

96%. This suggests that for the included participants, the matching task at exposure was not any more difficult in one condition over another. A mixed-model with a main effect of Timing is not a better fit to exposure-accuracy on ambiguous targets than a model including only random effects,  $\chi^2(1) = 2.38$ ,  $p = .123$  (Bayes Factor = 1.35). Of particular note is that, as in Experiment 1, high accuracy on ambiguous targets in the text-after condition suggests that participants held an intermediate representation over time between hearing the word and seeing the text. The content of this representation can only be revealed by examining the adaptation patterns from the test phase.

Data from the test phase appear in Figure 5 (split by Shifted Phone and Timing). As can be observed, adaptation was successful: the psychometric functions are different between shifted-/t/ and -/d/ ranges. Remarkably and again as predicted under AOC, such an effect was only observed in the text-before condition; the categorization functions are not reliably different in the text-after condition as a function of shift-direction, i.e. adaptation did not occur in the text-after condition. Unsurprisingly, and as in Experiment 1, this effect faded over time: the magnitude of adaptation was numerically larger in the first half of the test phase and diminished by the second half.

The results were confirmed in mixed-effects model comparisons. First, I compared models over all of the data. The best-fitting model was one including a main effect of VOT and a main effect of Test Half, with main effects and interactions of Shifted Phone and Timing. This model was a better fit than one that did not include the interaction of Shifted Phone and Timing,  $\chi^2(1) = 6.05$ ,  $p = .014$ . A model with interactions of Shifted Phone and Timing with Test Half did not improve the fit,  $\chi^2(2) = 4.45$ ,  $p = .108$ , nor did one with the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(3) = 5.04$ ,  $p = .169$ . These modeling results demonstrate that adaptation was higher in the text-before condition than text-after, and that the effect was relatively consistent throughout the test phase.

Next, although a model with the triple interaction of Shifted Phone, Timing, and Test Half was not a significantly better fit, I nonetheless tested for the effects of interest (Shifted

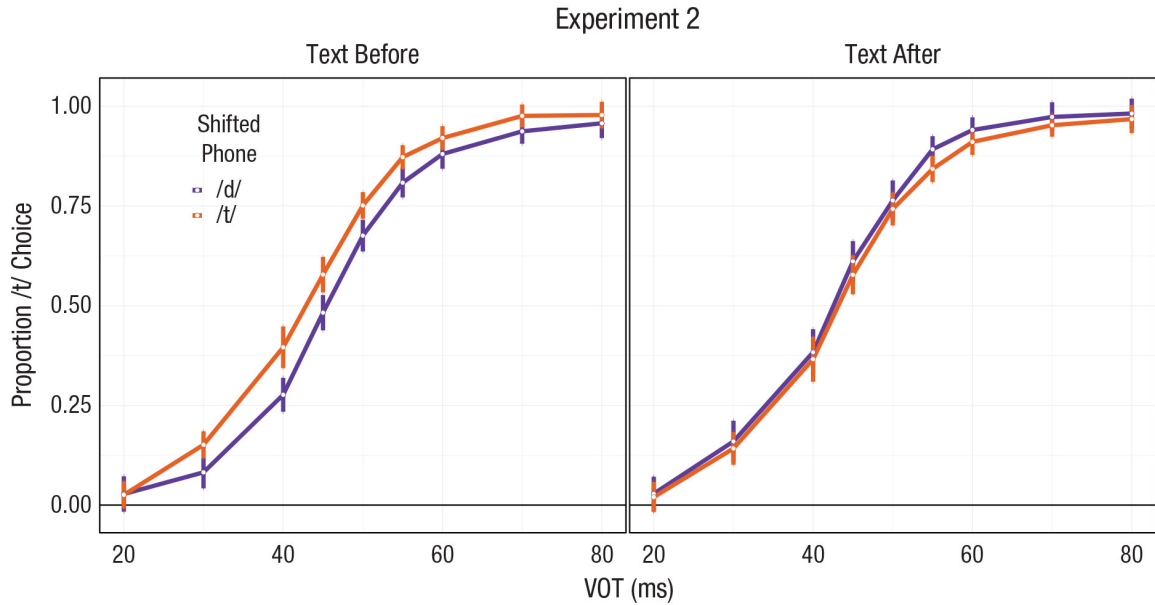


Figure 5: Psychometric functions for Experiment 2: proportion of /t/ choices as a function of voice-onset time (VOT) and shifted-phone condition (/t/ or /d/), plotted separately for the text-before and text-after conditions. Data points are the average of participant means, and error bars are within-subject 95% confidence intervals. Adaptation occurred in the text-before condition, but did not occur in the text-after condition.

Phone and Timing) in each test phase half separately, as was done for Experiment 1. In the First Half of the test phase, the best-fitting model was indeed one that included a main effect of VOT and main effects and interactions of Shifted Phone and Timing (Table 2). This model was a better fit than one that did not include the interaction of Shifted Phone and Timing,  $\chi^2(1) = 5.69$ ,  $p = .017$ , and better than one that included only a main effect of VOT,  $\chi^2(3) = 8.44$ ,  $p = .038$ . Likewise, in the Last Half, there was still an interaction effect of Shifted Phone and Timing: a model that included a main effect of VOT and main effects and interactions of Shifted Phone and Timing was significantly better than one without the interaction of Shifted Phone and Timing,  $\chi^2(1) = 3.95$ ,  $p = .047$ . However, this effect was more subtle; this model was not significantly better than one with only a main effect of VOT,  $\chi^2(3) = 6.23$ ,  $p = .101$ . Together, these modeling results confirm that the timing-specific adaptation effect was present in both halves of the test phase, although it was not as robust in the last half.

Predictor	Coefficient	z	p	Odds Ratio
(Intercept)	1.68 [0.09, 3.28]	2.06	.039	5.37 [1.09, 26.46]
VOT	4.02 [3.76, 4.28]	30.45	< .001	55.58 [42.91, 71.98]
Shifted Phone	-0.11 [-0.34, 0.13]	-0.89	.372	0.9 [0.71, 1.14]
Timing	-0.15 [-0.39, 0.08]	-1.28	.201	0.86 [0.68, 1.09]
Shifted Phone x Timing	-0.29 [-0.53, -0.05]	-2.4	.016	0.75 [0.59, 0.95]

Table 2: Output of the best fitting model predicting /t/ responses on the first half of test trials for Experiment 2. Bracketed values are 95% confidence intervals.

Next, as in Experiment 1, I directly compared the effect of Shifted Phone separately in the two Timing conditions, text-before and text-after, to confirm that the effect was indeed present in the text-before condition, but not in the text-after condition (First Half of test phase only). For text-before, the best-fitting model was one that included main effects of VOT and Shifted Phone. This model was a better fit than one that did not include the effect of Shifted Phone,  $\chi^2(1) = 6.18$ ,  $p = .013$  (Bayes Factor = 3.83). In contrast, in the text-after condition, the best-fitting model was one that included only the main effect of VOT. A model with the additional main effect of Shifted Phone was not a better fit,  $\chi^2(1) = 0.90$ ,  $p = .344$  (Bayes Factor = 0.92). While the Bayes Factor of 0.92 on its own is essentially ambiguous for or against the null model, the alternative model (i.e. one including an effect of Shifted Phone) actually contains a weak trend in the opposite direction of the original acoustic signal: the overall rate of “t-choices” is negligibly higher for text-after shifted-/d/ participants than it is for text-after shifted-/t/ participants. These modeling results demonstrate that the adaptation effect was not simply greater in the text-before condition than in the text-after condition, but that an adaptation effect was not statistically detectable in the text-after condition at all.

Lastly, the additional steps I had taken to address the asymmetry between shifted-/d/ and -/t/ conditions did not appear to succeed. When examining effect of Timing condition separately in the two Shifted Phone conditions, the effect of timing was significant in the shifted-/d/ condition,  $\chi^2(1) = 7.28$ ,  $p = .007$ , but not the shifted-/t/ group,  $\chi^2(1) = 0.66$ ,  $p = .416$ . While this may have been caused by residual voicing cues (e.g. vowel length), the

asymmetry does not interact with either of the primary theories (signal retention vs. AOC) under discussion.

Experiment 2 replicated the primary findings from Experiment 1. Adaptation to the exposure phase was observed when participants received disambiguating information before the acoustic signal (text-before condition) but not after (text-after condition). These findings were robust to a display timing change and the additional manipulation of pitch in tandem with VOT.

### 2.3. General Discussion

In two experiments, I observed that listeners can adapt to speaker-specific acoustic cues to phone perception (e.g. VOT), but only when disambiguating information is provided *before* rather than *after* hearing the ambiguous acoustic input. When disambiguating text appeared after the ambiguous speech, listeners could verify and accept either lexical alternative (e.g. “time” or “dime,” depending on condition) but they could not use this disambiguating text to learn the particular VOT-to-phone mapping. Only when the order was reversed (text-then-speech) could listeners both verify the intended word and adapt. This finding is consistent with AOC over the signal retention account of speech processing. According to AOC, graded activation of linguistic categories (e.g., phones, words) persists over time but not the acoustic evidence that gave rise to this probabilistic information. Maintenance of probabilistic information about linguistic categories permits the accurate lexical verification during the exposure phase of the text-after condition but blocks the ability to adapt because the acoustic cues were not retained. Even the most course-grained representation of acoustic cues would have been sufficient for adaptation (i.e. tracking “high” and “low” VOT values spaced far apart during exposure), yet adaptation did not occur.

Such a finding is consistent with the demands of real-time language processing. Consider how little is lost by not retaining VOT information compared with how much is gained in performance by storing probabilistic activation over higher-level categories. Indeed, I know of no linguistic phenomenon that requires the computation of “long-distance dependencies”



(over seconds) between acoustic cues and later arriving linguistic input, but dependencies abound for linguistic categories such as phonemes and words, over which phonological and syntactic systems traffic, respectively. This likely reflects a general property of perception and cognition over time: lower-level representations may be fast-changing and ephemeral, mirroring the input, whereas intermediate and higher-level categories are more persistent, given their need for inference and integration.

The presents experiments, though, can only speak directly to intermediate speech representations on the timescale of about one second and beyond. Indeed, neuroimaging studies (e.g. Toscano et al., 2010, 2018) indicate that acoustic detail is present during early cortical processing for up to 200ms. This suggests a more refined AOC account, under which early perceptual representations are built based on acoustic cues over the first few hundred milliseconds, with information passed on to higher-level categories beyond that. An alternative possibility is that while the fingerprint of acoustic cues can be detected during early cortical processing, this information is not available to the components of the cognitive system used for subsequent interpretation. Such a “modular” variant of AOC would provide a mechanism in support of previously identified limits on perceptual learning: Jesse and McQueen (2011) found that Dutch listeners adapt to speech when ambiguous targets appear word-medially (“bene[f/s]it”) or word-finally (“regre[ss/ff]”) but not word initially (“[f/s]reedom”). They suggested a “Timing Hypothesis” which proposed that relevant lexical knowledge must be available before hearing the ambiguous sound to support adaptation. AOC offers an explanation of why such a Timing Hypothesis would be true, namely that intermediate representations of speech consist of activated linguistic categories, not sub-phonemic or acoustic information. Future work is required to disentangle these two variants of AOC and related questions on a narrower timescale.

At a broader timescale, AOC clarifies the interpretation of listeners’ sensitivity to within-category acoustic variation. Past work showing that performance on memory tasks depends on acoustic clarity (Crowder and Morton, 1969; Frankish, 2008) or that sensitivity is main-

tained across syllables (Brown-Schmidt and Toscano, 2017; Falandays et al., 2020; McMurray et al., 2009) or integrated over a delay (Galle et al., 2019; Gwilliams et al., 2018), did not and cannot address the internal contents of the representations that support such sensitivity. Alas, the received wisdom in this area been that gradient behavior entails gradient representations. I argue that this is not the case. Downstream behavior (intermediate representations) display sensitivity to gradient input while not including a direct representation of the underlying signal. The present findings provide direct evidence in favor of the position that gradience is maintained through probabilistic uncertainty about potential categories. Within-category sensitivity to phonetic cues (VOT) is consistent with AOC and does not entail storage of those phonetic cues in intermediate representation.

Any cognitive representation of speech is subject to the “Immediacy of linguistic computation.” The structure of intermediate representations thus has substantial ramifications as a bottleneck for broader theories of phonological representation. For instance, the present findings raise important questions for exemplar accounts of phonology (e.g. Bybee, 2002; Pierrehumbert, 2001; Johnson, 2006): how could acoustic-phonetic detail be stored in stable representation when it is not active in an intermediate state over the span of even several seconds? Rather these results appears in line a traditional generative view of phonology which takes “exemplar” effects of articulatory and perceptual variation on a word-by-word basis as arising from mechanisms in psycholinguistic processing that are somewhat isolated from lexical and phonological representations themselves.

While acoustic maintenance may appear to be supported by findings that unsupervised exposure or time-delayed subtitles may attenuate the processing difficulties associated with unfamiliar accents (Bradlow and Bent, 2008; Burchill et al., 2018), such adaptation can also be accomplished under AOC through listeners’ use of contextual information to predict upcoming words and evaluate/adjust to the bottom-up mapping accordingly. Such a top-down mechanism finds support in recent electrophysiological evidence (Getz and Toscano, 2019). Likewise, infants’ difficulty processing unfamiliar variants of their native languages (Cristia

et al., 2012) is overcome when words are embedded within the context of highly familiar stories (van Heugten and Johnson, 2014). Thus while there are experimental conditions which prevent adaptation from occurring (i.e. our text-after condition), being able to predict and activate upcoming linguistic material before the corresponding signal arrives (Jesse, 2021) compensates for the restrictions imposed by the immediacy of computation. Category representations provide the bridge that supports listeners' adaptation to variability despite computational and structural restrictions around the ephemeral signal.

## CHAPTER 3 : Word Learning as Category Formation

The central task of a natural science is to make the wonderful commonplace: to show that complexity, correctly viewed, is only a mask for simplicity; to find pattern hidden in apparent chaos.

---

Herb Simon. *Sciences of the Artificial*

Children famously face ambiguity during of morphological and syntactic acquisition (Yang, 2002, 2016; Tyler and Nagy, 1989; Pinker, 1989; Rumelhart and McClelland, 1985): how does the learner deal with ambiguity when multiple grammars have extensions that are overlapping (Gold, 1967)? While words, unlike syntactic units, are often thought of as atomic, a fundamental question in word learning is strikingly similar: how, given only evidence about what objects a word has previously referred to, are children able to generalize to the total class? How does a child end up knowing that “poodle” only picks out a specific subset of dogs despite their overlapping extensions? Learners display surprising sophistication in their ability to perform statistical inference over the input distribution (Saffran et al., 1996). However, this does not specify a mechanism by which such computation are actually made. What’s more, learners appear sensitive to input conditions that are orthogonal to pure statistical input (e.g. the timing and relative order of stimuli) — motivating the exploration of an architecture which can explicitly *generate* hypotheses prior to their evaluation. The *Immediacy of Linguistic Computations* serves as an informative bottleneck on the acquisition process here, as it did for speech perception in Chapter 2. The learner is limited to evaluating the fit of whatever intermediate representations they posit (be they potential syntactic rules or possible word meanings). One hypothesis will end up winning out because it offers a sufficiently good fit to the data, but this does not mean that the final grammar or meaning is necessarily the “best.” Learners do the best job they can, not the best job possible. In this chapter, I first review part of the large literature on word learning from the perspective of generalization, including a reanalysis of the empirical data related

to the effect of stimulus *timing* on learning (Sections 3.1-3.2). The chapter then presents a new computational model of word learning (NGM) which is able to capture and explain empirical results related to input timing (Sections 3.3-3.5).

### *3.0.1. Word Learning and Generalization*

A crucial facet of language acquisition is the development of the lexicon. Language learners need to infer the set of vocabulary items belonging to their particular language based on the patterns of speech produced around them (see Bloom (2000), among others, for an overview). The lexical entries that learners need to store consist of numerous components. These include a word’s pronunciation, potential syntactic and morphological roles and marking, as well as its meaning (Carey, 1978). While a substantial literature addresses the problem of resolving referential ambiguity (Yu, 2008; Trueswell et al., 2013; Fazly et al., 2010; Medina et al., 2011)—e.g. does “wug” refer to the bird or the squirrel that was being pointed at in the park—the establishment of word meanings does not end there.

Learning even simple categories involves a difficult inductive problem (Quine, 1960). Consider a sample environment for learning the word “dog”: A child hears an adult speaker point at their pet and refer to it with the label /dɔg/. While from the perspective of referential ambiguity the situation is clear—the intended referent is the dog rather than the dishwasher—the space of possible meanings for the phonological label /dɔg/ is still quite large. The word may be the particular pet’s name, or it could mean pets generally. It might pick out the set of (all and only) dogs. But it also might select the set of poodles, or mammals, or animals. It might refer to the appearance of the dog: spotted or four-legged or tired. The list goes on. Moving beyond the reference mapping problem for word learning, this chapter focuses on the subsequent and thorny question of meaning generalization. It is a speaker’s intended conceptual characterization rather than any particular scene-dependent referent that should constitute a word’s meaning (Chomsky, 1957; Fodor, 1983; Gleitman, 1990, *inter alia*).

Experimental work on lexical acquisition has shown that language learners approach the

problem under strong biases with respect to meaning. This is functionally beneficial since it severely limits the size of the potential search space. For instance, learners generally assume new words refer to whole objects rather than sets of adjacent parts (Markman, 1989). There is a bias towards categorization by shape rather than color or size (Landau et al., 1988). Prior vocabulary knowledge has a strong effect through mutual exclusivity (Markman and Wachtel, 1988; Merriman et al., 1989). There are also guiding effects from a host of input signals: syntactic frame (Gleitman, 1990; Naigles, 1990; Snedeker and Gleitman, 2004), phonetic content (Gervain et al., 2008), social-attentive properties (Baldwin, 1991; Gillette et al., 1999), etc.

While vocabulary development is clearly guided by such constraints, an enumeration (or even a rich typology) of biases does not explain the underlying mechanism responsible for learning word meanings. How does this process function? What is the cognitive mechanism behind learners' remarkable ability to infer the meanings of words based only on one or a few instances of usage? A helpful conceptualization of this is that words are invitations to form categories (Waxman and Markow, 1995). It is striking that infants interpret a word as selecting members of some kind, rather than simply naming an individual referent. Put succinctly in Waxman (2003): *“Novel words invite infants to assemble together objects into categories that would otherwise (without linguistic context) be perceived as disparate and distinct.”* While category representations do not necessarily *require* explicit linguistic support, experimental evidence supports a tight link between categorization and word learning (Balaban and Waxman, 1997; Xu, 2002; Ferguson and Waxman, 2017; LaTourrette and Waxman, 2019; Pomiechowska and Gliga, 2019).

If hearing a novel word like “fep” can prompt the learner to create a category, we would like to know what knowledge ends up encoded by that process and how. Once a child has seen that “poodle” can refer to whatever instances of poodles they were exposed to, how does she know that “poodle” can refer to all (and only) items in the real class of poodles? This is in contrast to both failing to generalize sufficiently, e.g. erroneously positing that the word

only refers to their pet, as well as overgeneralizing that the word selects the set of all dogs.

### 3.0.2. Algorithms and Rational Behavior

One influential account of generalization in word learning is the Bayesian inference theory (Xu and Tenenbaum, 2007b). On this view, learners have some representation of many potential meanings for a word and engage in statistically sensitive calculations to select the hypothesis that is most probable given a distribution of attested exemplars. While some predictions of a Bayesian inference model are consistent with experimental outcomes (Xu and Tenenbaum, 2007b), these outcomes do not uniquely support the Bayesian view and are open to alternative explanations like I present here. The most discussed empirical finding in this area is the “suspicious coincidence effect” (SCE)—that an increase in sample-size corresponds to more narrow word meanings. However, as discussed below, other empirical findings (Spencer et al., 2011; Lewis and Frank, 2018) are unaccounted for by existing models. In particular, learner behavior additionally depends on the timing of stimulus presentation: whether training items are presented simultaneously or one-by-one—an effect that is consistent and robust over a range of related studies and domains (Carvalho and Goldstone, 2015; Gelman and Markman, 1986; Lawson, 2014a; Lupyan et al., 2010; Spencer et al., 2011; Lewis and Frank, 2018). This experimental manipulation maps well onto real-world conditions: sometimes a learner encounters examples in temporally distributed occurrences rather than grouped together, and one of the goals of this chapter is to clarify the role and effect of temporal *presentation-style* (PSE) along with SCE that has been obscured by previous literature’s focus on potential interactions (Spencer et al., 2011; Lewis and Frank, 2018).

It is worth noting that the goals of Bayesian approaches to word learning are “*at the level of computational theory (Marr, 1982) or rational analysis (Anderson, 1990) to understand in functional terms how implicit knowledge and inferential machinery guide people in generalizing from examples rather than to describe precisely the psychological processes involved*” (Xu and Tenenbaum, 2007b, pg. 250). We might then contrast the class of *computational* descriptions with *algorithmic* or *mechanistic* explanations of word learning. Even authors of Bayesian models of cognition note that (Bonawitz et al., 2014, pg. 60): “*Following the*

*procedures of Bayesian inference by enumerating and testing each possible hypothesis is computationally costly, and so could not be the actual algorithm that learners use... considering the algorithmic level of analysis more seriously can help to address these significant challenges for Bayesian models of cognition.”* This chapter aims to do exactly that.

Word learning is to construct mental representations of words. While statistical inference accounts of this process posit a global probability optimization over a (potentially large) set of hypotheses, I instead argue that word learning is an incremental process. Like other psycholinguistic processes, this is fundamentally constrained by the Immediacy of Linguistic Computation. From an algorithmic perspective, hypothesized representations are first generated and then only locally revised — as needed — based on input data. On this account, not all plausible hypotheses are simultaneously available. Meanings are built incrementally; any evaluation metric functions only over what is generated from input by the learner. This *Markovian* property is analogous to representations under AOC as outlined in Chapter 2: once the learner extracts a belief about what a word potentially means and the original stimulus disappears, then that representation can be updated in the future but the learner is not directly privy to the sequence of input which led to that belief. This kind of limitation to the domain of computation is similar to the divide between *global* and *local* models of referent mapping in word learning (Stevens et al., 2016).

The Naïve Generalization Model (NGM) presented in this chapter offers an explanation of word learning phenomena grounded in category formation (Smith and Medin, 1981; Medin et al., 1987). The NGM outlines a mechanism by which hearing novel words invites a learner to create a new category from component “features” or “properties.” Learners extract properties of objects and store a mental record of them. This is importantly different from statistical inference models of word learning because, under the NGM, word meanings are *generated* by the learner rather than only selected for. Once a representation for a novel word has been generated, the learner is able to evaluate subsequent labeled objects with respect to this hypothesized meaning; it is these mental representations that serve as the



basis of word meanings and generalizations. I term this process “naïve” in the sense that it does not optimize for any particular global value. Under the NGM, both the creation and evaluation of word meanings functions locally rather than in terms of a total distribution of input. While the NGM gives rise to *rational* input/output mappings such as SCE, Bayesian inference plays no explicit role in the internal algorithm underpinning the word learning process.

The NGM is able to capture both well-discussed (SCE) and previously unaccounted for (PSE) empirical findings in meaning generalization with respect to word learning. Under the NGM, word learning is fundamentally a local process by which mental representations of words are constructed rather than strictly evaluated. The generalization model does not function in isolation. The NGM is embedded within a larger understanding of word learning and is consistent with previous work regarding other aspects of learning required for vocabulary acquisition. See Trueswell et al. (2013); Stevens et al. (2016); Smith and Yu (2008) for mechanisms of referent mapping. The contribution of the NGM is to explain the way in which representations of meaning are created, updated, and maintained.

### *3.0.3. Organization of the Chapter*

In Section 3.1, I outline previous models, and in particular the Bayesian inference model, along with experimental paradigms, and major phenomena in word meaning generalization. Section 3.2 details a reanalysis of the experimental data reported in Lewis and Frank (2018) which serves as a replication of the presentation-effect on a large scale while disentangling the main effect of PSE from its lack of interaction with SCE. I aim for this to clarify recent findings (Spencer et al., 2011)—as well as their potential misinterpretation (Jenkins et al., 2021; Lewis and Frank, 2018) while highlighting the proper conclusions one can draw from them. Section 3.3 introduces the Naïve Generalization Model (NGM) and its internal mechanisms, as an implementation of the theory of word learning as category formation. The local computation of hypotheses within the NGM can account for both SCE and PSE in a unified way. Section 3.4 describes the output of the NGM compared with human performance on seven different experimental conditions for word learning, varying over presentation-

style, number, and hierarchical relation between training items. Based on two evaluation schemes, one parameter-free and one parameter-tuned, these results offer support for the NGM over statistical inference models of generalization in word learning, and provide a concrete mechanism for how words invite the creation of categories (Waxman and Markow, 1995; Waxman, 2003). Section 3.5 concludes and offers directions for future work grounded in studying the intermediate representations recruited during word learning rather than strictly a set of output conditions.

### 3.1. Models, Experiments, and Major Findings in Generalization

#### 3.1.1. *Word Learning as Bayesian Inference*

Some of the most popular approaches to generalization in word learning have been built on hypothesis comparison and global optimization (Xu and Tenenbaum 2007b and subsequent work): A large set of hypotheses compete based on the relative probability that each hypothesis would be generated by the attested input data. The task is then re-framed as choosing how words map onto those concepts by ruling out impossible or less probable hypotheses until a consistent hypothesis is reached. This stems from work on word learning by *hypothesis elimination* more broadly such as Pinker (1989) and Siskind (1996). Other approaches based on associative learning (e.g. Colunga and Smith, 2005; Landau et al., 1988; Regier, 2005; Roy and Pentland, 2002), have not been formalized to interface with the main experimental paradigm discussed in this paper (Xu and Tenenbaum, 2007b; Spencer et al., 2011; Lewis and Frank, 2018) and so I will not address them further here (which might otherwise require a chapter-length discussion on its own). Instead this chapter focuses primarily on comparison to the Bayesian inference model of word learning.

The Bayesian model posits that the learner keeps track of their observed sample of referents (out of a known domain of possible items) labeled by a novel word. By assuming that the learner has access to a hypothesis space over the possible concepts that this novel word might map to, the heavy lifting in word learning is understood to arise from a probabilistic model relating individual hypotheses to the observed sample of exemplars. The Bayesian learner evaluates all hypotheses for candidate word meanings according to Bayes' rule, by com-

puting their posterior probabilities (the likelihood of each hypothesis given the input data  $p(\text{hypothesis}|\text{referents})$ ), proportional to the product of prior probabilities  $p(\text{hypothesis})$  and likelihoods  $p(\text{referents}|\text{hypothesis})$ ).

This kind of model can be thought of as *global* in two ways (Stevens et al., 2016). First, calculations of hypothesis-fit to the data are taken over all input received. In the limit, the learner would need to track some record of every attested exemplar in order to compute probabilities over them. Second, all alternative hypotheses are also calculated for goodness-of-fit to the input data. This allows for global comparison, not only between total input and some temporary hypothesis, but between all hypotheses themselves (*contra* the Immediacy of Computation as a bottleneck on learning).

This Bayesian inference account makes an intuitive prediction dubbed the “suspicious coincidence effect” (SCE), that if a learner is exposed to some new word “fep” (adapted from (Xu and Tenenbaum, 2007b, pg. 249)): *“It would be quite surprising to observe only Dalmatians called feps if in fact the word referred to all dogs and if the first four examples were a random sample of feps in the world. This intuition can be captured by a Bayesian inference mechanism that scores alternative hypotheses about a word’s meaning according to how well they predict the observed data, as well as how they fit with the learner’s prior expectations about natural meanings.”* Even without explicit computation, it seems natural that a learner should be more likely to notice the uniquely “poodle” aspects of some set when shown many poodles to compare, rather than only a single poodle in isolation.

### 3.1.2. Immediate Generalization Paradigm

One paradigm for investigating learners’ behavior relating to generalization in word learning is a simple labeling task from Xu and Tenenbaum (2007b) which I term the *Immediate Generalization Paradigm*. Unlike in paradigms used to probe the referent-mapping problem (Trueswell et al., 2013; Yu and Smith, 2007), participants are provided with an unambiguous word-label for a set of one or more objects. Given a test-grid of other referents, the experimenter can probe what level of generalization a learner has posited for the novel word-label

by measuring subsequent selections from the test-grid.

The Immediate Generalization Paradigm introduced by Xu and Tenenbaum (2007b) consists of photographs of real objects distributed across three different broad categories or classes (animals, vegetables, and vehicles) to be used as stimuli. The test-grid for these experiments consists of pictures of twenty-four items. This is made up of eight items from each of three classes. For any particular item, it is typical to describe some “basic-level” term (Markman, 1990; Mervis, 1984; Rosch et al., 1976) as the label which would most likely be given to it in isolation (e.g. a dog). In relation to the basic-level term, that same item might also be referred to using a more narrow “subordinate-class” label such as “poodle” or a broader “superordinate-class” label such as “animal.” Within each class in the test-grid, objects exist within three hierarchical levels: two items which come from the same subordinate category (e.g. two jalapeños, or two dalmatians), two items which fit into the same basic-level category as the two “subordinate-level objects” (but which themselves each belong to a distinct subordinate category — e.g. a yellow and red pepper, or a poodle and a golden retriever). Lastly, there are four items which share the same broad class (e.g. animal) but which each belong to distinct basic-level categories (an elephant, a bee, a cat, etc.). The set of “test” objects is consistent across trials with only their position on the grid randomized. A sample of this type of learning trial and test-grid used in Xu and Tenenbaum (2007b) and subsequent studies (Spencer et al., 2011; Lewis and Frank, 2018) is shown in Figure 6.

The contextual grounding for this task is that participants are interacting with an alien puppet, ostensibly a monolingual speaker of “alien puppet talk.” On each trial, participants are presented with one or several *training* objects below the test-grid along with an accompanying monosyllabic nonce-word label. For instance, a participant may be shown a picture of a dalmatian with the label “fep” and asked to pick out all the other “feps” for the puppet from the simultaneously displayed test-grid. This paradigm was originally established by Xu and Tenenbaum (2007b) but has been replicated and extended several times. This includes investigating the effects of distributional structure of stimuli (Dautriche and Chemla, 2016),

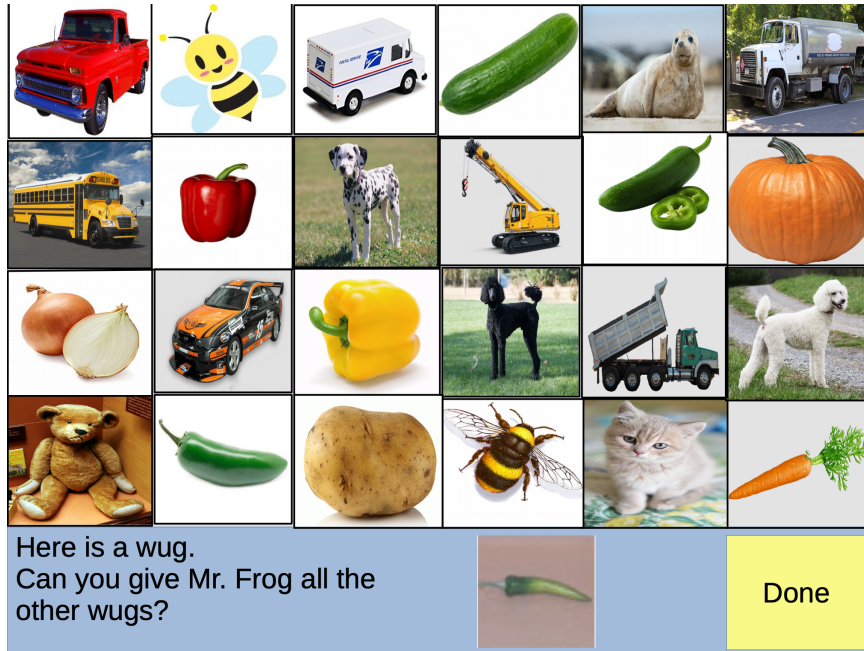


Figure 6: Example word learning trial with test-grid shown to participants in Xu and Tenenbaum (2007b); Spencer et al. (2011); Lewis and Frank (2018). Figure adapted from Spencer et al. (2011).

the role of prior vocabulary knowledge in generalization (Jenkins et al., 2015), and the influence of spatial distance in presentation (Axelsson et al., 2016). These general findings are consistent regardless of whether participants are 3-4 year old children or adults (Xu and Tenenbaum, 2007b) and whether stimuli consist of photos of naturally occurring objects or artificially created patterns (Dautriche and Chemla, 2016).

The nature of the timing with which items are presented to learners plays a material role in the outcome of learning as we will see. Adapting terminology from Spencer et al. (2011) these frameworks are “simultaneous presentation”, in which all training objects are displayed simultaneously directly underneath the test-grid. And “sequential presentation”, in which the training objects are displayed in slightly different locations across very close, yet disjoint times: i.e. the first exemplar is presented at the bottom left for a second and then removed (displaying nothing for a second), then the second exemplar is presented at the bottom middle for a second and removed (again displaying nothing for a second), and then the

third exemplar is presented at the bottom right for a second and removed. This sequence is repeated in full two times before the test set is displayed, for a total of six seconds of study time. Once the test set comes up, the training display continues to loop through until the participant has finished that trial. To be clear, the set of exemplars seen by participants in training are held constant between these two conditions of presentation-style. The difference is in the manner in which the training items are encountered — either all at once, or cyclically, one at a time.

### 3.1.3. *Experimental Phenomena*

When only a single object is presented with a label, then subjects most commonly generalize to the basic-level category, e.g. selecting all *dogs* rather than only *dalmatians* given that the single training item was a dalmatian (Xu and Tenenbaum, 2007b; Spencer et al., 2011). This is consistent with the robust effects of a basic-level bias (Markman, 1990), although see Wang and Trueswell (2019) for the effects of semantic contrast on this comparison. When multiple training examples are presented, then generalization is made narrower (e.g. participants selecting only dalmatians) with respect to the single-exemplar baseline. This “suspicious coincidence effect” (SCE), that category narrowness is linked to the size of the training sample, has been presented in favor of Bayesian inference in word learning. However this phenomenon is not uniquely consistent with Bayesian inference and is thus open to alternative explanations, such as laid out in this chapter. Furthermore, the SCE is not the only important factor which influences generalization in word learning. Spencer et al. (2011) notes that because the implementation in Xu and Tenenbaum (2007b) assumes inference is performed over an independently sampled distribution of attested referents, the Bayesian account predicts that the order or manner in which training instances are received by the learner should not have an effect on generalization. The likelihoods computed over some sampled set of training objects are agnostic as to the sampling order. Xu and Tenenbaum (2007b) does mention potential *pragmatic* effects of sampling on representations, “[a learner may require] a sensitivity to the intentional and epistemic states of the speakers whose communicative interactions produce the examples observed” (see Xu and Tenenbaum (2007a)),

but alluding to the idea of pragmatics does not actually specify a theory or mechanism by which is this actually achieved.

Even when the number and identity of training objects is held constant, the *timing* with which those items are presented to participants has a significant effect on behavior (Spencer et al., 2011; Lewis and Frank, 2018).<sup>1</sup> A series of experiments in Spencer et al. (2011) and Lewis and Frank (2018) test two basic presentation frameworks in the same word learning task. Under *simultaneous presentation* all training objects are displayed simultaneously along with the test-grid. This is the setup that Xu and Tenenbaum (2007b) used to originally measure the SCE: generalization is more narrow when given multiple training items, compared to the single-exemplar baseline. When the same training items are given a single label but displayed to participants in *sequence* rather than all at once, then generalization is significantly broader compared to the simultaneous-presentation baseline, i.e. it is more likely that all dogs are chosen rather than only dalmatians. Spencer et al. (2011) argued that presentation-style interacts with, and may explain, the effect of training-number (SCE). Lewis and Frank (2018) found no evidence for an interaction, but they did not analyze or highlight that there is an independent effect of presentation-style. Put simply, both the size of the training set as well as the temporal manner of presentation (PSE) have notable independent effects on the meanings posited by participants (a point which is discussed in depth in Section 3.2.) These two phenomena taken together (SCE and PSE) are difficult to explain via a global evaluation model or without taking into account the Immediacy of Computation and the mechanisms of visual comparison that this requires. I return to these results in comparison to the output of the NGM in Section 3.4.

While the Bayesian inference model correctly predicts the SCE, the other findings, concerning presentation-style in the data from Spencer et al. (2011); Lewis and Frank (2018), also warrant explanation. There is no mechanism inherent to Bayesian inference which can ex-

---

<sup>1</sup>While Lewis and Frank (2018) demonstrate the lack of interference of presentation-style on the SCE (i.e. the *interaction* between presentation-style and training-number on generalization) they do not test for the *main effect* of presentation-style on generalization which is present in their data as well as Spencer et al. (2011). See Section 3.2 for relevant analysis.

plain a narrowing of generalizations when subjects are shown objects in short succession as opposed to in parallel. If learners were applying Bayesian inference to maximize the probability of a hypothesized word meaning over global input, then a larger degree of subordinate training items should necessarily correspond to an increased probability of a subordinate word meaning independent of minor timing variations. The Bayesian account also predicts that the SCE should grow as a function of sample-size (Xu and Tenenbaum, 2007b). If it was suspicious to see three dalmatians given a single label, it should be far more suspicious to see twice that many be labeled with the same word. Yet, even when doubling the number of sequential training items from three to six, Spencer et al. (2011) found no significant difference in generalization. Such findings *“directly contradict the [Bayesian] model’s claim that the likelihood of generalizing at a particular level is scaled exponentially by the number of exemplars at that level”* (Spencer et al., 2011). This is suggestive that the mechanism underlying the SCE and PSE may not reside in reasoning over distributional statistics, but results from the Immediacy of Computation, and the mechanisms of visual processing and comparative reasoning (Spencer et al., 2011; Gentner and Namy, 2006) as under the NGM.

### 3.2. Robustness of the Presentation-Style Effect

The focus of recent literature (Spencer et al., 2011; Lewis and Frank, 2018; Jenkins et al., 2021) has been on explanations of SCE and the factors that do (or do not) influence that effect rather than on the question of word learning or generalization more broadly. It is thus important to clarify some terminology and properly distinguish between the relevant main effects from how they potentially interact. First, as defined in Section 3, the “suspicious-coincidence effect” (SCE) is the effect that when single-item training trials are taken as a baseline, then increasing the number of training items leads to significantly more narrow generalization. Second, the “presentation-style effect” (PSE) is the effect that when the number of training items is held constant, then the presentation-style (timing) with which they are exposed to participants significantly affects generalization. Namely, sequential presentation of training items leads to broader generalization than simultaneous presentation of the same stimuli. Lastly, the “number-timing interaction” measures whether PSE and SCE



interact, e.g. if SCE depends on PSE or not.

Lewis and Frank (2018) report the results from a number of variant experiments based on the paradigm from Xu and Tenenbaum (2007b) and Spencer et al. (2011). In addition to the two conditions discussed prior (training-number and presentation-style), another design factor to consider is the order in which participants are exposed to experimental trials. Since training-number is a blocked within-subject manipulation, each participant completes the experiment in one of two possible orders: single-item trials first and multiple-item trials second (1-3) or multiple-item trials first and single-item trials second (3-1). Lewis and Frank (2018) find that, in line with the supplemental experiments reported in Spencer et al. (2011), *block-order* has a significant effect on learning outcomes.

Lewis and Frank (2018) demonstrate the lack of interference of presentation-style on SCE and present this as a rebuttal to Spencer et al. (2011), but Lewis and Frank (2018) did not test for the *main effect* of presentation-style (PSE)<sup>2</sup>. In Section 3.2.1, I present an additional analysis of the data from Lewis and Frank (2018). My analysis replicates the findings reported in Lewis and Frank (2018), but additionally uncovers a robust PSE alongside SCE.

### 3.2.1. Analyzing data from Lewis and Frank

The data from Lewis and Frank (2018)<sup>3</sup> encompass a number of different experiments and the way in which these are analyzed or plotted can obscure crucial patterns. While originally run as twelve separate experiments, a number of manipulations do not have a significant effect on generalization behavior (e.g. same vs. different labels across words, trials grouped by stimulus category or interleaved) and are thus not of primary interest. Here I analyze all the data from Lewis and Frank (2018) together to evaluate the potential manifestations of PSE and SCE. I fit a mixed-effects logistic regression to predict of basic-level generalization on

---

<sup>2</sup>Unfortunately Lewis and Frank (2018) is not the only paper to misinterpret the data from Spencer et al. (2011). Jenkins et al. (2021) presents a model intended to capture the supposed “*reversal of the suspicious coincidence effect with sequential presentation.*” When analyzed properly, there is no such reversal of SCE. This only appears to happen if, as Lewis and Frank (2018) note, one were to confound timing and block order: comparing the results of first-block sequential trials against second-block single-item trials.

<sup>3</sup>All data used analyzed in Section 3.2.1 is available here: <https://github.com/mllewis/XMEM/tree/master/data>

each trial using fixed effects of presentation-style, training-number, and block-order (along with their interactions) and random effects for each subject and stimulus class (animals, vegetables, vehicles)—the output of which is summarized in Table 3. When analyzing all trials from Lewis and Frank (2018) (including both first and second-block trials) then there is a significant main effects of presentation-style (PSE), training-number (SCE) and block-order, along with a significant three-way interaction (Table 3). The same results hold if generalization is alternatively coded as a gradient outcome as in Appendix B.2.

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	-1.610	0.312	-5.155	<.001
Presentation-Style (PSE)	-0.445	0.221	-2.016	0.044
Training-Number (SCE)	1.508	0.129	11.654	<.001
Block-Order	-1.444	0.226	-6.387	<.001
Presentation x Number	0.214	0.246	0.870	.384
Presentation x Block	-0.253	0.440	-0.574	.566
Number x Block	-4.326	0.272	-15.903	<.001
Presentation x Number x Block	1.316	0.494	2.666	.008

Table 3: Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization on all trials. Mixed-effects logistic regression predicting generalization based on listed effects as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects along with a three-way interaction between Presentation-Style, Training-Number, and Block-Order.

In order to investigate the shape of the three-way interaction, and in line with the observation from Lewis and Frank (2018) that block-order has a large effect on generalization outcome, I held block-order constant for subsequent analyses. By looking only at the first-block trials (that is the “3” trials in 3-1 ordered experiments and the “1” trials in the “1-3” ordered experiments), we control for this ordering effect and can compare the differing resultant basic-level generalizations between other conditions. For second-block trials I fit a similar mixed-effects logistic regression (see Table 4) which shows no meaningful effects of any training condition. Neither training-number (1 vs. 3 items) nor presentation-style (sequential vs. simultaneous) has a significant effect on generalization during second-block trials. This is very likely a task-effect: once participants are accustomed to the potential hypothesis space and semantic contrasts (see Wang and Trueswell (2019) a direct manipu-

lation of semantic contrast) they generalize only narrow meanings regardless of condition. Adaptation to the generalization paradigm is, perhaps, unsurprising given language users ability to rapidly adapt to systematic cues in other domains such as speech perception (e.g. Caplan et al. (2021); Kraljic and Samuel (2006) as well as Chapter 2)) or syntactic processing (Fine et al., 2013). Since neither SCE nor PSE manifest on second-block trials (and thus diminish the size of those effects when analyzing all-block trials), I additionally analyzed first-block trials on their own.

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	-9.378	0.705	-13.297	<.001
Presentation-Style (PSE)	-0.220	0.547	-0.403	.687
Training-Number (SCE)	0.009	0.546	0.016	.987
Presentation x Number	-0.207	1.093	-0.190	.849

Table 4: Mixed-effects logistic regression predicting generalization outcome on second-block trials (data from Lewis and Frank (2018)) based on listed effects (Presentation-Style, Training-Number, Number-Timing Interaction) as well as random slopes for each subject and stimulus class. Neither SCE nor PSE manifest on second-block trials.

The same mixed logistic model fit to predict basic-level generalization on first-block trials (see Table 5) shows significant main effects of presentation-style (PSE) and training-number (SCE), with no significant interaction between the two. As with the model fit over all data these trends are robust to gradient vs. discrete coding (see Appendix B.2). While the magnitude of PSE is larger in Spencer et al. (2011) compared to Lewis and Frank (2018), the qualitative trend is consistently replicated between multiple labs and at high power considering the large sample-size in Lewis and Frank (2018).

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	-0.815	0.417	-1.957	.05
Presentation-Style (PSE)	-1.091	0.410	-2.664	.008
Training-Number (SCE)	4.950	0.680	7.276	<.001
Presentation x Number	0.246	0.792	0.311	.756

Table 5: Mixed-effects logistic regression predicting generalization outcome on first-block trials (data from Lewis and Frank (2018)) based on listed effects (Presentation-Style, Training-Number, Number-Timing Interaction) as well as random slopes for each subject and stimulus class. PSE and SCE emerge as significant main effects.

Lewis and Frank (2018) is correct in noting that “*SCE is robust to presentation timing*” since PSE and SCE do not interact in their data (contra Spencer et al. (2011) as well as Jenkins et al. (2021)). However the findings from Spencer et al. (2011) are not moot: presentation-style and training-number nonetheless both have robust and significant independent effects on generalization, motivating a unified explanation from an explicit computation model such as the NGM.

### 3.2.2. *Presentation-Style and Learning in Similar Domains*

Similar presentation-style effects have been observed in other generalization problems, suggesting that presentation style is a robust phenomenon worthy of explanation. Children are skilled at performing inductive reasoning by means of generalizing a limited piece of evidence about one or a small sample of individuals (e.g. “this peach has a pit”) to an entire category (“peaches have pits”) (Gelman and Markman, 1986). In such *property projection* tasks, participants are provided with some fact about a set of objects in the same general category (e.g. “These animals have *type-Z* blood inside”) as training. During testing, participants are asked whether or not they think that same property is also present in some novel object (e.g. “Does this other animal also have *type-Z* blood inside?”). Lawson (2014b, 2017) show that even when the training stimuli are held constant, the manner in which they are presented to participants (3 year-olds in this case) induces a significant effect on the outcome. In the Sequential condition, pictures of animals are presented individually, attributed a novel property (“this animal has [property 1/2]”), and placed into a matching pile. In the Simultaneous condition, items are not presented individually. Rather, both samples are presented at the same time in two piles divided by property and described as a total group: “these animals have [property 1],” while “those animals have [property 2].” As in the word learning domain, sequential presentation of exemplars leads to higher rates of broad generalization compared to an otherwise equivalent simultaneously presented set (see Figure 7). When learners study two or more instances of the same concept side by side, transfer to more remote instances or acquisition of a new category (Gick and Holyoak, 1983; Meagher et al., 2017; Omata and Holyoak, 2012; Gentner and Namy, 1999) is more likely

than when only one instance is studied at a time.

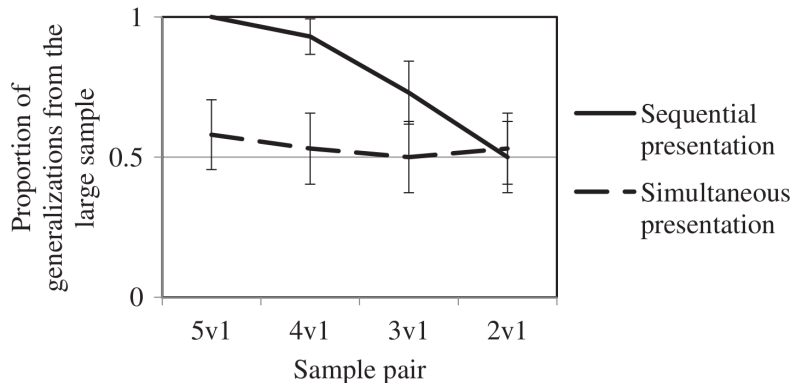


Figure 7: Proportion of broad property-projections based on sample size and presentation-style (figure reproduced from Lawson (2014b) with permission). When presented simultaneously, the size of training has no effect on projection. When presented in sequence, the rates of broad property-projection quickly approach a ceiling condition as the size of training increases.

In fact, a wide range of cognitive tasks exhibit a difference in outcome based on the timing or presentation-style of exemplars. In addition to the timing effects in Chapter 2 this includes, but is likely not limited to, inductive category learning (Carvalho and Goldstone, 2015), visual pattern differentiation (Lappin and Bell, 1972), the link between conceptual and visual processing (Lupyan et al., 2010), visual identification (Shifflin et al., 1973), discrimination learning (Lipsitt, 1961), relational reasoning (Son et al., 2011), orthographic processing (Williams and Ackerman, 1971), and sensory preconditioning (Rescorla, 1980), etc. All of these studies show important differences under sequential vs. simultaneous presentation of stimuli. Taken together, these findings on presentation-style seem to reflect more general issues in category learning, such that we should aim for a unified solution.

### 3.3. Naïve Generalization Model

Word learning is to construct mental representations of words. In considering the process by which learners encounter input materials and extract potential meanings from that signal (whether in natural settings or experimentally controlled contexts), I would like to highlight a few design principles. This section highlights these properties (in line with the discussion from Section 3.0.2) and introduces the Naïve Generalization Model (NGM) as a formal

implementation of the word learning as category formation theory.

Under the NGM, word learning is a dynamic process whereby potential meanings are built incrementally. The set of potential hypotheses that *might* be derived from a given input is large, but in practice, learners are quick to converge on the correct one most of the time (and only a limited number of active hypotheses are actually considered). I argue that this is due not to the inherent virtue of individual meanings in context, but arises from largely mechanistic means. Statistical trends present in the input may not end up manifesting in cognitive representations depending on properties of the learner (e.g. visual attention) or the learning environment (e.g. presentation-style). Hypothesized representations are generated and only locally revised (as needed) based on input data.

The NGM operates by generalizing semantic representations over “features.” These representations are compared against any subsequent data — also implemented as bundles of features — and potentially updated. This internal representation is consulted whenever the word is heard, and in experimental forced choice tasks like the present paradigm, used to select closely matching objects. I discuss these components of the model in turn in Sections 3.3.1 through 3.3.3. By specifying a mechanism by which mental representations of words are both constructed and evaluated, the NGM is able to capture both supposedly *rational* phenomena such as SCE as well as comparatively arbitrary phenomena like PSE.

### 3.3.1. Features

The NGM implementation of “features” follows the classic literature on categories (Smith, 1979; Smith and Medin, 1981; Medin et al., 1987) by representing concepts as salient features/properties. What I call “features” are simply properties that hold for some item (Murphy, 2004). While any two properties may be equally true of an object, in the sense that they are formal operators, it should be clear intuitively that some properties are more salient than others. Consider the number 73. It is probably easier to determine that 73 is odd than it is to determine that 73 is a prime; it is not that its prime-ness is less valid than its being odd, rather it is simply a matter of salience (i.e. how noticeable it is in a particular

context).

To simulate the degree to which a property is noticed by a learner, I model two Gaussian distributions over salience. These “salience distributions” differ only in mean; one for features with elevated prominence (here the driving force behind the basic-level bias)<sup>4</sup> and one for all other features. Not to be misconstrued as an explanation of visual processing itself, these salience distributions are akin to an abstract placeholder: a way of formally implementing the notion that some levels of generalization are privileged compared to others. More complete featural or visual theories (see Johnson and Mervis (1997); Schyns and Gosselin (2002) for instance) or the shifting distribution of attention based on communicative grounding (Steels and Belpaeme, 2005) may offer insight into what drives some level of generalization to manifest as “basic” rather than others, but the NGM captures the way in which categorization and word learning functions to be more narrow or more broad with respect to this baseline condition, regardless of how it is defined. The main point here is that not all features are created equal.

When a learner encounters a new word, the model samples from the appropriate salience distribution for each feature present. The result is a mental representation as a gradient vector of features (Equation 3.1). Features are discrete, but learners represent them by assigning and updating probabilities over such feature values. As with any probability values, these mental records are allowed to range on a gradient between zero and one. The upper-bound of one is intuitively important because, conceptually, this corresponds to the feature being as present mentally as it is in the physical world. The learner iterates over the items displayed (if more than one present) and each feature present in the real world will be stored in mental representation at a proportion relative to that feature’s salience.

Such discussion of “features” or properties need not be limited exclusively to low-level perceptual information. Rather, the NGM is able to operate over any kind conceptual or visual

---

<sup>4</sup>This is not intended as an *explanation* of the basic-level bias per se but is encoded much the same way that basic-level hypotheses are given higher *prior* probabilities under the Bayesian inference account. Both models attempt to explain the conditions under which learners deviate from this baseline.

units which may be further built up throughout development. It is the process of word learning to create rich category representations from existing parts. I would stress though that this generation process is local and non-deterministic, so not every possible competitor hypothesis is necessarily activated given a fixed set of stimuli.

To state the process of feature extraction more formally, a vector representation  $R$  is computed for a label  $w$  based on an example set of training items  $T$  by sampling all features  $f_i$  in  $T$  with salience  $S(f)$ . This is adapted from classic approaches to category membership calculation (Smith and Medin, 1981).

$$R_w = \sum_{t \in T} S(f_i), \quad 0 < i < |t_p| \quad (3.1)$$

$t_p$  is the set of features (or *properties*) of the item  $t$ .  $S(f)$  is the *salience function* for a feature  $f$ , which returns a value sampled from the normal distribution with mean  $\mu$  (dependent on  $f$ ). While features for an object in the world are implemented as formal operators, the mental stored values for a given feature are gradient. The NGM sums the values of each present feature until reaching the ceiling condition (of 1.0). This is in line with previous featural implementations of categories, e.g. (Kruschke, 2008, pg. 287): *“the simplest way [to learn associative strengths] is adding a constant increment to the weight whenever both its source and target node are simultaneously activated.”* I discuss this weighting of features further in Section 3.3.2.

The only restriction placed on learning by feature sampling in the NGM is one of consistency. Since some features are necessarily in conflict with one another — no object is both [+ROUND] and [+SQUARE] — we would like a learner to not represent both such features within the meaning of a single word. Once a feature has been sampled with salience above a “semantic incompatibility” threshold, any other features which would be in semantic conflict with that are not added to representation. An outline of the feature sampling process and subsequent category determination for novel objects is diagrammed in Figure 8.



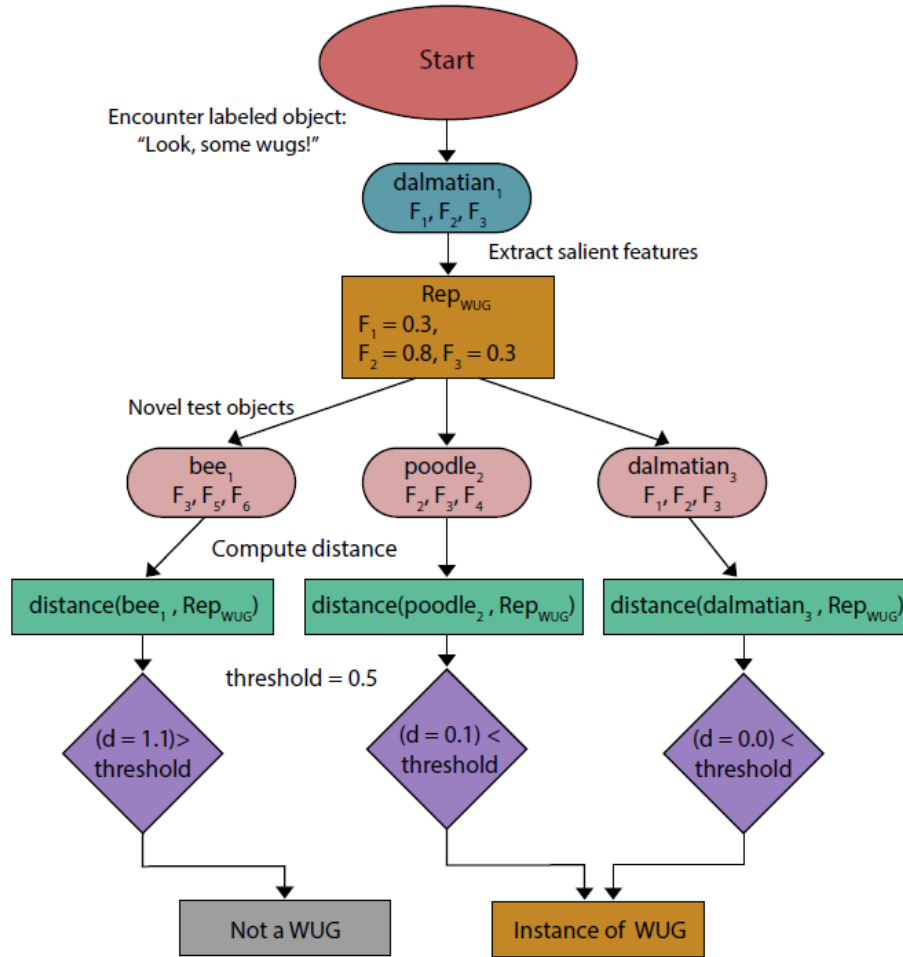


Figure 8: Computation of mental representation from single training example and subsequent comparison to test objects. Values are schematic and for illustration only.

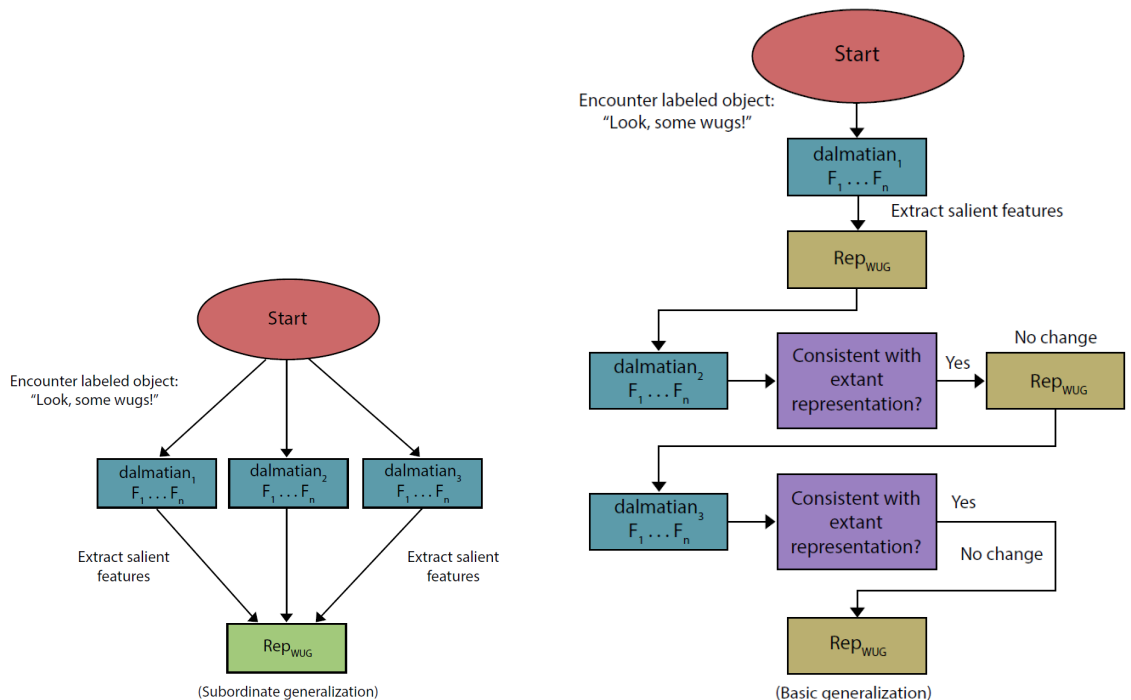
### 3.3.2. Learning

When trained on a single exemplar, the experimental finding (Xu and Tenenbaum, 2007b; Spencer et al., 2011; Lewis and Frank, 2018) is that learners' most likely generalization is to the basic-level. This is driven by the privileged status of certain features for generalization over others (and is effected by factors such as the prototypicality of features (Emberson et al., 2019)). When training objects are initially presented simultaneously then a hypothesis category can be formed in a single shot. Thus, when they are co-present, the function which extracts features from a scene is able to compare exemplars directly to exemplars. When

features are activated multiple times, they are more likely to be encoded in the category representation (Smith and Medin, 1981; Kruschke, 2008). The notably rapid ability of people to extract accurate estimates from a set is well-discussed in the literature on ensemble perception (Ariely, 2001; Whitney and Yamanashi Leib, 2018). Properties which, when encountered in isolation, would not have a significant effect on stored meaning can, through this combination, lead to more narrow generalization. The NGM’s mechanistic account of featural weighting thus makes the same predictions as Bayesian inference with respect to SCE under simultaneous presentation. A typical path from labeling to representation in parallel presentation trials is diagrammed in Figure 9a.

When the same stimuli are presented in sequence rather than in parallel, learners’ generalizations are more broad (Spencer et al., 2011). Even though training objects may be shown to learners multiple times, the learner can construct an initial hypothesis only once. After an exemplar has disappeared from view then, due to the Immediacy of Computation, the learner can only refer back to it by consulting some mental representation for the presented word. Once a mental representation exists, there is no onus to change it significantly so long as subsequent objects picked out by the word are congruent with what has been stored. This process is analogous to localist models of referent mapping (Stevens et al., 2016; Trueswell et al., 2013). Learners generate a hypothesis and either stick with it if evidence is consistent, or move to a new hypothesis (or otherwise incorporate updates) when faced with inconsistent evidence. When subsequent training instances appear, the original exemplar(s) are no longer present, with only the generated category representation remaining. This means that learners are comparing new exemplars to a category representation rather than directly comparing exemplars with each other. Since all of these trials concern levels of generalization, no new training item will disprove an over-generalized hypothesis. Therefore, learners will simply continue along with whatever initial hypothesis was created. Repeat exposures increase a learner’s *confidence* in the hypothesized meaning rather than triggering any change in the word’s internal contents. This continues until some “convergence point” is reached and a semantic representation is more or less fixed. Such a convergence point is a required

component of any model of word learning. The cause of the “basic-level bias” on sequential presentation trials is the same as in the single-exemplar trials: certain types of features lead to privileged levels of generalization. A typical path from labeling to representation in sequential presentation trials is diagrammed in Figure 9b. While statistical trends may be latently present in the input distribution, such patterns may remain unnoticed unless they relate back to whatever initial guesses were hypothesized by the learner.



(a) Typical path of meaning extracted from parallel-presentation trial. All exemplars contribute to initial hypothesized meaning.

(b) Typical path of meaning extracted from sequential-presentation trial. Subsequent “training” stimuli do not affect initial hypothesis so long as consistent. If inconsistent, then a new hypothesis is generated (not depicted here).

Figure 9: Algorithmic flow charts highlighting some possible paths of NGM behavior. This illustrates the common difference in experimental outcome under parallel (left) and sequential (right) presentation of stimuli.

In conceptualizing of word learning as a process of category formation, processing of stimuli is qualitatively different before and after an initial hypothesis representation has been generated—much as we saw with the processing of speech relative to temporal order in Chapter 2. Upon initially encountering referents, when there are no prior hypothesized

meanings to compare against, a representation must be created. At future instances of the word’s usage, however, the learner must decide whether this new token is consistent with the current mental representation or not. If the prior hypothesized representation is inconsistent with current input, then an alternative hypothesis is created. When subsequent input is consistent with the prior hypothesis, then the multiple “trials” across sequential presentation do not necessarily impart any change to the internal contents of the word’s represented meaning. Rather it is the learner’s *confidence* in that hypothesis which gets increasingly solidified.

When items are presented in sequence, only salient features are likely to be encoded in representation. This “sparse” representation corresponds to a broad category generalization. Simultaneous presentation, on the other hand, joins all the shared features between presented items (Gentner and Namy, 1999; Rescorla, 1980; Lawson, 2017). This combined weighting of otherwise non-salient features leads the representations to correspond to more specific, narrow categories.

### 3.3.3. Computing distances

The NGM makes a distance calculation between any new objects and extant mental representations. The comparison of that distance value to a fixed parameter threshold determines category membership. In linear algebra terms, this is formally equivalent to projecting the vector of object features  $t$  into the subspace of the representation  $r$  and taking the  $l_1$ -norm. This calculation, adapted from Smith and Medin (1981), can be written out more mechanically as follows in Equations 3.2 and 3.3:

$$D(r, t) = \sum_{n \in r_{indices}} q(r_n, t_n) \tag{3.2}$$

$$q(r_n, t_n) = \begin{cases} r_n - t_n & \text{if } r_n \geq t_n \\ 0 & \text{if } r_n < t_n \end{cases} \tag{3.3}$$

Unpacking Equation 3.2 a bit, this simply states that to compute a distance between a mental representation  $r$  and some (test/training) item  $t$ , the NGM sums up the gap between features in mental representation ( $r_n$ ) and corresponding properties ( $t_n$ ) in the at-issue item. In Equation 3.3 we see that there is a distance penalty for any feature present in the mental representation  $r$  that is missing in the test object  $t$  under consideration. The size of this penalty is the salience of that feature in mental representation. However, there is no cost incurred for features which are present in a test item but are missing in the mental representation of a class. This asymmetry should be intuitive: representations are by definition abstractions, and thus more sparse than actual items (which can be defined under any number of properties). Every object in the world is going to be perceived as having some color value, but that color may play no role in these items membership in the of various categories being learned here. See Figure 10 for a schematic of this evaluation function. An example of the full word learning experimental calculation is given in Table 6.

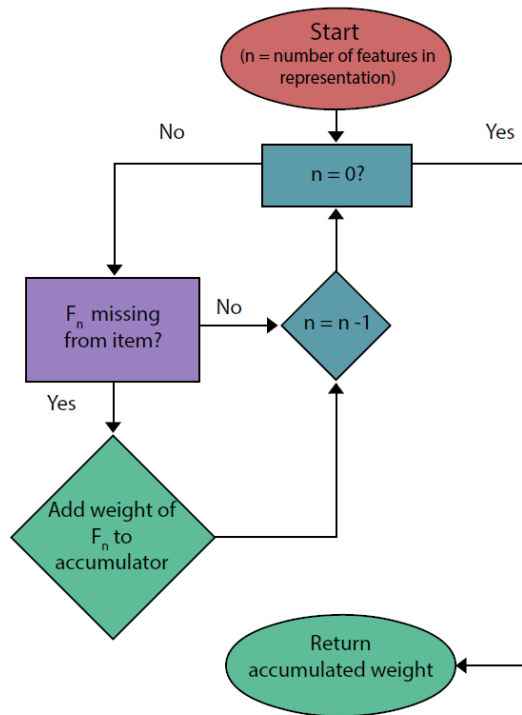


Figure 10: Implementation of distance computation between an object and mental representation under the NGM

	Object	Features	Distance
<b>Training</b> ("Look, a wug")	Dalmatian	(A:1, B:1, C:1, D:0)	N/A
<b>Stored</b>	Category Representation <sub>wug</sub>	(A:0.3, B:0.8, C:0.3, D:0)	N/A
<b>Testing</b> ("Select the other wugs")	Bee	(A:0, B:0, C:1, D:0)	1.1
	Dalmatian	(A:1, B:1, C:1, D:0)	<b>0.0</b>
	Poodle	(A:0, B:1, C:1, D:0)	<b>0.3</b>
	Truck	(A:0, B:0, C:0, D:1)	1.4

Table 6: In this toy example, the initial training set contains a single instance of a dalmatian with features (A:1, B:1, C:1, D:0). From this, the learner extracts a mental representation of (A:0.3, B:0.8, C:0.3, D:0). During testing, a few potential items are all compared against mental representation in order to select category members. Only values present in mental representation but missing from the evaluated items incur a penalty. If the maximum category cutoff were 1.0, then both the dalmatian and the poodle (shown with shaded background) would be selected in this case.

### 3.4. Modeling Results

#### 3.4.1. Scoring

The crux of the category generalization problem is that items within each narrow class are hierarchically nested in other, more broad, categories as well. A “poodle” is a “dog” is an “animal.” However, the scoring methodology adopted by previous work on this task (Xu and Tenenbaum, 2007b; Spencer et al., 2011; Lewis and Frank, 2018), assesses the matches to a specific level of generalization (subordinate, basic, or super) independently.<sup>5</sup> On the test-grid in such experiments, a given training object will then correspond to two uniquely subordinate-matches, two uniquely basic (non-subordinate) matches, and four uniquely superordinate (non-basic) matches. The mean proportion of generalization to each category is then computed over those totals. *“For example, if a participant picks the 2 subordinate matches, it is scored as 100% subordinate generalization. Independent of that, if a participant additionally picks 1 basic match, it is also scored as 50% basic generalization. Likewise choosing 3 superordinate matches is tallied as 75% superordinate generalization”* (J. Spencer, personal communication). For consistency and comparability of evaluation, I implemented

<sup>5</sup>I would like to thank John Spencer (PC) for helpful communication clarifying the methodology used for scoring.

the same scoring methodology as previous work (Xu and Tenenbaum, 2007b; Spencer et al., 2011; Lewis and Frank, 2018). Scoring for each level of category generalization is done by, in each trial, dividing the number of objects selected by the participant (or model) by the total number of possible objects for that level in the test-grid. Below I report two evaluation schemes to compare NGM performance to the empirical results.

#### *3.4.2. Parameter-independent Evaluation*

In order to properly evaluate a computational cognitive model, we should note which aspects of the empirical data we deem important for theoretical explanation. The evidence that results from experiments such as Xu and Tenenbaum (2007b); Spencer et al. (2011); Lewis and Frank (2018) is informative largely on the basis of indicating which experimental conditions drive a significant difference in participant performance, rather than the exact percentages involved. Whether the magnitude of PSE or SCE is 15% or 30% is of little relevance, since it is the presence of the effect that we are primarily concerned with. Thus just as it is important for a cognitive model to be able to capture precise output given the right parameter values, it is also crucial to determine the degree to which qualitative effects of model performance are driven by factors internal to the model itself or dependent on specific parameter settings.

To investigate the parameter-independent performance of the NGM, I measured the proportion of parameter configurations which result in qualitatively the same trends as human empirical output from Spencer et al. (2011); Lewis and Frank (2018). This is assessed in two parts, one for SCE and one for PSE. First, following Spencer et al. (2011), I measured SCE on simultaneous-presentation trials. I defined the SCE as “present” if the proportion of basic-level selections was substantially lower in the multiple-item trials compared with single-exemplar trials. Second, PSE is defined such that sequential presentation results in substantially more basic-level selections compared parallel presentation (while holding constant the number of training items.) “0.15” was chosen as the cutoff for a “substantial” difference for these tests as it is representative of empirical standard deviations in this paradigm. Both required conditions for qualitative evaluation are summarized in Table 7.

Trial Type	Level of Basic Generalization	Interpretation
Single Object	Baseline	Basic-level Bias
Simultaneous	At least 0.15 lower than Single	Suspicious Coincidence Effect
Sequential	At least 0.15 greater than Simultaneous	Presentation-style Effect

Table 7: Major patterns to be captured by models of word learning and generalization. Both the size of the training set (SCE) as well as the temporal manner of presentation (PSE) have reliable effects on the meanings posited by learners. “0.15” represents the typical standard deviation from results in Spencer et al. (2011)

With multiple parameters in the NGM, a large number of configurations are possible for the model to be seeded with. I evaluated 1024 different parameter configurations each run with 1000 simulated “participants.” The output trends of the NGM matched the above criteria for human performance on all runs. The qualitative trends required to be captured by the model are, on the whole, independent of individual parameter setting.

### 3.4.3. *Parameter-tuned Evaluation*

Parameter tuning, and subsequent testing, of the NGM was performed by feeding in abstracted versions of the same input data from Spencer et al. (2011) and scoring the resultant output like the empirical findings. There are seven different trials types (single exemplar trial, three trials with objects presented in parallel, and three trials with objects presented simultaneously) to model, and each experimental condition includes three output proportions (subordinate, basic, and superordinate level generalizations) for a total of twenty-one output means to be compared. To ensure fair evaluation (and avoid over-fitting), I trained the NGM on three of the seven different trial types — training over a single exemplar, training over three basic-level matches in parallel, and training over three basic-level matches in sequence. These conditions are shown in red in Table 11. Testing was then performed on all experimental conditions from Spencer et al. (2011) varying the hierarchical organization and presentation-style of the input. Parameter tuning was performed by running a five-way step-wise (*step size* = 0.1) grid search of 1024 configurations (two salience distributions means, salience standard deviation, distance threshold, semantic incompatibility parameter).<sup>6</sup>

<sup>6</sup>Parameter tuning here is simply a method for assessing model fit and relative power, and should not be interpreted to reflect any measure of cognitive development. That being said, I report the final tuned







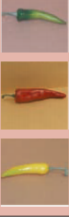


Trial	Single Exemplar	3-Subordinate Simultaneous	3-Subordinate Sequential
Objects			
3-Basic Simultaneous	3-Basic Sequential	3-Superordinate Simultaneous	3-Superordinate Sequential
			

Figure 11: Chart of all seven training configurations. Conditions used for parameter tuning shown in light red. Time during training is indicated within each block vertically; the objects in the parallel condition are co-present at the same time, while the “sequential” trials training objects are never co-present.

For each trial, there are three different generalization levels (sub, basic, super), each with a different proportion. To compute the distance from a parameter setting for the model and the empirical data, I summed the absolute value of the difference for the proportion for each level. Each trial configuration was run with 1000 simulated “participants” in the NGM. Even when tuned on only three out of seven of the experimental trial configurations, the model fit is very strong as shown in Figures 12 through 14. The mean divergence per trial between the experimental data and the output of the model is 0.049. 96% of trial configurations were within a single standard deviation of the empirical finding.

Overall, the output of the NGM is strongly consistent with human performance on generalization tasks in word learning. The model matches the general empirical trends of note independent of individual parameter values. When tuned on a training set, the output of the NGM has a mean divergence of approximately 0.05 per trial condition compared to the empirical finding. This is within a single standard deviation of human performance 96% of parameter values in Appendix B.1 (Table 24) as they are potentially useful for replication.

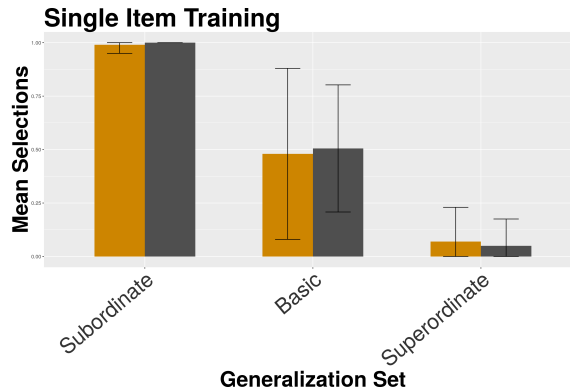
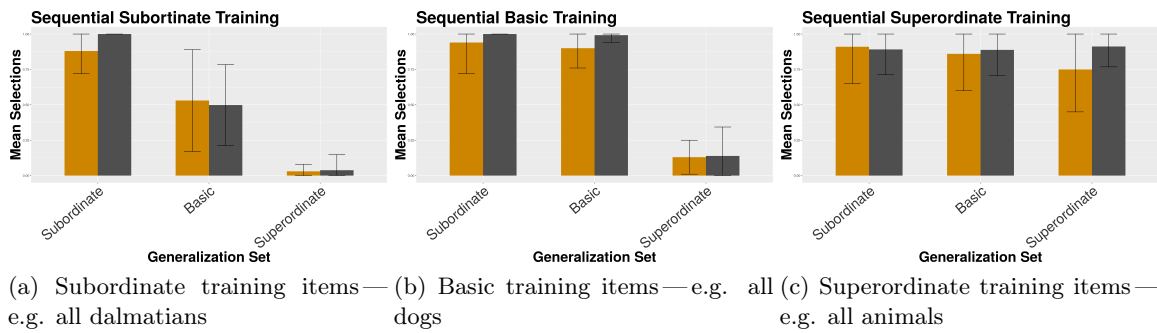
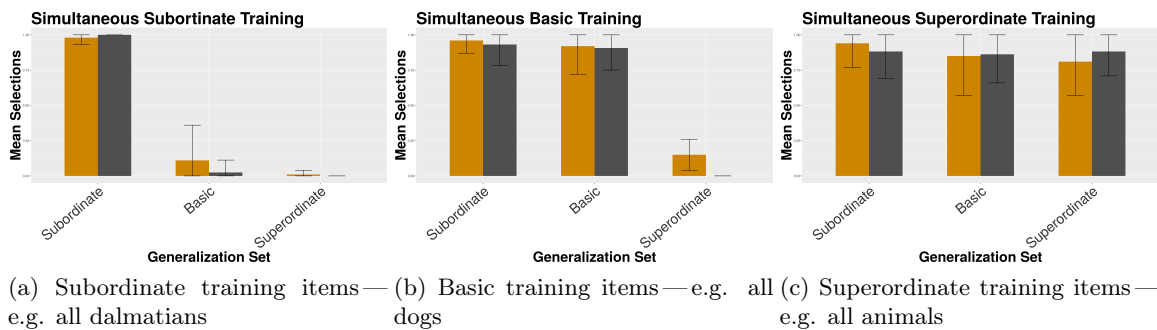


Figure 12: Training on a single item. Experimental results from Spencer et al. (2011) are shown in gold. Output of NGM in grey. Bars indicate standard deviations.



(a) Subordinate training items — e.g. all dalmatians (b) Basic training items — e.g. all dogs (c) Superordinate training items — e.g. all animals

Figure 13: Training items presented in sequence. Experimental results in gold. Output of NGM in grey. Bars indicate standard deviations.



(a) Subordinate training items — e.g. all dalmatians (b) Basic training items — e.g. all dogs (c) Superordinate training items — e.g. all animals

Figure 14: Training items presented simultaneously. Experimental results in gold. Output of NGM in grey. Bars indicate standard deviations.

the time. Several general patterns are captured here; the strong basic-level bias in generalization from a single, labeled training instance, the SCE that an increase to training-number leads to narrower generalization, as well as the PSE that sequential-presentation facilitates broader generalization than simultaneous-presentation. While for practical reasons the NGM was evaluated on a set of seven particular experimental conditions<sup>7</sup>, the underlying trends in generalization are robust under numerous related conditions (Gentner and Namy, 1999; Lawson, 2014b; Spencer et al., 2011; Lewis and Frank, 2018).

There are a few cases in which the NGM shows some divergence from the empirical output. The model’s rate of superordinate generalization given a parallel presentation of basic objects is near zero; while the experimental data for such a case is around 0.15. Additionally the model’s rate of generalization to any level given multiple superordinate objects as input is roughly uniform (at approximately 0.85 - 0.92). The empirical pattern is a slight declination from 0.91 subordinate-generalization to 0.75 superordinate-generalization. One potential explanation for this might be in the variance over both individual items as well as by participants. In particular, the range of variation in human participant responses is quite large. The standard deviation for superordinate object training cases varies from a low of 0.17 to nearly 0.3. This is a result both of cross-participant variation as well as item-level differences. Baseline rates of broad generalization are highly variable and governed by the prototypicality of the individual subordinate objects with respect to their containing class (Wang and Trueswell, 2017; Emberson et al., 2019). In Wang and Trueswell (2017) this ranges from a SCE gap of 50% for “dalmatians” being generalized to “dogs” whereas that effect is as low as 10% or even absent in the case of “goldfish” compared to “fish” or “monarch butterflies” compared with “butterflies.” A better grounding of individual item prototypicality and feature salience (Johnson and Mervis, 1997; Schyns and Gosselin, 2002) may explain these performance gaps.

---

<sup>7</sup>All the modeled experimental data was taken from *first-block trials* in order to avoid the effect of block-order uncovered by Lewis and Frank (2018) as well as the supplemental experiments in Spencer et al. (2011)

### 3.5. General Discussion

The problem of word learning highlights many of the interesting complications in language acquisition. From speech segmentation, to morphological analysis, to referential disambiguation, to the current question of semantic generalization, this is a task which *should* be extremely difficult (Quine, 1960) yet, in general, proceeds smoothly. The problem is certainly constrained by a range of cognitive biases which serve to effectively narrow the search space (Markman and Wachtel, 1988; Merriman et al., 1989; Markman, 1989; Gervain et al., 2008; Baldwin, 1991), yet an enumeration of such biases does not provide a mechanism underlying the acquisition process.

One of the contributions of the Bayesian program to cognitive science is that it *formally implements* a set of computations that can capture intuitions about statistical reasoning and makes related predictions. However, computation is contingent on the representations that have been posited so far. While the Bayesian inference account correctly predicts SCE, it cannot straight-forwardly account for other empirical effects such as PSE. This problem arises because the computations under Xu and Tenenbaum (2007b) are fundamentally a method for “hypothesis evaluation” without specifying the internal contents of the representations or how they are derived. In the end, it is the combination of representations and the computations performed over them that drive learning. These mental representations are unavoidably derived from the input but may diverge in systematic ways, particularly in light of the Immediacy of Linguistic Computation. Accounts at the computational-level that operate only via statistical trends will naturally miss some of the puzzle.

The Naïve Generalization Model (NGM) presented in this chapter offers an explanation of word learning phenomena grounded in category formation (Smith and Medin, 1981; Medin et al., 1987). As I argue, word learning is fundamentally to construct mental representations of words rather than strictly evaluate them. This does not necessarily maximize global probability of the output vocabulary, but rather the evaluation metric for meanings functions only over what is generated from input by the learner. The NGM explains the

mechanisms behind generalization for word learning in a manner that is consistent with and complementary to localist models of referent mapping (Stevens et al., 2016; Trueswell et al., 2013). Taken together, a more complete picture of word learning begins to emerge.

While the NGM was evaluated over the output of a particular experimental paradigm, the phenomena captured here (SCE and PSE) are a natural function of real-world learning contexts. Sometimes a child will encounter an object in isolation or across time, in other cases a group of referents is encountered together. This effect of *timing* in particular should not be surprising in light of the Immediacy of Computation. This is not to overlook the limits to the experimental paradigm employed by Xu and Tenenbaum (2007b); Spencer et al. (2011); Lewis and Frank (2018). Since the test-grid was always co-present with at least some training item, it is unclear how long the learning from this paradigm actually persists. In fact, data from Wang and Trueswell (2019) indicate that the particular competitors present in the test-grid have a notable impact on learning outcomes. Looking at a variety of inputs that learners receive (MacWhinney, 2000; Smith et al., 2015) can establish how such paradigms, and the findings based on them, map well onto or deviate from naturalistic contexts. Future work should also aim to connect the NGM as a model of word learning to the more general case of the mechanisms used for hypothesis evaluation for “structural” problems like the acquisition of argument structure (Naigles and Terrazas, 1998; Matsuo et al., 2016).

The training and testing images used in these studies were also pictures of natural objects. This presents a few potential limits. First, participants were already familiar with these natural objects, so it is difficult to factor out how big a role prior knowledge played in their test selections. Second, the “features” that can be described for such natural categories, and were available to the model, are essentially placeholders—and it may never be fully solved what the internal contents of conceptual representations really consists of. New studies using artificial stimuli or an independent measure of visual salience would allow the NGM to posit more precise predictions. Looking ahead, Chapter 4 presents findings from a novel eye-tracking paradigm which aims to address many of these issues and connect more

closely to studies of category learning and visual attention (Rehder and Hoffman, 2005a). Despite these limitations, the NGM provides a concrete mechanism for how words invite the creation of categories (Waxman and Markow, 1995; Waxman, 2003) and makes clear, testable predictions not only about the end state of word learning but also how and when learners update their intermediate representations over time. We do after all aim to reach past the *what* of word learning to understand the *how*.

## CHAPTER 4 : Selective Attention and the Intermediate Representation of Word Meanings

The trouble is that an observer who notices *everything* can learn *nothing* for there is no end of categories known and constructible to describe a situation.

---

Lila Gleitman.

Concepts are the building blocks of cognition. They are also an invisible yet inexorable filter: we do not have direct access to the external world, save through the lens of our mental encoding of it. This is particularly salient in light of the Immediacy of Linguistic Computation. Perceptual signals are inherently ephemeral: if cognitive computations are not made over transient and shifting information as it occurs, they cannot be made at all. Although this process appears seamless, we can interact with and reason over only the mental representations that emerge from the other side of this bottleneck.

I will argue in this chapter (as I do in the dissertation as a whole) that this local computation, limited to a category-based representation, is a benefit rather than a detriment to language processing and acquisition. As Lila Gleitman noted eloquently: “*The trouble is that an observer who notices everything can learn nothing, for there is no end of categories known and constructible to describe a situation*” (Gleitman, 1990). Our cognitive systems’ ability to represent and process input, to organize the sensory rip tides, the “blooming, buzzing confusion” (James, 1890) of perceptual experience and delineate sensible mental order is accomplishable only through the abstraction of categorization.

Perhaps the most concrete avenue to the study of mental categories is through the learning of words. Chapter 3 introduced a model of word learning grounded in category formation—the Naïve Generalization Model (NGM). The NGM explains the mechanism by which hearing novel words invites a learner to create a new category from component “features.” This is importantly different from the Bayesian theory of word learning because, under the NGM,

word meanings are *generated* by the learner rather than only selected for. Once a representation for a novel word has been generated, the learner is able to evaluate subsequent labeled objects with respect to this hypothesized meaning. This process is “naïve” in the sense that it does not optimize for any particular global value, with attention paid only to the local creation and evaluation of word meanings rather than statistical inference calculated in terms of a total distribution of input. While the NGM is able to capture a range of experimental findings from the Immediate Generalization Paradigm (Xu and Tenenbaum, 2007b; Spencer et al., 2011), that particular paradigm faces certain limitations which I hope to address here. At the same time as addressing those experimental issues, I aim to test further predictions of the NGM—relating to the time-course of learning and the intermediate stages of representation—in contrast with statistical or *evaluation*-based accounts.

#### 4.0.1. Design Constraints on Word Learning

In previous experiments regarding generalization in word learning (Xu and Tenenbaum, 2007b; Spencer et al., 2011), participants were provided with an unambiguous word-label for a set of one or more objects. Given a test-grid of other referents, we can probe what level of generalization a learner has posited for the novel word-label by measuring subsequent selections from the test ‘world’. While informative, this suffers from several limitations.

First, there is no time-course of learning in such experiments. Since at least some of the training stimuli are always co-present with the test-grid, it is unclear if semantic generalizations made in these experiments persist across time. In such cases, we would like to make inferences of the underlying mechanism used for word learning but have access only to the end result of generalization. We lack information about the *path* which brought a participant to such a generalization. Separating training and testing phases should aid in this as well as likely providing a more realistic approximation of real-world word learning conditions. What’s more, the particular “filler” objects available in the test-grid impose different meanings indirectly based on their contrastiveness to the word being learned (Wang and Trueswell, 2019).



Second, since previous experiments have used photographs of common objects as stimuli, this has the potential to conflate the creation of a semantic category from the act of assigning a novel label to an existing category. Particularly, since the NGM hopes to explain the *category formation* aspect of word learning, it motivates the use of artificial stimuli which participants would have no prior exposure to or semantic categories over. While some work has shown that behavior regarding homophony is consistent when learning from either natural or artificial stimuli (Dautriche and Chemla, 2016), this has not been directly studied for the presentation-style effect. I would like to further advance this by using stimuli built from a limited set of spatially separated features. This allows the use of selective attention via eye-gaze to be taken as a measure of semantic representation during learning.

It is important that the measurable output on experiments like Xu and Tenenbaum (2007b) are object *selections* at the end of learning. For instance, in one condition participants tend to select all the dogs from a general test domain, while in another condition they select only the *dalmatians*. Researchers can make inferences regarding the internal representations corresponding to a learned word, but this paradigm does not provide direct evidence as to the mental path or time-course over which such semantic generalizations emerge. This is particularly important given the significant effect which temporal presentation has on learning (Spencer et al., 2011; Carvalho and Goldstone, 2014; Lawson, 2017).

The current experiment aims to test predictions made by hypothesis generation accounts like the NGM in comparison to the statistical *evaluation* view through the use of eye-tracking in a novel word learning task. The experimental paradigm uses artificially created stimuli with spatially distributed features. Through this we can probe the way that novel categories are formed (during a process of word learning) rather than simply a mechanism by which phonological labels are assigned to existing categories. As attention and eye-movement are tightly linked (Deubel and Schneider, 1996; Henderson, 1992), we can use eye-gaze to probe the encoding and processing of individual features, enabling the measurement of the time-course of intermediate representations (intentions) throughout learning rather than simply

their selections (extensions) at the end of learning. The particular manipulation of interest is the role of presentation-style of stimuli (either in parallel or in sequence) and its role on generalization in word learning.

This chapter is hardly the first to employ this kind of stimulus structure to study category learning. Since at least Shepard et al. (1961) there has been an effort to understand how category structure affects learning and categorization. In a classic paper aimed at distinguishing exemplar and prototype representations of concepts, Medin and Schaffer (1978) constructed the “5-4 category structure” which represents stimuli as objects comprising two opposing categories and varying along four binary-valued dimensions. Since then, many subsequent papers have utilized the same kind of category structure; see Smith and Minda (2000), for an extensive list of references.

This chapter is also not unique in using eye-gaze as a measure of selective attention; see for instance (Rehder and Hoffman, 2005a,b, among others). However Rehder and Hoffman (2005a,b) and studies like it, following Shepard et al. (1961), typically involve dozens of repeat exposures to instances of the to-be-learned category along with supervised feedback (including the labeling of instances when the learning is wrong in their categorization). This crucially differs from the type of input from which words — and natural language more broad — are acquired. It is well understood that children receive virtually no actionable negative evidence acquisition (Brown and Hanlon, 1970; Braine et al., 1971; Bowerman, 1988; Marcus, 1993). Explicit feedback — in the form of overt corrections or other behavioral cues — is relatively rare, and crucially unreliable. Thus it is important for studies of word learning to reflect the real constraints and design principles of the problem at hand. This motivates a new paradigm which combines the desirable elements of category learning studies with the constraints of word learning, letting us probe the generation and evaluation of meanings as they unfold over well-controlled stimuli.

#### 4.0.2. Hypothesis Generation vs. Evaluation

Beyond the particular details of the NGM or the Bayesian inference theory in particular, I contrast two broad classes of theories: hypothesis *generation* and statistical *evaluation*. While of course any account in some sense requires both the generation and evaluation of hypotheses, these theories differ in where the “heavy lifting” resides — whether this is primarily in the representations, or the types of computations that are performed over them.

On the *generation* theory, learning is constrained by the immediacy of linguistic computation. Learners do not and cannot store a record of all the underlying signal. Occurring features are only retained as part of an initial categorization decision (e.g. the initial hypothesis in the NGM, or the way that speech is represented in terms of an activation over discrete categories in Chapter 2). The subsequent evaluation procedure is “lazy,” operating only over the intermediate representation (current hypothesis) under a “good-enough” metric: if future referents appear consistent with my guess, then keep it, otherwise try something else. There is no global optimization to find the “best” hypothesis over the entire input set.

On the statistical *evaluation* or accumulation theory, learners attempt to sample a record of the total input data. When encountering subsequent referents, all the features could be attended to in order to update the statistical occurrence record. This record can then be used to perform subsequent inference (whether via Bayesian updating or another scheme). This is unlike an exemplar model, which actually represents the final category as a combination of all the instances that belong to that category (Brooks, 1978; Estes, 1994; Hintzman, 1986; Kruschke, 1992; Lamberts, 2000). Yet, the accumulated evidence required to evaluate different hypotheses (the intermediate representation for a Bayesian model) looks quite similar: a sample taken from the whole input distribution.

These two accounts thus differ quite starkly in the intermediate representations that they predict to occur throughout the learning process. On the hypothesis generation account, learners posit and evaluate a single representation, without storing the direct evidence which

led to that belief. The generation of hypotheses from locally available input is the primary bottleneck to word learning rather than the method by which such hypotheses are evaluated. On the hypothesis evaluation account, learners need to accumulate evidence throughout learning prior to its subsequent filtering.

In the remainder of the Chapter, I introduce and present results for a new eye-tracking paradigm for semantic generalization in word learning. This paradigm combines many of the desirable properties from previous work on category learning while imposing the described constraints required for word learning, including limited training instances and no supervised feedback. By using artificial stimuli with spatially distributed features and a division between training and testing phases, I can probe the way that novel categories are formed during word learning, rather than simply the mechanism by which phonological labels are assigned to existing categories. The use of eye-gaze as a proxy for selective attention to individual features serves as a method to study the time-course of intermediate representations throughout learning. This supports more directly testing the predictions of a hypothesis generation account (such as the NGM) in contrast with a statistical evaluation/accumulation account of word learning.

## 4.1. Experiment

### 4.1.1. *Design*

Participants were taught a total of eight words. The learning set was composed of two words in each of four “blocks” — one for each domain of stimuli (birds, bugs, houses, submarines) in counter-balanced order. Each block was divided into several phases: familiarization, learning/training (in one of two experimental conditions), and testing. The cover story told to participants was that they are going to virtually visit the lost world of “Atlantis” and learn some of the language used there. In this way, subjects knew they are going to be exposed to a language which differs from English but without preconceived notions of the properties of such a language. What’s more, they could expect to see novel or foreign objects but understand that their grouping by words reflects a reasonably human and non-arbitrary pattern.

In the familiarization phase, participants were shown a “scene from atlantis”: a grid consisting of a random sample of twenty-four objects from the current domain. This was displayed for a period of twenty seconds over two slides (ten seconds per slide, and twelve objects per slide). This was intended to give participants a sense of the potential contrasts and features present in each domain prior to their exposure to any word learning instances.

In the training stage, each subject learned two words per block (for a total of eight words overall) in an ostensive labeling task. Every word learned is underlyingly defined in terms of two relevant features (RF), with the other three features in free variation (or NF for “not-relevant feature”). For example, a type of bug might have [+RoundHead,+FourLegs] while varying in terms of Tail, Wings, and Body. See Section 4.1.2 for details on the stimulus construction. The spatially centered “middle” feature was always in free variation, and thus never a defining characteristic. Each word was assigned a randomly chosen disyllabic nonce-word audio label (see Appendix C.1 for the full list) and presented to participants in alternation. Subjects heard/saw each word in interleaved order: i.e. Word<sub>1</sub>, Word<sub>2</sub>, Word<sub>1</sub>, Word<sub>2</sub>, etc. Each audio label was played twice (one second apart) and participants were shown an instance of an object(s) from the corresponding category. The training for each word consisted of five different labeled instances. Objects were ordered such that the first three never accidentally shared the same value on any of the features that were in free-variation (NFs).

The definitions for each word were randomized, but configured to never overlap so as to minimize and properly control for ambiguity. Thus if Word<sub>1</sub> had RFs defined over the Head and Legs, then Word<sub>2</sub>’s RFs would need to consist of the Back and the Tail — a single spatial dimension was never a relevant feature for both words. Since each word had two (binary valued) RFs, and the RFs between the two words were always in complementary distribution, then out of the 32 possible objects there were 6 unique object matches for each word, 18 objects that were matches of neither category, and 2 objects that would be matches of both categories. In order to avoid the potential confound of mutual exclusivity, participants were

never exposed to objects that could be matches of both categories. Instead, the five training objects were sampled from the six possible unique matches for each word.

There was one between-subject condition of “presentation timing:” parallel presentation and sequential presentation. In the parallel condition (3-1-1), subjects were initially exposed to three labeled instances for each word appearing on screen simultaneously. They were then subsequently given two more labeled training instances, one at a time. While in the sequential training condition (1-1-1-1-1), each of five instances is displayed separately one at a time. The presentation-style training conditions are visualized in Figure 15.

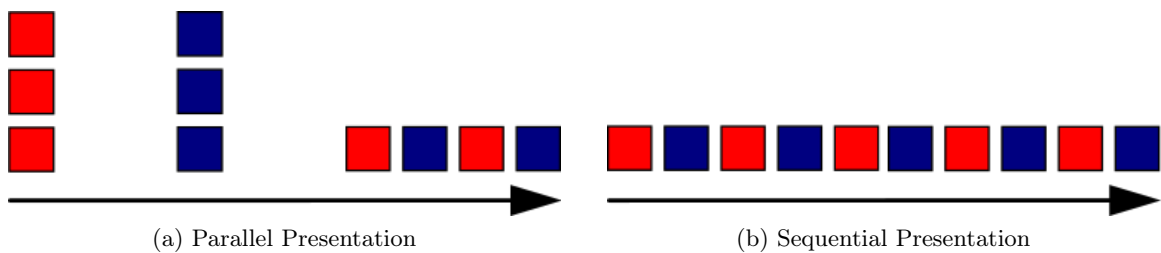


Figure 15: Visualization of parallel vs. sequential training conditions. Word<sub>1</sub> in red and Word<sub>2</sub> in blue. The total number of exemplars and display time remained constant across conditions.

The location of training objects was randomly chosen on each trial from among four potential boxes (located on the left, right, top or bottom of the screen). Each box was 640 by 480 pixels in size. A cross-hair was displayed in a randomly chosen box between each trial. The number of training objects per word was held constant at five regardless of training condition. Exposure time per instance was also held constant at five seconds. This means that the initial trial for participants in the parallel condition was fifteen seconds, so that total exposure time was consistent across groups.

Since the testing instructions were somewhat long, a brief “re-familiarization” phase immediately followed training (and testing instructions) but preceded testing itself. This was intended to prevent participants from accidentally flipping the labels for the two words they just learned. In this re-familiarization, an example item (always the first training instance encountered) was re-presented to the participant with instructions for the upcoming test

phase. For example, “This is a *Gronen*.” (sees example *gronen* on screen) “Now, please select all of the other *gronens*.” This re-familiarization exposure provided an opportunity to take another measure eye-gaze after the learning phase has completed.<sup>1</sup>

In the testing phase, participants were shown a sequence of four cards, each containing six objects (arranged in a 2x3 grid), asked to select all the matches to the learned-word by clicking. Since words were underlyingly defined in terms of two features RFs, call them A and B, then each test card contained one *Narrow* match (+A+B), two *Broad* matches (+A-B and -A+B) as well as three non-matches (-A-B). Test items are assigned a random location on the 2x3 grid for each test card. For each word, participants make selection from four cards, with words tested in blocked sequence (i.e. all of the Word<sub>1</sub> test cards were completed before moving on to the Word<sub>2</sub> test cards). The name of each word was displayed in text before each card, with audio then played twice (one second apart). The test card remains on-screen until participants have finished making selections.

Before any real test trials are undertaken, participants were shown a pair of “practice” test cards with pictures of real objects (e.g. “click on all the dogs”). One of the test practice cards included just a single match, while the other practice card included multiple matches to be selected. This was intended to emphasize to participants that they might need to click on different numbers of objects depending on the configuration of the actual test cards and their hypothesized meanings for each of the learned words.

#### 4.1.2. Stimuli

The stimuli in use in this experiment here have numerous benefits. Using artificially created stimuli rather than pictures of naturally occurring objects ensured that participants did not have prior category knowledge of the distributions/concepts being learned. These stimuli additionally ensured that objects were defined in terms of a pre-configured, finite number of distinctive features. This kind of structure resembles those which learners naturally impose

---

<sup>1</sup>Since the re-familiarization item was always first object encountered during training, this introduced a confound between initial and final exposure objects. This is unfortunate, but does not interfere with the primary analyses or theories of interest in this Chapter.

when asked to categorize stimuli of this type (Medin et al., 1987). The stimuli also control for the degree of similarity or dissimilarity between any two objects. For instance, it is simple to compute whether a pair of referents share two feature values while differing on a third. This sort of particular feature computation is difficult to estimate over natural categories like those modeled by Xu and Tenenbaum (2007b) and in Chapter 3. Finally, by constructing stimuli such that distinctive features are separated spatially and of equal size, we can estimate which feature is being attended to throughout learning via eye-gaze.

The stimuli were sets of objects from one of four separate domains termed “bugs,” “birds,” “houses,” and “submarines.” Each object is an instance from a class with five potentially distinctive, binary features; this means there are  $2^5$  or 32 possible objects per domain. The features are spatially distributed and non-overlapping. This ensures that each gaze location corresponds to the spatial region associated with exactly one semantic feature. A listing of the features defined for each domain is provided in Appendix C.2. Stimuli from each of the four domains are given below in Figure 16. The objects on the left and right are the maximally divergent pairs (differing on all features) within each domain.

Since participants were exposed to training samples from a word defined by exactly two features, there are several possibilities of the resultant semantic representation that they might posit. An individual learner might properly encode the entire shared distribution, i.e. both features [+A+B] to make a *Narrow* generalization. The learner may instead end up encoding only a single feature, i.e. either [+A-B] or [-A+B] to make a *Broad* generalization. The more dense the semantic representation becomes, the more *narrow* the generalization—the cardinality of the set of selected test objects becomes smaller and vice versa. For example, if only a single distinctive feature were encoded, that semantic representation maps onto 16 possible test items. Whereas if both distinctive features were encoded, that semantic representation maps onto only 8 possible test items. The intuition that sparse concepts correspond to broad categories (with more members) while dense representations correspond to narrower categories (with fewer members) is perhaps clearest when considering a concrete



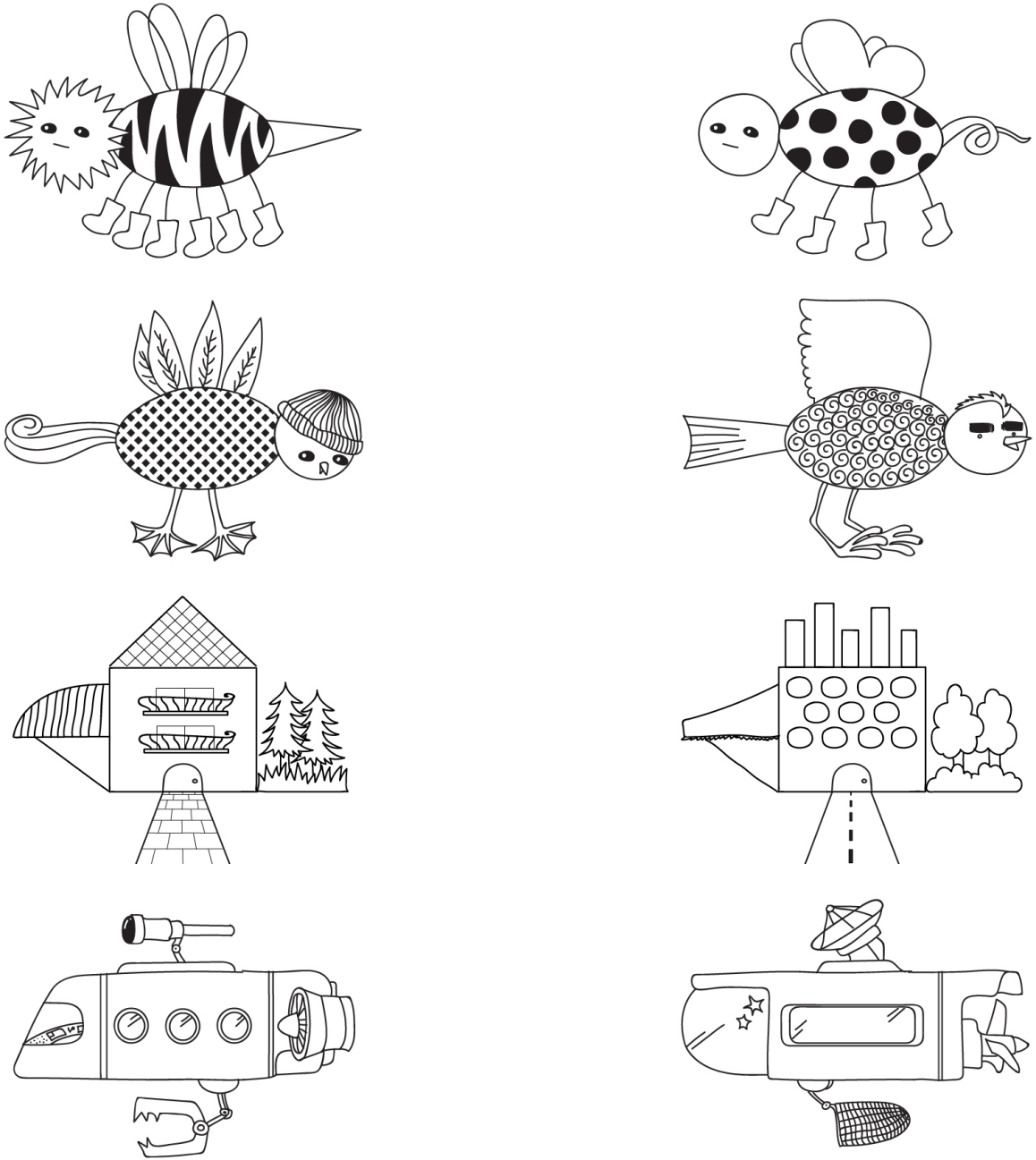


Figure 16: Sample pairs of maximally divergent stimuli (differing on all five features) for each domain.

example. For instance if one learned a word that means [+SQUARE], that clearly picks out more members of the world than a word which is learned as [+RED,+SQUARE].

#### *4.1.3. Procedure*

Sixty participants were recruited through the University of Pennsylvania Experiment Registration, and randomly, evenly assigned to one of two experimental timing conditions (parallel or sequential timing). Participants were paid \$10 for participation and limited to monolingual native speakers of English, who were 18-years or older.

Participants were tested individually, using a Tobii tx300 eye-tracking for stimulus presentation and data collection recording at 120 Hz. This was subsequently downsampled to 60 Hz for analysis. The display screen was 1920 by 1080 pixels. The experiment script was coded in Matlab and run on a separate laptop controlling output to and taking data from the Tobii eye-tracker. Verbal instructions were also read aloud by the experimenter prior to starting.

#### *4.1.4. Measures for Analysis*

Two classes of measurements were taken: final generalizations based on the overt (clicked) selection of referents during testing, and the intermediate representation of words based on eye-gaze to individual dimensions throughout training/learning.

In order to score final learning/generalization, each test card was scored as corresponding to one of four possible outcomes. Narrow (if only the +A+B was selected), Broad<sub>1</sub> (if both the +A+B and +A-B objects were selected), Broad<sub>2</sub> (if both the +A+B and -A+B objects were selected), or “Outside” (if anything else, including any of the -A-B objects were selected). A proportion (of Narrow vs. Broad vs. Outside) was then computed for each “deck” of the four test cards for each word. This was done so as not to throw out participants who might have accidentally clicked on one of the test objects incorrectly. Overall, within-deck agreement was quite high, with 92% of all test decks sharing an outcome on 3 of the 4 cards. In such cases, the word was then labeled as having been learned as either Narrow or Broad based on the deck outcome. For words whose test decks contained something else (some “outside” -A-B selections), I computed the learned definition by intersecting the feature values present for all of the selected items in the deck. Many of these words were given a

consistent definition, just one that was not strictly supported based on the training input. Any words whose intersected definition did not share any features was labeled “inconsistent” and excluded from analysis.

In order to extract eye-gaze measures, I defined AOIs for each feature as the smallest bounding box that totally encapsulated the feature’s spatial region, plus an additional buffer of 20 pixels on each side. I defined the RF (relevant feature) set as the combination of both AOIs for the two underlying consistent features defining that word. For instance, if a word in the Submarine domain were underlyingly defined by the front and bottom, that is indicated in blue on Figure 17. The NF (non-relevant feature) set is the combination of both AOIs for features that are in free variation for a given word. I did not include the center region, which was in free-variation for every word, in the NF set. A few factors motivate this. First, I initially hypothesized that the center position in the image may be privileged for initial fixations<sup>2</sup>. Second, excluding the center region from analysis of gaze allows AOIs to remain non-overlapping. Lastly, and most importantly, the present analysis scheme holds both the number and total area of AOIs in the RF and NF constant with one another, which enables a fair direct comparison.

This is primarily reported as the total proportion of gaze time spent looking at the RF-set and the NF-set (based just on the time when participants were looking inside of the stimulus bounding box). However, the proportion of looking time to the RF-set does not distinguish between cases in which total fixation time appears high because it is split more or less uniformly between both  $RF_1$  and  $RF_2$  or a case in which fixations are made exclusively to a single feature. I thus also computed a measure of “RF-skew.” Since the total size of the difference between gaze-time to  $RF_1$  vs.  $RF_2$  will vary as a function of the looking time to the RF-set, we defined RF-skew as follows:

---

<sup>2</sup>This turned out not to be the case. See Appendix C.3 for heatmaps showing the distribution of eye-gaze to different regions for each domain throughout the experiment

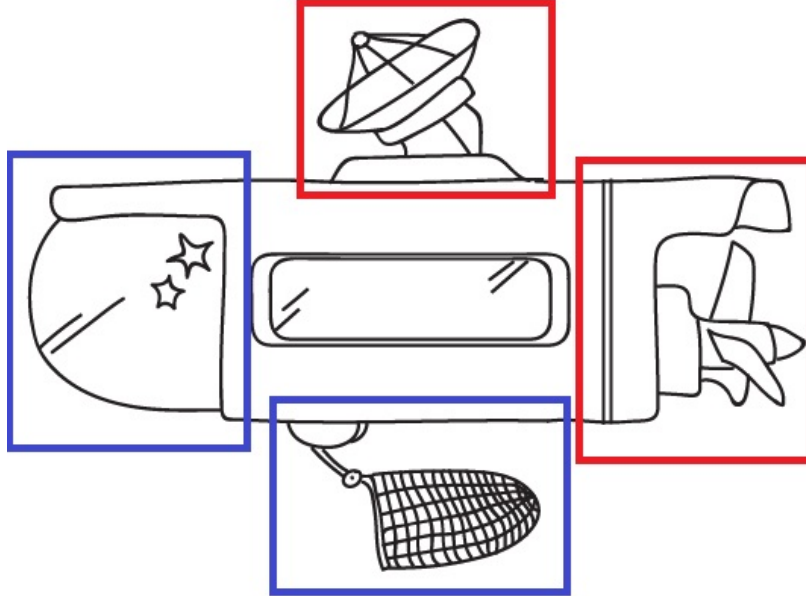


Figure 17: Example AOI calculation. The RF might be the front and the bottom in blue. While the NF are the tail and the top in red.

$$Skew = \frac{abs(f(RF_1) - f(RF_2))}{(f(RF_1) + f(RF_2))} \quad (4.1)$$

Skew scales between 0 and 1, where 0 indicates uniform looks to  $RF_1$  and  $RF_2$  while a value of 1 corresponds to looks exclusively to only one of the two features.

#### 4.1.5. Exclusions

I define a word as “mislearned” if at least two of four test cards were scored as “Outside” (i.e. given neither Narrow nor Broad interpretations). Subjects who mislearned more than 50% of words were dropped from analysis. Additionally, words whose trackloss during the re-familiarization phase was greater than 30% were dropped. For words that were mislearned, I computed each participant’s posited representation by intersecting the feature values from their set of selections (i.e. finding the maximal set of feature values that their selections all share). Words which did not have a consistent definition under this feature-intersection scheme, were dropped as being “inconsistent.” This resulted in a total of 315 learned words post-exclusion across 53 subjects.

#### 4.1.6. Predictions

On a hypothesis generation account, presentation timing is predicted to affect generalization outcome. In particular, parallel presentation should lead to higher rates of narrow generalization than sequential presentation. When multiple stimuli from the same category are initially presented together, then it is comparatively easy to notice and encode the features that they all have in common. Conversely, when the same stimuli are presented in sequence, then the learner can construct an initial hypothesis only once (and on the basis of just the first exemplar). After the first training object has been removed, then due to the immediacy of computation, the learner can only refer back to their intermediate hypothesis. If they happen to have encoded one of the RFs, then there is no onus to change that representation or update it to go searching for what else the referents might all have in common. Whereas if the initial hypothesis instead included one of the NFs, then the learner may subsequently realize that their initial guess was wrong, but they no longer have access to the full set of information about the initial referent. As with the comparison of speech representations in Chapter 2, this is a *Markovian* process: learners encode a state of belief, but do not retain the precise experiential statistics which led to that belief. The learner is left to make a new hypothesis based primarily on the current referent.

The statistical evaluation account does not predict this kind of sensitivity to temporal presentation. On this view, learners need to accumulate a sample of evidence from the input distribution so it should be immaterial whether the same stimuli are displayed all at once or distributed over a sequence.

The two theories also differ in their predictions about the distribution of eye-gaze during training. On the hypothesis generation view, attention should be limited just to those dimensions which correspond to a particular hypothesis. Once a hypothesis has been generated, the primary goal of the learner is to look for confirmatory evidence of that hypothesis. The statistical evaluation view requires learners to attend to multiple dimensions throughout learning in order to accumulate evidence towards one meaning over another.

## 4.2. Results

### 4.2.1. Effect of Timing on generalization

The hypothesis generation account predicts that parallel presentation should lead to higher rates of narrow generalization than sequential presentation, whereas the accumulation/evaluation account predicts no difference in generalization level across timing conditions.

A mixed effects logistic regression analysis was conducted on trial-level data. The main dependent variable was Narrow-Generalization: whether participants chose items during testing that correspond to a narrow definition for that word. The independent variables were experimental condition Timing (parallel vs sequential), Block number (first through fourth), and Word number (either the first or second word in the current block), as well as their interaction. I used the maximal random effects structure that converged; this structure included random intercepts for participants and domain. I tested for significance of factors in the model by using likelihood ratio tests on the  $\chi^2$  values from nested model comparisons with the same random effect structure (Matuschek et al., 2017). The best fitting model is shown in Table 8. This model was a better fit than one that did not include the main effect of Timing,  $\chi^2(2) = 11.46$ ,  $p < .001$ . These results demonstrate that there was significantly more narrow generalization in the parallel condition compared with the sequential condition. The full range of learning outcomes is shown in Figure 18.

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	-3.37 [-4.69, -2.04]	-4.98	< .001	0.03 [0.01, 0.13]
Timing	1.48 [0.53, 2.42]	3.06	.002	4.38 [1.7, 11.27]
Block	1.3 [0.73, 1.86]	4.50	< .001	3.66 [2.08, 6.44]
Word	-0.19 [-0.66, 0.28]	-0.78	.432	0.83 [0.52, 1.32]
Timing x Block	0.99 [0.32, 1.65]	2.92	.004	2.68 [1.38, 5.2]
Timing x Word	0.43 [-0.04, 0.91]	1.80	.073	1.54 [0.96, 2.47]
Block x Word	0.23 [-0.23, 0.7]	0.99	.324	1.26 [0.79, 2.01]
Timing x Block x Word	-0.11 [-0.57, 0.35]	-0.47	.639	0.9 [0.56, 1.42]

Table 8: Output of the best fitting model predicting Narrow generalization. Bracketed values are 95% confidence intervals.

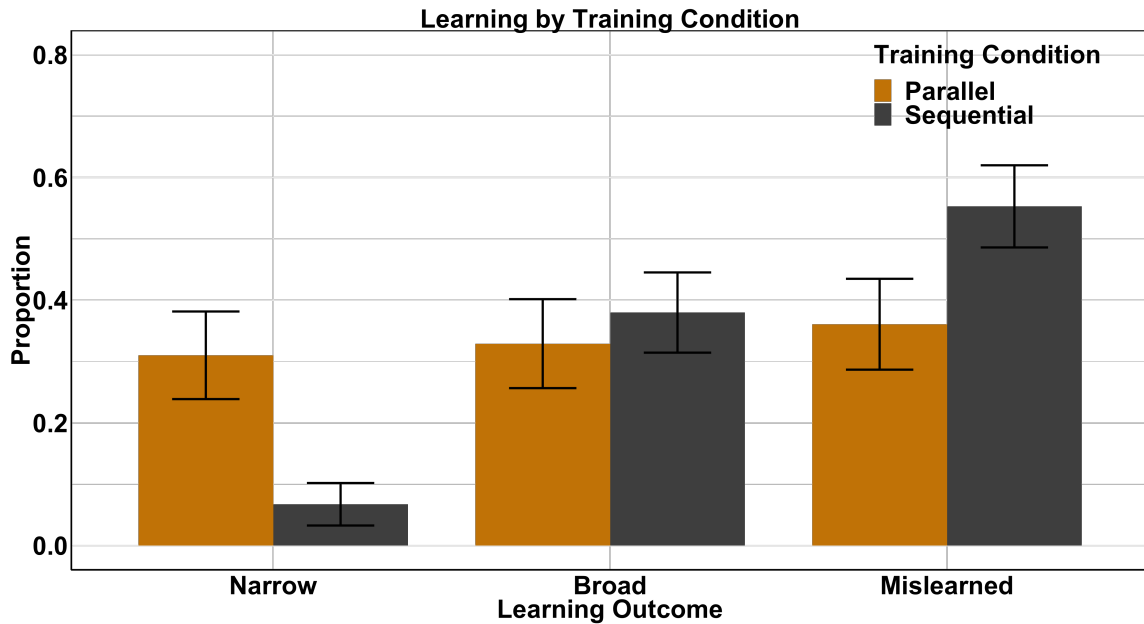


Figure 18: Bar graph of proportion of learning outcomes as a function of training condition (parallel vs. sequential)

#### 4.2.2. Relationship between eye-gaze and learning outcome

The hypothesis generation account predicts that selective attention (as measured through eye-gaze) should be directed primarily to features encoded in the current hypothesis (intermediate representation) for each learner, even if that hypothesis turns out to be wrong. Thus words that were correctly learned (either under a Narrow or Broad) interpretation should have gaze-time allocated to RFs compared with NFs. Likewise, when participants end up mislearning a word, they should be spending time evaluating the hypothesis that they believe to be correct (even though it's not strictly supported by the data); therefore mislearned words should have more gaze-time allocated to NFs than RFs.

While the accumulation/evaluation account might reasonably predict that learned words garner increased looks to the RF-set, in the case of mislearned words this should not occur if participants extracted a representative sample of information from the input distribution. The evaluation account therefore would predict that “mislearning” trials are ones during which participants are still searching for evidence and thus distribute their gaze uniformly across the different features (attempting to sample more information).

I conducted a pair of mixed effects linear regression analyses on trial-level data. In the first, the main dependent variable was gaze-time to the RF-set (during refamiliarization). The independent variables were Learning Outcome (binary learned vs. mislearned), Timing (parallel vs sequential), and Block number (first through fourth), as well as their interaction. I used the maximal random effects structure that converged; this structure included Domain name as random slopes for Word number. In the second the main dependent variable was gaze-time to the NF-set (during refamiliarization), while the independent variables did not differ. The maximal random effects structure that converged included random intercepts for Definition (which spatial regions were mapped to RFs vs. NFs), and Domain name as random slopes for Word number. I tested for significance of factors in each of the best fitting models (predicting RF-gaze and NF-gaze) by using likelihood ratio tests on the  $\chi^2$  values from nested model comparisons with the same random effect structure. The best fitting RF-gaze model was a better fit than one that did not include the main effect and interaction terms of Learning Outcome,  $\chi^2(2) = 51.813$ ,  $p < .001$ . The best fitting NF-gaze model was a better fit than one that did not include the main effect and interaction terms of Learning Outcome,  $\chi^2(2) = 32.319$ ,  $p < .001$ . This trend is visualized in Figure 19.

Since *Narrow* interpretations require encoding both RFs, whereas *Broad* interpretations only require encoding a single RF, then we further predict a difference in RF-skew based on these different learning outcomes. For Narrow-learned words, attention should be distributed uniformly between the two relevant features, whereas Skew should be higher for Broad-learned words (which only prompt attention to a single feature). To test this I performed



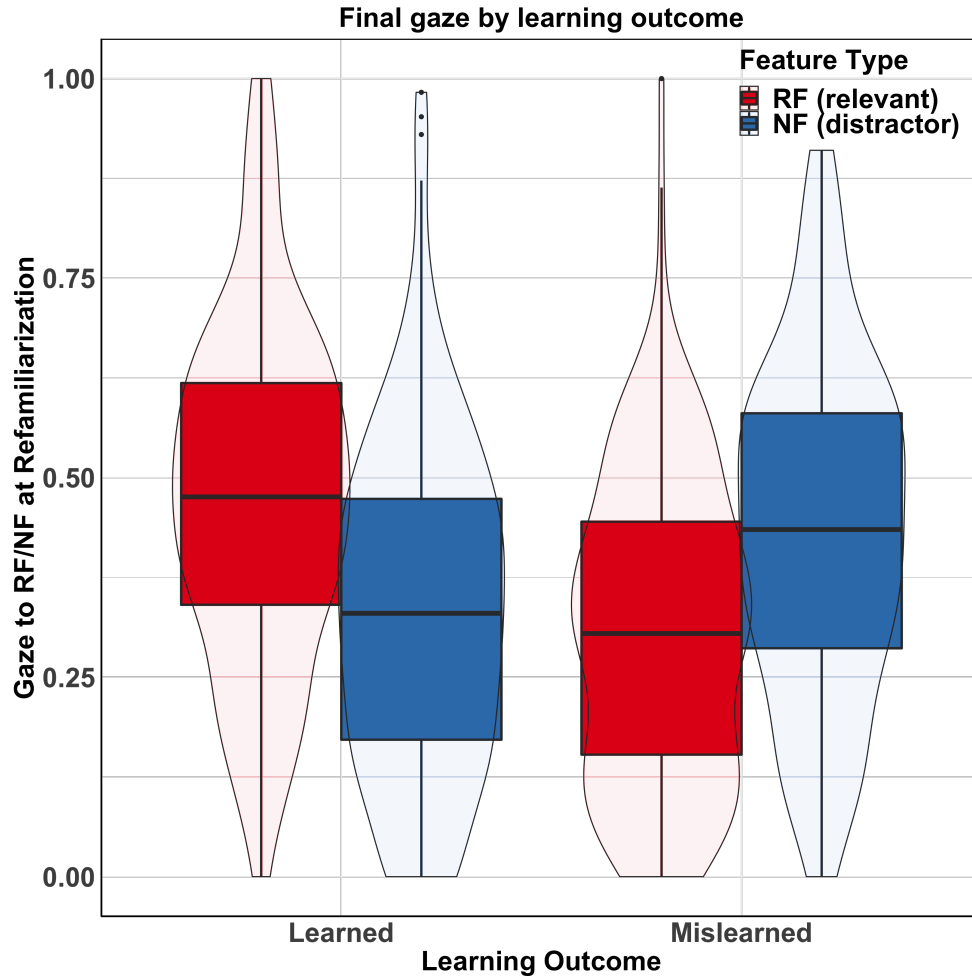


Figure 19: Violin plot of the proportion of gaze-time allocated to RFs vs. NFs as a function of Learning Outcome (learned vs. mislearned)

a mixed effects logistic regression analyses on trial-level data. The main dependent variable was Narrow-Generalization: whether participants chose items during testing that correspond to a narrow definition for that word. The independent variables were Timing (parallel vs sequential), Block number (first through fourth), and RF-Skew, as well as their interaction terms. I included the maximal random effect structure that converged; this consisted of Subject and Definition (which spatial regions were mapped to RFs vs. NFs) as random intercepts. The best fitting model was a better fit than one without the main effects and interactions of RF-Skew,  $\chi^2(2) = 8.2247$ ,  $p = .084$ . This trend is visualized in Figure 20.

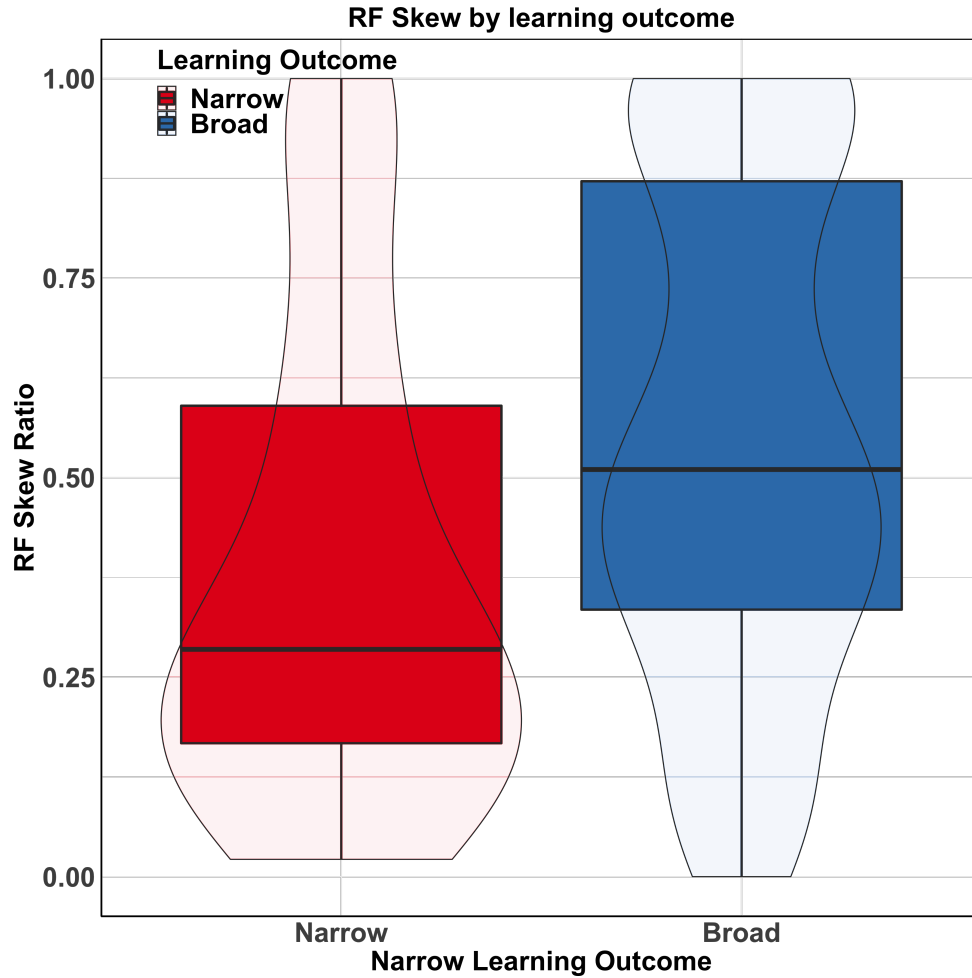


Figure 20: Violin plot showing RF-Skew as a function of learning outcome (Narrow vs. Broad)

The hypothesis generation account predicts that gaze should be primarily directed to features involved in evaluating the current hypothesis, regardless of whether that will turn out to be correct (or is even supported by the data in a global sense). Conversely, an evaluation account predicts that unless the learner is already confident in the meaning, then they should continue to search/sample the whole set of features in order to uncover the underlying statistical pattern. Thus, I tabulated the proportion of looking time spent attending to the features which define each participant’s eventual test selections. The hypothesis generation account predicts that this measure — gaze time to “posited features” — should not vary based on whether that hypothesis will turn out to be correct. Whereas the statistical evaluation

account predicts that gaze-time to posited features should be greatest when learners have already converged on a correct interpretation. I did not conduct a formal analysis of this prediction, but I have included a plot of this trend in Figure 21.

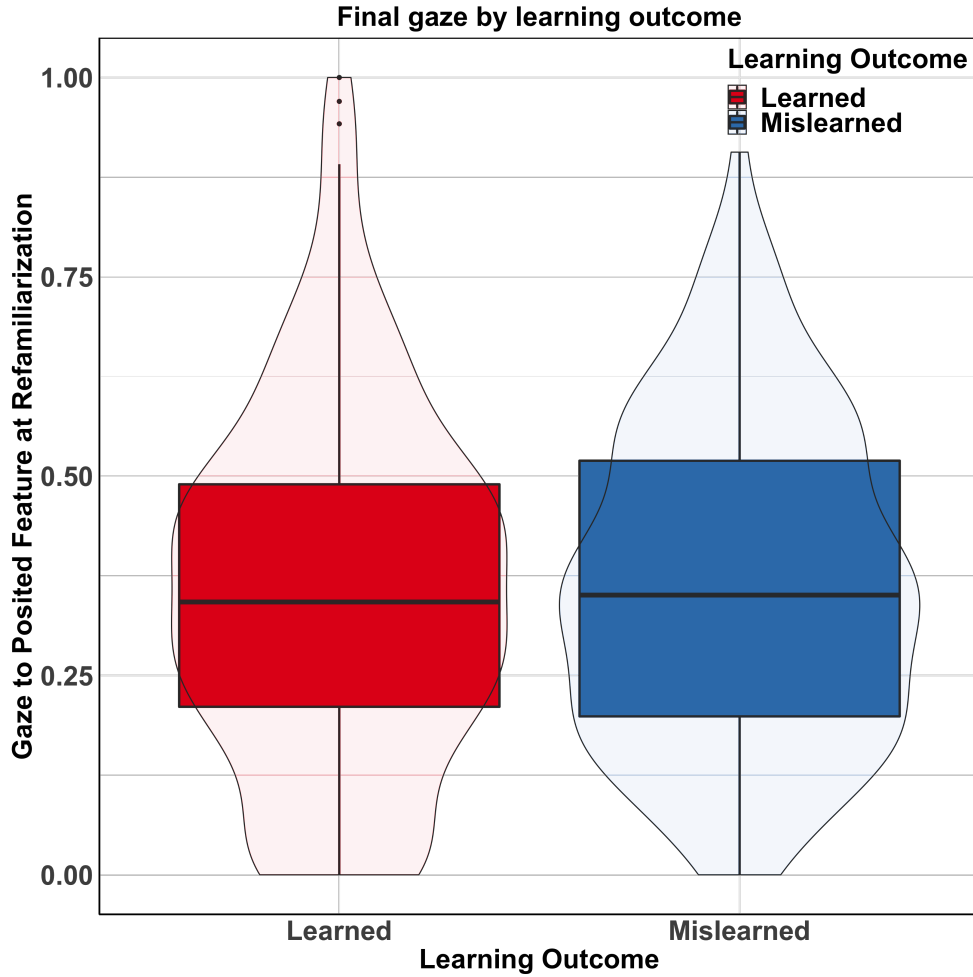


Figure 21: Gaze to posited features is not affected by learning outcome (learned vs mislearned).

Finally, the hypothesis generation account predicts that attention during the first occurrence will govern the initial hypothesis that gets created. If more attention is allocated to a relevant feature, this will increase the likelihood of successful learning on that trial. Conversely, if more attention is allocated to a distractor feature (NF), the likelihood of successful learning will be substantially lower. The accumulation/evaluation theory does not make this prediction, since information extracted during initial exposure does not occupy

a privileged position relative to any other source of evidence. I did not conduct a formal analysis of this prediction, but the trend is clearly visible in Figure 22.

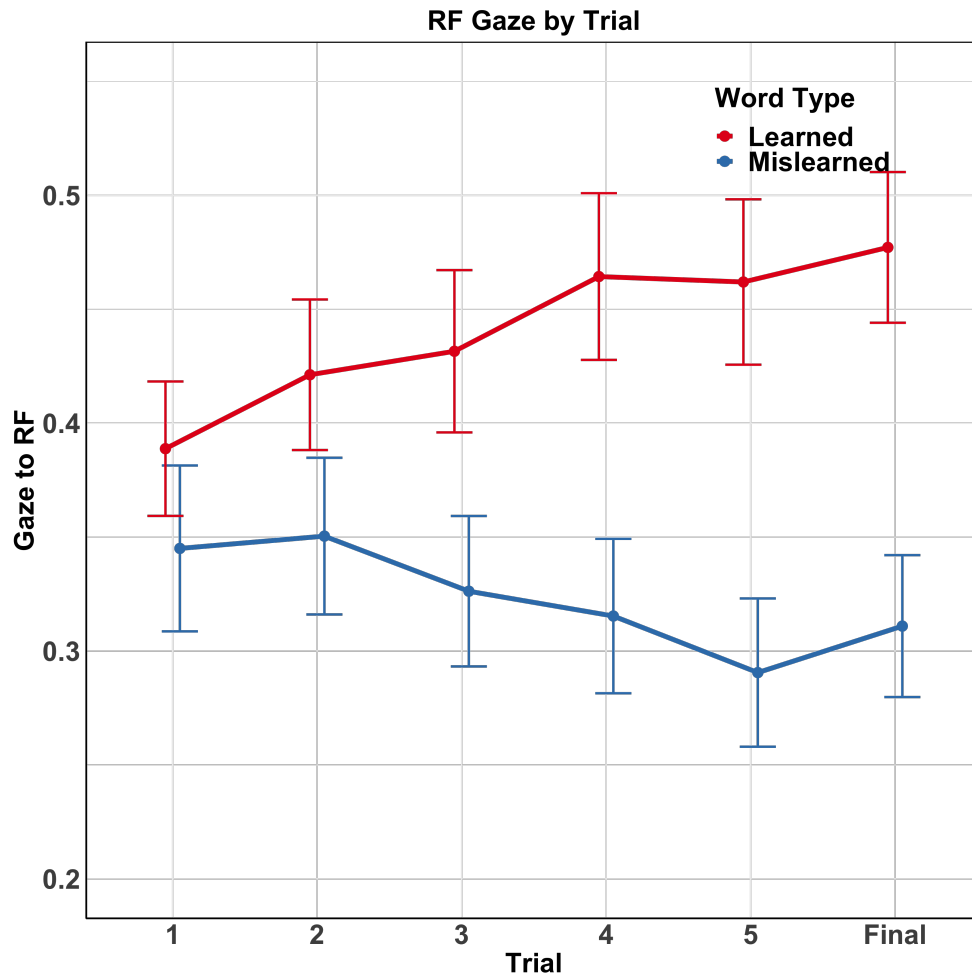


Figure 22: Gaze to RF-set during each training exposure. Participants, in aggregate, are likely to converge on their initial hypothesis. This figure includes trials from both parallel- and sequential-participants—however, for parallel-participants the first three “trials” were actually objects on the screen simultaneously.

### 4.3. General Discussion

The present set of results supports a hypothesis generation account of word learning over the statistical accumulation/evaluation view. In particular, when the exact same stimuli are presented in sequence rather than in parallel, learners’ generalizations are more broad. Even though sequentially encountered referents represent the same degree of statistical evidence, learners construct an initial hypothesis only once. After a reference has disappeared

then, due to the immediacy of linguistic computation, the learner can only refer back to it by consulting some mental representation for the presented word—which does not include the visual input that generated this hypothesis. Learners generate a hypothesis and either stick with it if evidence is consistent, or move to a new hypothesis (or otherwise incorporate updates) when faced with inconsistent evidence. The results also indicate that, consistent with this hypothesis generation view, learners’ attention *during* learning is limited only to evaluating the intermediate representation they have posited up to that point in time. This provides evidence against a framework under which learners evaluate stimuli holistically and perform an explicit optimization for the most probable meaning. Taken together, Chapters 3 and 4 highlight the utility of studying algorithm-level causal mechanisms operating within the learning process rather than high-level computational descriptions of the input and output. Effective models generate novel predictions which, when rigorously evaluated experimentally, provide much deeper insights than would be possible from either methodology alone.

## CHAPTER 5 : The Incremental Mechanisms of Functional Design: Language Production and the Immediacy of Computation

A model should be your friend, not your lover. That's  
because a friend can tell you when you're wrong, whereas  
a lover just tells you that you're wonderful

---

Mike Tanenhaus.

A major question in language production is what drives the order in which we say things.<sup>1</sup> While speakers generally produce utterances in a spontaneous and more or less fluent fashion, a rich cognitive architecture underlies the chain operating from semantic concepts to articulated utterances. When studying an opaque cognitive process such as language production, it is imperative to formally and concretely specify our level of analysis (Marr, 1982). Under Marrian terms we should make a distinction between a *computational*-level of analysis, which identifies what terms and materials are computed—i.e. a description of the system's Input/Output (I/O) goals—from an *algorithmic*-level of understanding which specifies the particular mechanism by which the high-level I/O description of the system is implemented (i.e. *how* data is computed rather than *what*).

With that in mind, this chapter has two main goals. The first is to build on the previous production literature (Bock and Levelt, 2002; Levelt, 1993; Ferreira and Dell, 2000; Levelt, 1992; Ferreira and Swets, 2002; Pechmann, 1989; Smedt, 1990) to lend support to a particular mechanism-level framework of language production which I term Incremental Generation (IG). Under IG, production functions through a number of forward-feeding incremental modules operating in semi-parallel. Since the input to downstream modules is generated by the output of upstream components, differences in the order in which information is delivered from one component to the next may manifest as eventual order in speech when such ordering changes do not drastically alter the meaning. I evaluate this framework using

---

<sup>1</sup>This chapter was born out of many lengthy and informative discussions with Tony Kroch, who contributed immensely to the development of the ideas contained here.

a set of statistical models over a very large database of naturally occurring data on the English verb-particle construction.

The second aim of this chapter is to explore how IG interfaces with proposals such as Uniform Information Density (Levy and Jaeger, 2007; Jaeger, 2010). I argue that, at a descriptive computational-level, IG and the present verb-particle data are consistent with the predictions of Uniform Information Density construed broadly, henceforth termed “UID as a computational-theory” or UIDC. Crucially, this pattern emerges as a result of the IG model (which makes a superset of predictions), rather than an explicit optimization in its own right. One might alternatively construct an algorithm-level understanding of UID, or UIDA. Under such a view, UIDC arises because individual speakers impose an ordering preference for information at a local-level. I present evidence that no algorithm-level optimization such as UIDA is taking place. UIDC is in a sense supported, but it does not involve the actual computations performed in the minds of individual speakers. In cases that something estimatable by predictability or surprisal plays a role in the study of language, this needs to be approached through the lens of explicit algorithms. By connecting explicit mechanisms to the computational study of language production, we can both uncover the limits and issues with information theoretic descriptions but also supply possibilities for improving them.

In the remainder of Section 5 I outline the Incremental Generation (IG) framework of language production. Section 5.2 introduces the English verb-particle construction and the methodology for data extraction, while Section 5.3 outlines the predictions that IG makes with regard to output order. In Section 5.4 I present the primary set of statistical analyses which evaluate, and offer support for, IG predictions. Section 5.5 describes the ways in which this relates to information theoretic approaches to language production (Jaeger, 2010), and in particular a view of Uniform Information Density both at a descriptive computational-level (UIDC) and an algorithmic-level (UIDA). While UIDC patterns are largely attested in the verb-particle data, there are particular conditions under which they disappear (Section 5.6); supporting the view that UIDC is an emergent property of IG rather than UIDA.

Section 5.7 concludes.

### 5.1. Language Production

That the planning and realization of speech occurs at a number of hierarchical levels is a widely supported view with a long history (Wundt, 1904; Lashley, 1951; Fromkin, 1971). More recently this has been popularly divided into a few major processes: conceptualization, grammatical encoding (also called “formulation” on some accounts (Roelofs, 1992, 1997)), and phonological encoding/articulation (Bock and Levelt, 2002). Conceptualization can be thought of as generating and mapping some intention onto an abstract “message.” This results in some conceptual information which needs to be encoded linguistically in order to accomplish the speaker’s communicative goal. The message contains the required lexical concepts and the relationship between them, but this may still be continually updated even while subsequent processes (grammatical encoding, etc.) have started operating.

The focus here, and the heart of the language production system, is grammatical encoding (Ferreira and Dell, 2000). Encoding starts with a message (or part of a message which is being continually developed) and outputs the word forms which make up the phonological content handled by the articulatory-phonetic system. Inside of grammatical encoding, a process retrieves the linguistic units which correspond to these lexical concepts and assigns them required thematic and functional structure (Levelt, 1992, 1993). Such lemmas are then linearized in accordance with syntactic restrictions. The end result is an articulatory plan for the utterance which can be realized as overt speech. This general framework is diagrammed in Figure 23.

What is important here is not only the presence of such distinct sub-systems to language generation, but the sequentially ordered relation between them (Ferreira and Swets, 2002). The most crucial organization is that language production functions *incrementally* (Pechmann, 1989; Roelofs, 1997, 2013; Ferreira and Swets, 2002; Smedt, 1990): only after a piece of information is made available at the immediately higher level of processing does it become available to trigger activity at the next level down the production stack. This means that a



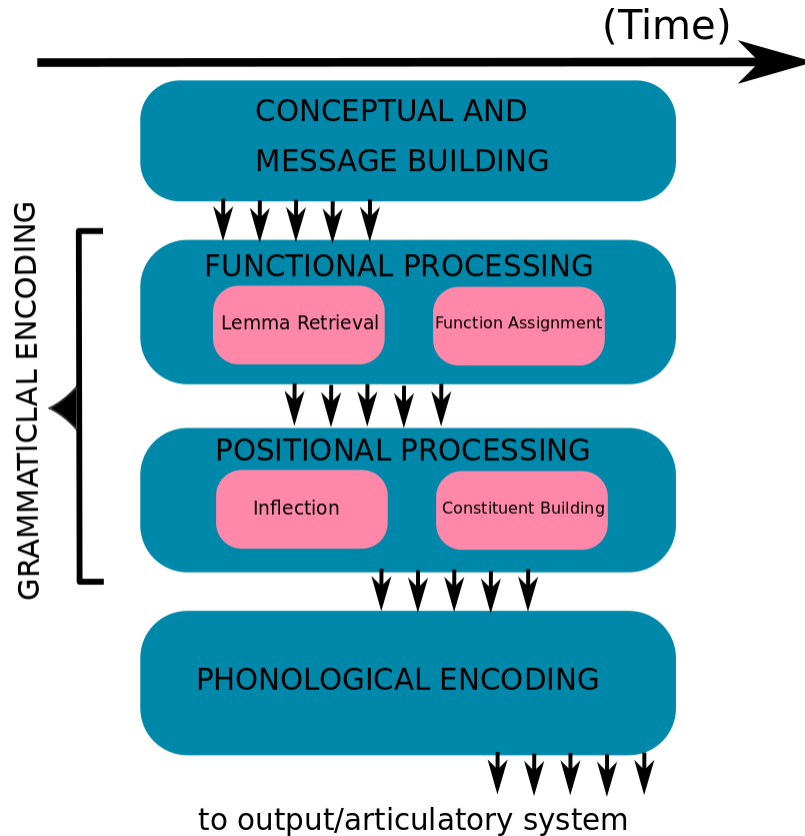


Figure 23: Basic outline of the language production architecture—adapted from Bock and Levelt (2002). Time is indicated from left-to-right.

word cannot be assigned phonological structure until it has received a semantic role. Phonetic articulation of an element cannot begin until it has been assigned morphological form, etc. While it may seem intuitive, this ‘vertical’ incrementality (Bock and Levelt, 2002) might be contrasted with a view in which a syntactic tree is first built, with individual words being slotted in afterwards (see Ferreira (1996) for support of incrementality of lexical retrieval over “competition” models). Additionally, processing in the system as a whole functions in parallel. Once a lemma has been sent off for morphological inflection, the system does not need to wait in order to start retrieving subsequent lemmas. *“At the same time that a piece of information works its way from idea to articulation, other pieces are constructed and make their way through the system as well.”* (Ferreira and Swets, 2002). Patterns in speech errors strongly indicate that functional and positional processing are able to function in parallel

(Dell, 1985; Hoppe-Graff et al., 1985)—yet despite the parallel nature of the system as a whole, individual elements (e.g. lemmas) are constrained sequentially and incrementally. Such assumptions about the incremental nature of language production have wide empirical support with evidence ranging from natural speech errors, to elicited productions in paired-association and picture naming tasks, etc. (Roelofs, 1998; Brown-Schmidt and Konopka, 2014; Roelofs, 2013, 1997; Ferreira and Swets, 2002; Smedt, 1994; Schriefers et al., 1990; Levelt et al., 1991, 1999), although see Stallings et al. (1998) for some complications.

Because language production is incremental, variations in the order in which information is delivered from one component to the next can affect the order in which elements appear in speech. Speakers begin articulating a sentence before the whole semantic content, let alone the syntactic structure, has been determined. Since the output of higher-level processing systems feeds the input of the lower-level system, this view of incremental production naturally predicts that higher-level modules do not need to complete their work on an utterance before the next level begins. A speaker may start the linearization, phonological encoding, and output of one constituent while simultaneously processing more content to be incorporated in the sentence. The system thus functions on a “first-in first-out” basis, a property I call “Minimized Buffering”<sup>2</sup> because without it, whatever lemmas are retrieved first would need to wait in a buffer for slower elements before continuing down the stack. There are explicit computational accounts of this principle (Smedt, 1990) as well as empirical support from visual-world naming tasks (Pechmann, 1989).

What I term the Incremental Generation (IG) framework—building on and consistent with Kempen and Hoenkamp (1987); Smedt (1990), among others—builds pieces of phrase structure as the lemmas become available and, owing to Minimized Buffering, fragments are fit together as quickly as possible so long as syntactic constraints are not violated. Taken together this makes a prediction about resultant constituent ordering: whatever factors correlate with initial conceptual retrieval or faster lexical access should correlate with that

---

<sup>2</sup>This is alternatively termed the “principle of immediate mention” (Ferreira and Dell, 2000, among others)

constituent being linearized first in the output. An elucidating example of this is given in Ferreira and Dell (2000):

“Assume a speaker wishes to describe the outcome of the race between the tortoise and the hare in Aesop’s fable with a verb such as defeat. Furthermore, assume that the lemma for the word hare is quickly activated and selected. Given the early selection of the hare lemma, the most efficient strategy is for the speaker to produce the passive, The hare was defeated by the tortoise, rather than the active, The tortoise defeated the hare, since only with the passive can the already-selected hare lemma be immediately mentioned. If the speaker produces the active, then one of two inefficient processing strategies must be adopted: Either the already-selected hare lemma must remain active in a buffered state until the sentence-final position arrives for production (while other words are selected and produced in earlier sentence positions), or the already-selected hare lemma must be deactivated and subsequently reactivated. From this, a general principle can be induced: Production proceeds more efficiently if syntactic structures are used that permit quickly selected lemmas to be mentioned as soon as possible.”

This type of mechanism is additionally supported by work on the link between visual-conceptual processing and linear order (Gleitman et al., 2007; Bungler et al., 2013), which manipulated not the lexical access speed for words, but the order in which concepts were activated through the visual system and found a similar effect on output production order. A summary of the Incremental Generation framework is given below:

1. Lexical retrieval functions incrementally
2. Different modules (functional vs. positional assignment) function in parallel
3. Due to Minimized buffering, lemmas are not held in a buffer longer than required by constraints of the grammar

4. Any factors which speed up lexical access are thus also proxies for linear order in the output

How are we able to evaluate the predictions which IG, or other frameworks, make on linear order? A difficulty in attempting to study the mechanisms of language production is that, most of the time, the largest contributor to spoken output is the meaning which speakers want to convey in particular contexts<sup>3</sup>. As such, any theory would attempt to make the same predictions in the majority of cases, i.e. output the grammatical and attested sentences while being unable to output the ungrammatical or otherwise impossible sentences. Thus a major testing ground for accounts of language production is the study of “syntactic optionality”: given multiple potential syntactic encodings for equivalent semantic sentences, what factors govern the use of one form rather than another. In Section 5.2 I describe the English verb-particle alternation (Gries, 2003; Lohse et al., 2004; Aarts, 1989) and the extraction of large-scale observational data to study optionality.

## 5.2. Verb-Particle Construction

The syntax of verb-particle constructions is well-researched (Farrell, 2005; Aarts, 1989; Gries, 2003; Thim, 2012, among others), although I will attempt to take a view which is as theory-neutral as possible. Consider the alternation in 1:

- (1) a. Julian picked up the book.  
b. Julian picked the book up.

This alternation includes a transitive verb, a morphologically invariant word which we’ll call a particle, and a direct object noun phrase. The sentence in 1a shows what I call “particle-first” order, while 1b exhibits “object-first” order. The verb-particle alternation in (1) should not be confused with superficially similar constructions such as in (2a). While the basic word order between (1) and (2a) appears similar, the underlying structure is quite dissimilar, as we can see by lack of the grammatical alternation as in (2b). The particles

---

<sup>3</sup>The ways in which syntactic alternations can differ in meaning are often quite subtle. See Caplan and Djärv (2019) for an example of syntactic alternations reflecting differences in meaning in Swedish.

in verb-particle constructions are not prepositions (i.e. they do not take objects), though words used as particles tend to be used as prepositions as well. Also, note that pronominal objects are unacceptable in particle-first order as in 3.

- (2) a. Patrick went into the lake.  
b. \*Patrick went the lake into.
- (3) a. Ryan put it down.  
b. \*Ryan put down it.

Previous work on ordering preferences of verb-particles have been framed as capturing disjoint accounts (e.g. NP-length (Hawkins, 1994) as opposed to discourse status (Dehé, 2002)). Gries (2003) invoked a multifactorial analysis of verb-particle data. However, the interpretive power of his analysis is limited by the fact that while multiple independent factors were in his statistical model, the explanation given to them required appeal to a variety of different accounts of processing cost. These range from performance theory (Hawkins, 1994) to apparent cost in identifying discourse referents (Givón, 1992), etc. Lohse et al. (2004) (and related experimental work (Wasow and Arnold, 2003)) successfully tied together a number of previously disparate factors in verb-particle order under the unified theory of “domain minimization” (Hawkins, 2004). This represents a computational-level theory, which IG as an algorithm should be consistent with, but an in-depth discussion of their connection is outside the scope of this Chapter.

### *5.2.1. Data Extraction*

I extracted verb-particle alternation instances from the Corpus of Contemporary American English (COCA) (Davies, 2009). This was an an ideal source as it is extremely large, allowing us to test multiple (interacting or related) predictions simultaneously. Additionally, it consists of five distinct genres (subcorpora): Academic texts, speech, news, magazines, and fiction. Unlike in the case of “that”-omission (as in 6), only a relatively minor fraction of total variance is explained by register differences across genres: The overall rate of particle-first order ranges only between 65% and 85%. This is far smaller than the amount of

sociolinguistically conditioned variance present in the case of “that”-omission, which ranges between 1% and 85% (Elsness (1984); Biber (1999)).

The code of automatically extracting verb-particle instances relies on possible “candidate verbs,” which were identified based on their co-occurrence with particles and pronominal objects. Because a sentence like “John picked him up” is grammatical but not \*“(John picked up him”)), thus if a verb *could* take part in a verb-particle construction, then we would expect to see it attested taking a pronoun object given a sufficiently large sample. This is straightforward to search for without getting false positives. I searched for all phrases matching the form “VERB PRONOUN PARTICLE.” I subsequently identified instances of such verbs containing neither pronouns nor prepositional phrases, e.g. “John put down the book”, but not \*“(John put down on the table the book”. Output data was tagged for a number of relevant factors including: order of particle, frequencies, conditional probabilities, constituent length, definiteness, source corpus, etc. These statistical measures were estimated over the entirety of COCA rather than only the verb-particle sentences which my regression model will eventually attempt to predict.

Data was limited to verbs which appeared as taking an object/particle combination (in either order) at least ten times and were not exclusively categorical in their particle-first or particle-second order. This was done to help remove idiomatic expressions which contain no optionality in their output order. The data was additionally filtered by running all extracted sentences through the Stanford lexicalized PCFG parser (Manning et al., 2014). I discarded any cases for which the parsed output did not contain exactly a phrasal verb followed by two daughter constituents tagged as a noun phrase (NP) and particle (PRT) in either order. After cleaning up, there are 67,905 unique sentences to be predicted. This includes 99 unique verbs, and 296 unique verb-particle pairs distributed over 33,085 unique triples (tuples of VERB, PARTICLE, OBJECT-HEAD). The magnitude of the extracted data is valuable since it allows for the evaluation of multiple hypotheses even over extremely rare events. These include 54,955 particle-first and 12,950 particle-second sentences for a total particle-first rate

of approximately 81%. This general rate of particle-first order is in line with previous work involving the manual extraction of verb-particle cases (Kroch and Small, 1978).

### 5.2.2. *Data Quality*

Any automated syntactic annotation or extraction scheme will necessarily contain some number of errors. No gold-standard corpus of verb-particle sentences exists against which the present extraction scheme can be evaluated. To evaluate the approximate quality of extracted data, I took a sample of 100 tagged instances and manually judged their status (true potential verb-particle alternation rather than a prepositional phrase for instance.) On this sample, approximately 90% of cases were true positive examples of the verb-particle alternation. The 10% of error cases are sentences which do not allow the alternate constituent order for syntactic reasons rather than facts of the processing mechanism. These were fairly evenly split between particle-first and object-first order. For example see below:

- (4) a. We pushed our way out.
- b. \*We pushed out our way.
- c. We pushed out a quality research project.
- (5) a. Why do I have to get out a period later than you?
- b. \*Why do I have to get a period later than you out?
- c. Let's get the birthday cake out.

Such cases are difficult to automatically avoid since the verb-particle pairs themselves can surface in the optional order some of the time. But these particular instances are only available on an alternate semantic interpretation which does not allow re-ordering.

The syntactic annotation scheme used for the Penn tree-bank was not designed with such fine-grained lexical semantic distinctions in mind. It is a necessary consequence that, based on these parse trees alone, some limited errors in construction identification arise. This would be true whether extraction were performed based on hand-annotated structures or algorithm-based. So long as the rate of errors is relatively low and, importantly, unbiased,

we can simply acknowledge it and proceed without it interfering in subsequent analysis. Nevertheless future work would benefit from confirming the robustness of the automatic extraction (and potentially uncover opportunities for improving it) by manually annotating a larger subset of the data for analysis.

The benefit of such naturally occurring and distributed data should also not be overlooked. There is widespread evidence that speakers and listeners hold representations and perform computations that are sensitive to aspects of statistical exposure (Hale, 2014; Levy, 2008; Bell et al., 2003). Even more importantly, adaptation to shifting distributions of use can be quite rapid (Saffran et al., 1999; Wells et al., 2009; Fine et al., 2013; Caplan et al., 2021), as with the adaptation to shifted speech in Chapter 2. Thus while large-scale observational data sacrifices experimental control and may introduce a certain degree of noise, it avoids a potentially dangerous confound from traditional experimental approaches to studying production: exposure to the experiment has the potential to (temporarily) interfere with the very representations that are being probed.

### 5.3. IG Predictions on the Verb-Particle Construction

In this section, I outline the major independent variables identified from previous studies (Gries, 2003; Lohse et al., 2004), and predicted by IG. This describes why each factor is predicted to correlate with output order and reports simple correlations. See Section 5.4 for the primary mixed-effects model results including the combined effects of these factors. When building statistical models like those under discussion, we need to keep in mind that a model which predicts linguistic output with high accuracy is not, in and of itself, an explanatory theory. It is insufficient to model the outcome of syntactic optionality (or any phenomenon) unless we can move towards understanding *why* correlated variables have predictive power. After all, statistical tools can only verify, but not produce, empirical hypotheses.

As it relates to IG and the verb-particle construction, the intermediate representation at the moment immediately following the output of the verb is an important inflection point.



The speaker still needs to output both the relevant object and particle, but at this stage, the grammar allows those two elements to be linearized in either order. This triggers a “race condition” between the two constituents. If the object is retrieved and constructed first, then it will be linearized first, and vice versa for the particle. Any factors which speed up lexical access or constituent assembly will also be proxies for output order. Figure 24 shows a schematization of this process. This intermediate representation is subject to the immediacy of linguistic computation: speakers cannot wait to consider what will be more efficient for the utterance as a whole, the race condition which minimizes buffering here is fundamentally a local computation.

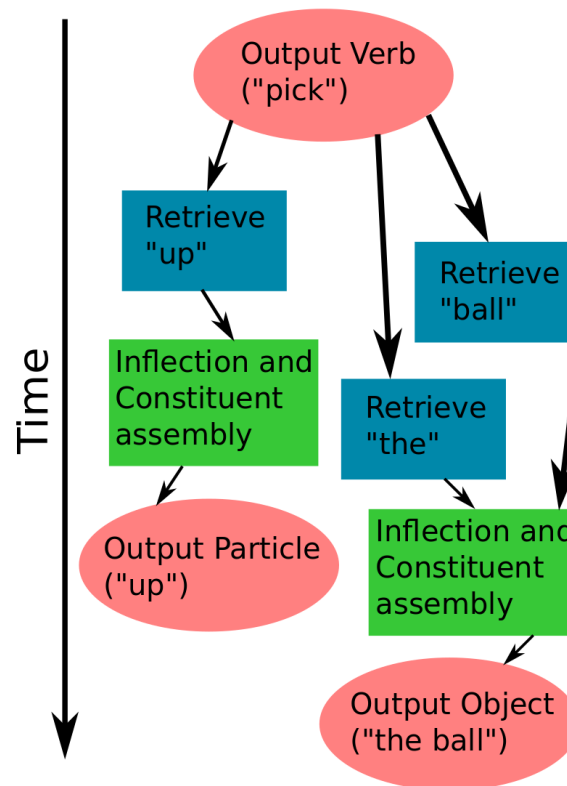


Figure 24: Example illustration of IG output of verb-particle construction. Variations in lemmas access or constituent assembly speed manifest as comparable variations in linear order (when permitted by the grammar): whichever element is retrieved and constructed first is sent off to positional processing first.

This is consistent with the reasoning behind “availability-based” accounts such as (Ferreira

and Dell, 2000; Bock, 1987), however I hope to expand on such accounts by evaluating a more complete set of cohesive factors than could be manipulated experimentally. Both unconditioned and conditional frequency play an important role here, and this is discussed further in relation to UID in Section 5.5.2. Frequency, predictability (computed several ways), definiteness, constituent length, and prior mention are discussed in turn below. As they are all independently known to affect lexical access speed or constituent assembly, IG predicts that they should correlate with linear order in the verb-particle construction. This is not an exhaustive enumeration of all potentially relevant correlates: speech rate, social register, disfluencies, stress patterns within the verb-phrase, etc. may all have some effect on output order. Technical limitations make it impossible to estimate and include all such factors in a model. More centrally, the goal of the chapter is not to predict verb-particle order with the highest possible accuracy — measures should only be taken if they are informative to testing clear predictions of the theories under consideration or differentiate between them.

### 5.3.1. Frequency

Higher frequency words have faster average retrieval times from memory (Bock and Levelt, 2002; Paap et al., 1987; Forster and Chambers, 1973; Inhoff and Rayner, 1986; Rayner and Duffy, 1986). Under IG, if the frequency of the object and the particle are pitted against one another — as frequency of an element goes up, so should the likelihood that it is linearized first. Since the object constituent commonly consists of multiple words, the frequency computation is made over the head noun. For instance, for a longer constituent like “the big green ball that was kicked into the yard” then the operative measure would be the total number of occurrences of “ball.”

In the data that follows I generally use the term “frequency” as a stand in for what is actually the inverse log transformation over frequency. This choice of term is purely for convenience, but the log transformation rather than raw count is widely known to better correlate with access times (Murray and Forster, 2004; Lignos, 2013). The *inverse* log frequency can intuitively be thought of as the additional processing cost/time resulting for lower underlying frequencies: as raw frequency goes down, then the inverse-log (cost/time)

goes up. Lower frequency objects correlate with particle-first order, and vice versa for particles.

$$Frequency(word) = -\log p(word) \quad (5.1)$$

### 5.3.2. Predictability

Rather than computing the baseline likelihood of a word or constituent overall (unconditioned frequency, as in Section 5.3.1), we might estimate the conditional frequency or predictability of a word in a particular local context. This captures important information because, for instance, the probability of “the” is high overall (it is the most frequent word in English), yet the likelihood that “the” occurs directly following “The coach said that the...” is essentially zero.

Under IG, predictability is predicted to correlate with linear order for the same reason that frequency does. “Predictability” here should be thought of as a form of priming, rather than the speaker actively trying to “predict” what they are about to say next. As the predictability of the object in context is lower, then lexical access times are correspondingly slower. The slower it is to retrieve and construct the object, the more likely it is for the particle to win the linearization race and be sent off for positional processing first (and vice versa). Like unconditioned frequency, predictability is a strong correlate of lexical access times (Staub, 2011; Rayner, 1998; Ehrlich and Rayner, 1981; Rayner et al., 2004) — an effect measurable in reading speed (Rayner, 1998), distribution of scanpaths (von der Malsburg et al., 2015), as well as several neurolinguistic measures (Frank et al., 2015; Willems et al., 2015; Henderson et al., 2016). While the connection between predictability and processing speed is a robust empirical effect, IG is not tied to any particular mechanism which enables this connection to emerge.

I take two measures of predictability at the *inflection point* after the verb has been produced:  $P(\text{object}|\text{verb})$  and  $P(\text{particle}|\text{verb})$ . Conditioning on the verb tabulates total occurrences

of either the object or the particle within up to five words to the right of the verb. As with the case of unconditioned frequency, and following Jaeger (2010), the “object” is actually a tabulation of the noun head rather than a multi-word phrase. As the predictability effect on access times is logarithmic rather than linear (Smith and Levy, 2013), I actually take the inverse log transformation over these probabilities. For consistency with the discussion of UID (Section 5.5), this is labeled as “Information.” Like in the case of frequency, the “Information” value can be thought of as the additional processing cost/time resulting lower in probability/higher surprisal, see Eq. 5.2.

$$\text{Information}(\text{word}|\text{context}) = -\log p(\text{word}|\text{context}) \quad (5.2)$$

Rather than compute conditional predictability solely following the verb, i.e.  $P(\text{object}|\text{verb})$  or  $P(\text{particle}|\text{verb})$ , I also compute the “downstream” effects of production after the inflection point. For instance, if the object were linearized first, then  $P(\text{particle}|\text{object})$  would immediately become relevant to processing, or vice versa for  $P(\text{object}|\text{particle})$ . These ‘downstream’ predictability values may be hugely asymmetric and hence have an effect on total constituent construction. For the same reasons as in Section 5.3.2 IG predicts that as  $\text{Information}(\text{particle}|\text{object})$  goes up, that should correlate with object-first order (due to slower processing) and vice versa.

### 5.3.3. *Definiteness*

Definite articles presuppose identifiability or familiarity within a context or discourse. This should correlate with faster lexical access times, and thus an IG framework predicts that definite objects should be more likely to occur in object-first constructions compared with indefinites.<sup>4</sup> Constructions containing definite articles occur with particle-first order 82.8%

---

<sup>4</sup>It is worth noting that the case of pronominal objects is, in spirit, similar to the effect of definiteness under an IG account. Since pronouns are highly frequent and, in context, refer to some typically salient individual their lexical access speed would in general be quite fast, consistent with object-first order. However, the apparent categoricity of the pronoun output order results from this being grammaticalized, which leaves underspecified the mechanism by which this grammaticalized happened historically and how pronoun order is stably acquired now.

of the time, compared with 77% for indefinite objects.

This is unsurprising given previous studies of definiteness and linear order (Ransom, 1977). Another view is that definiteness serves as a reasonable proxy for the given vs. new status of conceptual information within a discourse. In the English dative ordering alternation<sup>5</sup> Collins (1995); Bresnan et al. (2007) find an overwhelming effect of discourse status (given vs. new) on constituent ordering. While it is not possible to estimate the discourse status of constituents directly from a corpus with unknown speakers and contexts, definiteness is a reasonable, albeit limited, proxy for such discourse structure. IG offers a single mechanism of action by which these ordering effects on discourse structure emerge in language production in tandem with frequency, predictability, etc.

#### *5.3.4. Object Length*

The effect of phrase length on constituent ordering, so-called “heavy-NP shift”, is well-studied (Kimball, 1973; Stallings et al., 1998; Arnold et al., 2004, etc.). Under IG, this results because objects consisting of more words simply have more pieces to combine, and thus should take more time to build within the production system. IG predicts longer object length (which I measure here strictly by number of words) to correlate with particle-first order based on the slowdown in constituent construction time.

#### *5.3.5. Prior Mention*

Repeated mention of a constituent has been shown to facilitate faster lexical retrieval (Forster and Davis, 1984). Previous work on “that”-omission has seen only mixed effects on this regard; Ferreira and Dell (2000) found a strong effect within a sentence-recall paradigm, while that effect disappeared under more naturalistic dialogue settings (Ferreira and Hudson, 2011). Here I measure “prior mention” as a categorical outcome when the noun head of the verb-particle object had occurred previously in the same sentence.<sup>6</sup>

---

<sup>5</sup>The alternation illustrated in (1a) vs. (1b).

- (1) a. Santa gave [toys] [to the children]  
b. Santa gave [the children] [toys]

<sup>6</sup>In natural dialogue it would be more natural to repeat only the head rather than an entire object verbatim. Contrast the naturalness of (1a) compared to (1b).

If repeat mention does significantly speed up lexical access, then an IG framework predicts that it should correlate with object-first order. I find a notable correlation overall as the proportion of particle-first sentences is approximately 3% lower under object repeat-mention cases compared to the base condition (0.772 compared to 0.807). Likewise particle-first order is more likely in particle repeat-mention cases compared to the base condition (0.839 compared to 0.807). One possibility for this difference between the previous data on “that”-omission and the present verb-particle data is the rarity of overt repeat-mention (approximately 1% of sentences in the current data) which make any effect relatively difficult to uncover.

#### 5.4. Primary Model

I fit a mixed-effects logistic regression in order to evaluate the factors representing predictions of the IG framework of language production (listed in the second column of Table 9). The dependent variable was the binary outcome of linear order (particle-first ordering rather than object-first). The model included all aforementioned independent predictors as fixed-effects, along with random intercepts for each verb-particle pair and COCA genre. Output of the model is shown in Table 9

The results of this model on large-scale verb-particle (Table 9) provide strong evidence in favor of IG. Except in the case of “particle prior mention,” we see a strong significant correlation, in the expected direction, between every independent variable predicted by IG and the dependent linear order. In contrast with previous corpus-based studies of language production (Gries, 2003; Lohse et al., 2004; Jaeger, 2010, etc.), IG represents a single framework accounting for the surface variables under consideration. A unified mechanism of action offers an explanation of why this large number of correlates operate the way that they do.

- 
- (1) a. Find the biggest and most brightly colored apple in the display, and now hand me the apple.  
b. Find the biggest and most brightly colored apple in the display, and now hand me the biggest and most brightly colored apple.

<b>Factor</b>	<b>IG Prediction</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>p</b>
(Intercept)	N/A	12.49	2.18	5.74	< .001
Frequency(object)	Positive	0.37	0.01	25.54	< .001
Frequency(particle)	Negative	-1.53	0.24	-6.48	< .001
Information(object   verb)	Positive	0.12	0.01	12.50	< .001
Information(particle   verb)	Negative	-0.34	0.10	-3.52	< .001
Information(object   particle)	Negative	-0.38	0.01	-28.35	< .001
Information(particle   object)	Positive	0.18	0.01	19.61	< .001
Object prior mention	Negative	-0.35	0.10	-3.60	< .001
Particle prior mention	Positive	-0.02	0.19	-0.09	.93
Object Length	Positive	0.98	0.02	44.00	< .001
Definite Object	Negative	-0.83	0.03	-28.42	< .001

Table 9: Output of primary logistic regression model where the dependent variable was particle-first order.

### 5.5. Efficiency, Optimization and Uniform Information Density

The IG view of language production is egocentric and algorithmic: the output of the system looks the way it does on the primary basis of mechanical properties of the speaker’s cognitive architecture rather than displaying overt sensitivity to the potentially differing needs of a listener. On an alternative “audience-design” view of production, such communicative considerations are central. In order for linguistic communication to be successful, a hierarchical syntactic/semantic representation needs to be encoded, transmitted, and subsequently decoded. Since real-time language processing functions not over an entire sentence in batch, but over an incrementally received sequence of partial information over time, listeners are constantly faced with a large number of local ambiguities. While these are generally resolved either from prior context or additional content, some temporary syntactic ambiguities are biased toward an analysis that will eventually turn out to be wrong (i.e. “garden paths”) (Trueswell et al., 1993; Traxler and Pickering, 1996). Particularly when extra-syntactic properties (context, frequency, etc.) lead to a strong garden path, it may temporarily hinder communication (Ferreira and Henderson, 1991; Trueswell et al., 1999).

In contrast with IG, some theories of production posit that systems involve an important degree of computational oversight in order to behave in such a way that limits garden

paths and inefficient output (Clark and Tree, 2002; Temperley, 2003; Hankamer, 1973). Yet, however intuitively appealing such audience-design theories may seem, evidence of ambiguity avoidance through syntactic optionality has been lacking or mixed (Ferreira and Dell, 2000; Arnold et al., 2004; Wasow et al., 2005).

A related line of prominent reasoning to the role of communication in language and production is the the “Uniform Information Density” hypothesis (UID) (Jaeger, 2010; Levy and Jaeger, 2007). UID proposes that syntactic optionality is driven by a speaker’s implicit managing of computable information content to maximize communicative efficiency. This follows the same “audience-design” principle: if language has a primary function of communication, we might imagine each utterance to convey some particular “amount of information.” The hypothesis of Uniform Information Density is that, agnostic to the actual implementation of the language processing system, there must be an upper bound to the amount of information that can be processed within a fixed time. In order for communication to be efficient, a speaker should not want to convey too much information at once (which would be difficult to process) nor would they want to produce speech that is overly redundant (since that is a waste of potential bandwidth for information transfer). As languages allow some degree of optionality of expression for given semantic/pragmatic content, then in order for language processing to be efficient, the amount of information conveyed over time should be relatively more uniform than non-uniform. In this way there is an intuitive relationship between “information” and “predictability”: The more an event can be reasonably expected, the less we have learned upon its occurrence. Conversely, if an event is assumed very unlikely, then its occurring gives the listener a great deal of new information. Jaeger (2010); Levy and Jaeger (2007) take an inverse log transformation over conditional probability to serve as a representative proxy for information (Eq. 5.3). Note that this is not the typical notion of “information” in linguistics as “some meaningful propositional content” but rather a particular estimate of Shannon information (Shannon, 1948). In the context of UID, “information” is simply the inverse log transformation of predictability/probability of some word conditioned on a previous word. A prose definition of UID is provided by Jaeger



(2010) as follows: “*Within the bounds defined by grammar, speakers prefer utterances that distribute information uniformly across the signal (information density). Where speakers have a choice between several variants to encode their message, they prefer the variant with more uniform information density (ceteris paribus).*”

$$\text{Information}(word|context) = \log \frac{1}{p(word|context)} = -\log p(word|context) \quad (5.3)$$

Original support for UID came from optional “that”-omission (Jaeger, 2010). UID makes predictions in this case based on the fact that individual predicates vary in terms of how likely they are to introduce a complement clause. A verb like “thinks” is very likely to introduce a complement clause, and so hearing the overt complementizer “that” provides listeners with relatively little new information. Conversely “confirmed” is notably less likely to occur with complement clauses, and so hearing “we” directly follow it is a fairly surprising event (since the morphology indicates it must be an embedded subject).

- (6) a. The coach knew (that) the players were tired.  
 b. The teacher confirmed (that) we wouldn’t get our tests back.

Jaeger (2010) tested the UID hypothesis by extracting a sample of complement clause instances (approx. 7,000)—1,173 (17.5%) of which surfaced with an overt complementizer, while 5,543 (82.5%) did not—from the Penn Treebank (Marcus et al., 1993) subset of the Switchboard corpus of telephone dialogues (Godfrey et al., 1992). A multiple regression model was used to predict the binary outcome of overt “that,” and included a number of independent factors intended as proxies for different theories of language production. Jaeger (2010) found that “information” (defined as he does) is a significant predictor of “that”-omission on top of previously identified effects. This finding in support of UID (and the efficient language framework more broadly) has generated a large resultant literature with claims ranging from language production (Frank and Jaeger, 2008), and speech rate (Pelle-

grino et al., 2011), to word order typology (Maurits et al., 2010), distributional properties of the lexicon (Piantadosi et al., 2011), and language acquisition (Fedzechkina et al., 2012). With such apparent centrality to language, it should seem imperative to understand what is included in a theory like UID, what it does and does not predict, and what mechanism is responsible for such findings.

### *5.5.1. UID and Levels of Analysis*

In relation to Marr’s levels of analysis (Marr, 1982), there is an important distinction to be made between some aspects of how UID is described in prose (Jaeger, 2010) compared with the methods employed to evaluate it (Jaeger, 2010) and its assumed role at the individual (rather than community) level (Frank and Jaeger, 2008; Fedzechkina et al., 2012). The most general interpretation of UID is as a computational description or typology of language output, henceforth “UID as a computational-theory” or UIDC. Under UIDC, we simply posit that the distribution of Shannon information transmitted by each word/phrase is more uniform than it would otherwise be by chance. UIDC is not a causal theory, it represents a general statistical phenomenon — a high level description about optionality — but does not specify a process by which this pattern emerges. In other words, UIDC is not a theory involving the actual computations performed in the minds of individual speakers.

Every computational-level theory (including UIDC) necessarily requires some algorithm underlying it. While it is tempting to assume (perhaps implicitly) that some computational description was generated by the optimal, simplest, or most straightforward algorithm for generating such output, this is not a safe assumption. In many cases there may be a large set of naïve, local, or greedy algorithms which produce output that is very close to the optimal solution without reference to or explicit optimization over the function we describe at a computational-level (see Caplan et al. (2020) for an application of this distinction applied to the evolution of the lexicon). I argue that IG is the underlying causal algorithm responsible for giving rise to the patterns of UIDC.

While a log transformation over conditional probability is taken to be the proxy for in-

formation content under UID (Jaeger, 2010; Levy and Jaeger, 2007), the fact that such conditional probability (and contextual predictability more generally) should correlate with the output of grammatical optionality is not a unique prediction of UID. On Jaeger’s model, the major factor encoded to represent ease of lexical access in supposedly competing production algorithms is frequency rather than predictability (De Smedt and Kempen, 1991; Ferreira and Dell, 2000). I argue that this represents too limited a view of such theories and disregards Marr’s levels of analysis. Despite the claim from Jaeger (2010) that: “*Given the definition of information, UID assumes that speakers have access to probability distributions over linguistic units (segments, words, syntactic structures, etc.). This distinguishes UID from most existing production accounts, which make different architectural assumptions and do not predict information density to affect speakers’ preferences,*” there is no reason why a production algorithm, including IG, would be incompatible with probability-sensitive representations. In fact, predictability is a strong predictor of lexical access times (Staub, 2011; Rayner, 1998; Ehrlich and Rayner, 1981; Rayner et al., 2004) much like raw word frequency (Inhoff and Rayner, 1986; Rayner and Duffy, 1986). This effect of predictability (also commonly called “surprisal” in the parsing literature) is robustly attested across a number of different methodologies (See Hale (2014) for a review of such effects and effective models in sentence comprehension.) Scanpaths during reading are more irregular when predictability is low (von der Malsburg et al., 2015). Within studies of event related potentials, predictability is negatively correlated to the amplitude of the N400 component (Frank et al., 2015). The timecourse of activation in functional MRI is correlated with predictability values estimated from language models using both surface n-grams (Willems et al., 2015) and more rich phrase-structure (Henderson et al., 2016). Given that, IG (or an “availability-based” production theory more broadly (Ferreira and Dell, 2000)) also predicts conditional probability to correlate with output order through lexical access times. This is consistent with the results reported in Jaeger (2010) on “that”-omission: if uttering a complement-clause taking predicate introduces phrase structure either with or without an overt complementizer, then there is a similar output “race” between the complementizer and

the beginning of the complement clause.

### 5.5.2. UIDA

On my view, IG and our understanding of the factors influencing lexical access are sufficient to explain UIDC trends. An alternative account, UIDA (“uniform information density as an algorithm”), would posit that individual speakers impose a direct preference for uniform information ordering in a way that is disjoint from IG or other “production-side” theories. Here I discuss how UID(C/A) relates to factors previously introduced in Section 5.3.

The effect of (unconditioned) frequency is a unique prediction of IG which is not made by UID. Even though Shannon information could just as easily be computed using frequencies rather than conditional probabilities, our frequency values are global, unconditioned, and thus stable throughout the sentence. Whatever the gap in frequency is between the particle and the object within a verb-particle construction is, this difference remains constant regardless of the order with which the words are linearized. UID makes no particular prediction with respect to frequency and linear order. This is not to say that UID is *incompatible* with attested frequency effects, simply that UID is orthogonal to them—it neither predicts nor offers an explanation of frequency.

If we make the reasonable assumption that, on average, predictability of elements increases monotonically as a function of growing context, only then does UID predict a relation between predictability and linear order in verb-particle sentences. Under this assumption, and given whatever initial asymmetry existed between  $P(\text{particle}|\text{verb})$  compared with  $P(\text{object}|\text{verb})$ , then whichever is linearized first *increases* the baseline predictability of the second element. When the initially more-predictable element is output first, this reduces the asymmetry in information content at each time. Conversely, mentioning the low predictability element first would exacerbate the information asymmetry. For example, imagine that after producing the word “pick” then the predictability of “up” is 0.5 and the predictability of “book” is 0.1. If the next word uttered is in fact “up” that the subsequent predictability of “book” would rise higher, say to 0.6; while the verb imposes some selectional restrictions on

the object, it would seem natural that a verb-particle pair would further reduce the upcoming search space making predictability greater. Even if the selectional restrictions placed on the particle by the object are on average less drastic, it should be intuitive that predictability doesn't get *reduced* based on additional information. So while the gap in predictability directly following the verb was 0.4 (0.5 for "up" minus 0.1 for "book"), the eventual gap in predictability would be reduced if the more predictable element were linearized first ("up" is output at 0.5 predictability while the predictability of "book" rises to 0.2, hence the gap is reduced to 0.3) and it would be increased if the less predictable element were linearized first ("book" is output at 0.1 predictability while the predictability of "up" rises to 0.6, hence the gap is increased to 0.5).

In the cases of object-length and definiteness, UID makes no explicit prediction on linear order. Since probabilities (following Jaeger (2010)) are computed on the noun head, changing the article from definite to indefinite or adding a modifier does not effect that computation. UID is not incompatible with such attested effects, but UIDA would predict the effect of information density (predictability) to remain robust in the face of altering these orthogonal factors. Section 5.6 introduces a pair of additional experiments using object-length and model comparison to differentiate between IG and UIDA explanations of UIDC.

### 5.6. Object Length Experiment between IG and UIDA

An argument presented in Jaeger (2010) to support UID is that, even after controlling for correlates of other processing theories, the effect of UID manifests in his regression model. Jaeger additionally attempts to make a direct comparison of the relative theoretical status between such factors by examining the size of coefficients for such actors:

“To put the effect in relation to two theories of sentence production that have received considerable attention in psycholinguistic work on syntactic variation, availability-based sentence production and dependency processing accounts: the effect associated with the only parameter fitted for information density outranks the effect of all three parameters associated with dependency length effects in the

model. The effect of information density also is much larger than the combined effect of accessibility related parameters in the model. That is, information density emerges as the single most important predictor of complementizer that-mentioning.”

However, comparing the coefficient sizes between variables is not a good way to judge their relative “status” within the processing architecture. Because more sentences happen to contain short objects (Figure 25), it follows naturally that the effect of object length appears smaller. This alone does not serve as evidence to support UIDA or afford particularly elevated status to UIDC. By looking at the subset of sentences with somewhat longer objects, we give a fairer chance to compare the two theories. Under UIDA, if speakers directly manage information density as posited, then the effects of predictability on linear order should remain present when examining multi-word objects. This is not to say we shouldn’t expect the coefficient size to change (based on an interaction with other factors), simply that a meaningful interpretation of UIDA predicts information density to remain a significant factor in the regression model.

Alternatively under IG, predictability is only one of many factors which correlate with linear order by way of the single mechanism of lexical access times. On this view, there is no particularly special status for conditional probability compared to frequency, definiteness, or object length. Each factor has some effect on lexical retrieval speed (potentially in opposing directions). On this equal footing, we might expect the predictive power of information density to disappear when looking at medium-length objects—any boost in retrieval speed that increased conditional predictability offers would be overshadowed by the increased amount of time it takes to build and process a multi-word object. To evaluate these differing predictions of IG and UIDA, I re-ran similar regression analyses as in Section 5.4 but limited the evaluation set to sentences whose objects are at least N words long for various values of N. Compare the output of evaluating cases of N=2 or more words in Table 10 (58,628 instances at a ratio of 82.05% particle-first) with what happens when limited to somewhat

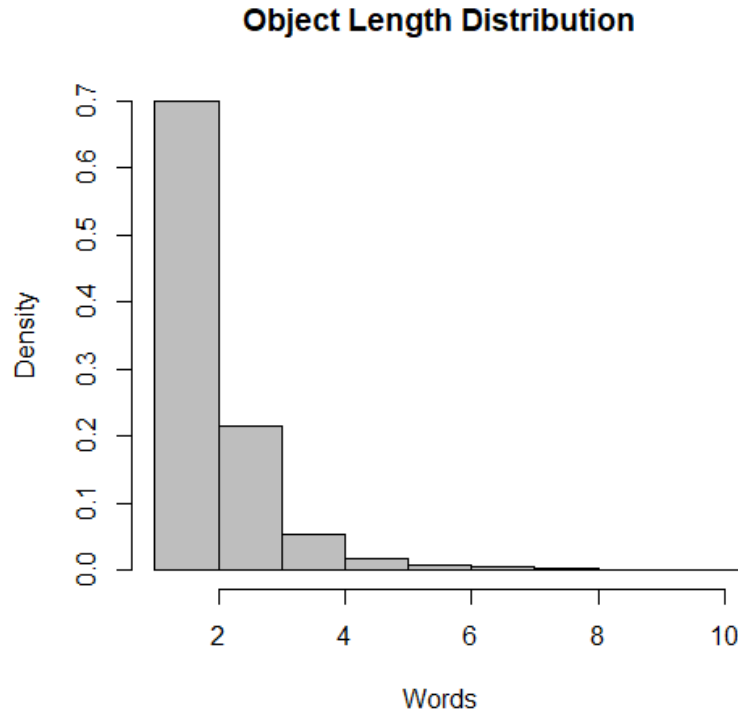


Figure 25: Distribution of object length (in words) within the present sample of verb-particle data (67,905 sentences)

longer objects as in Table 11 (N=4 or more words, 5,688 instances at a ratio of 96.6% particle first)

When restricted to evaluating even medium-length objects (four or more words), there is no effect of information density. Importantly, while the proportion of particle-first cases is approaching categorical here (driven by the length of the object), other processing factors remain significant predictors of linear order. The lack of significance for the “information” factor is not simply an effect masked by distance, since the frequency factor (like Information) is computed on the same object head at an equal distance from the verb. This represents evidence against UIDA and in favor of IG. Information theory is no silver bullet for language processing. When effects describable via information theoretic formalisms emerge as a result of algorithms with many moving parts, it should be expected that such effects interact with

Factor	Coefficient	Std. Error	z	p
<b>(Intercept)</b>	-1.38	0.23	-5.94	< .001
<b>Frequency(object)</b>	0.05	0.01	5.48	< .001
<b>Information(object   verb)</b>	0.06	0.01	7.34	< .001
<b>Object prior mention</b>	-0.3	0.10	-3.04	.002
<b>NP Length</b>	1.05	0.03	37.82	< .001
<b>Definite Object</b>	-0.72	0.03	-27.03	< .001

Table 10: Evaluating cases of N=2 or more words

Factor	Coefficient	Std. Error	z	p
<b>(Intercept)</b>	0.83	0.74	1.11	.27
<b>Frequency(object)</b>	0.16	0.05	2.92	.004
<b>Information(object   verb)</b>	-0.08	0.06	-1.42	.16
<b>Object prior mention</b>	0.16	0.57	0.27	.78
<b>NP Length</b>	0.59	0.13	4.55	< .001
<b>Definite Object</b>	-1.05	0.17	-6.11	< .001

Table 11: Evaluating cases of N=4 or more words. The effect of conditional probability is absent, while the effects of frequency, object length, and definiteness remain.

other factors, and may disappear under the right (or perhaps wrong) circumstances.

## 5.7. General Discussion

In this chapter I have attempted to build on that production literature (Bock and Levelt, 2002; Levelt, 1993; Ferreira and Dell, 2000; Levelt, 1992; Ferreira and Swets, 2002; Pechmann, 1989; Smedt, 1990) by outlining Incremental Generation (IG) as an explanatory mechanism underlying production choices around optionality. IG functions by a set of incremental forward-feeding modules linking from conceptual representation, to functional processing (lemma retrieval, etc., to positional processing (inflection and constituent assembly), and finally to phonological and articulatory encoding. Since (A) we have an incremental module responsible for lexical retrieval in this stack ahead of positional assignment, (B) operation between different modules functions in parallel, and (C) input to downstream modules is generated by the output of upstream components, then this has the measurable effect that differences in the order of conceptual access (Gleitman et al., 2007), lexical retrieval or constituent-building output manifest as comparable differences in output order.



It is because the architecture adheres to minimized buffering that whatever factors speed up upstream access, etc. will also be predictors of linear order among constituents. IG is strongly supported by the verb-particle data presented here, and is consistent with previous data on “that”-omission (Roland et al., 2006; Jaeger, 2010). Beyond that, IG offers a single, unified mechanism of action to explain the correlation between linear order and frequency, predictability, definiteness, and object length.

While a great deal of attention has been paid to information theoretic approaches to language and production in the last decade (Jaeger, 2010; Piantadosi et al., 2011; Frank and Jaeger, 2008; Pellegrino et al., 2011; Maurits et al., 2010; Fedzechkina et al., 2012), it is important to recognize the application of different levels of analysis (Marr, 1982) to the study of cognition. Uniform Information Density (and more specifically UIDC) is a descriptive tendency of data relating to optionality and should not be confused with a cause for it. IG provides an explanation for why UIDC emerges. A stronger information theoretic account under which speakers directly impose an ordering preference for information at a local-level, UIDA, is not supported. While UIDC systems may have properties beneficial to communication, this results without an explicit optimization for it in the minds of individual speakers. To whatever degree we can characterize the output of the language production system as efficient in information ordering, this is an emergent property of a simple, incremental generation system. In other words, we can “win” without trying.

A significant portion of discussion in this Chapter focused on the interface between IG and UIDC, but this is not the only set of information-theoretic findings which IG offers a mechanical explanation of. Mahowald et al. (2013) provides experimental evidence that, in a production task, speakers choose to use the shorter variant of a semantically equivalent pair (e.g. “chimp” as opposed to “chimpanzee”) in more predictable contexts. This is presented as evidence that speakers are (implicitly) manipulating information rates to adhere to constraints of communicative efficiency. However, such a finding would seem to fall out directly from the view of IG presented in this Chapter. Once a concept has been activated within

the production system, the lexical retrieval module attempts to retrieve the corresponding lemma. In this case there is a race between forms compatible with the concept (in this case “chimp/chimpanzee” or “math/mathematics”) which is analogous to the race between the object and the particle in the verb-particle construction — whichever lemma is retrieved more quickly is sent off for positional processing. Since predictability supports faster lexical retrieval (along with frequency, etc.), it follows that speakers are more likely to pronounce the shorter (and more frequent) element in the pair in predictable contexts compared to unpredictable ones.

I hope that future work continues along the lines laid out here. It is important that work in cognitive science (of language and otherwise) stay true to a tradition in which we study the computational systems responsible for individual human behavior. The mathematical tools of information theory, probability theory, etc. have had huge utility in our development of psycholinguistic models, but the identification of statistical phenomena is not an explanation of them. We cannot achieve a thorough account and understanding of language production and processing without reference to underlying causal mechanisms.

## CHAPTER 6 : Conclusions

The goal of this dissertation is to explore the wide ranging implications of a simple fact: language unfolds over time. Whether as cognitive symbols in our minds, or as their physical realization in the world, if linguistic computations are not made over transient and shifting information as it occurs, they cannot be made at all. This idea, the *immediacy of linguistic computation*, motivates the study of the *intermediate representations* that are constructed during online processing and acquisition. While ultimately extracted from linguistic input, such intermediate representations may differ significantly from the underlying distributional signal. Language is fundamentally constrained by how and when learners generate linguistic hypotheses, subject to the immediacy of computation, above and beyond whatever description appears to be the best statistical fit to the input data.

Chapter 2 applied these ideas to the domain of speech processing. Using a novel perceptual learning paradigm I observed that listeners can rapidly adapt to variation in phonetic input, updating their mapping between perceptual cues (e.g. VOT) and phonological categories. This adaptation is only able to occur however, when the relevant disambiguating information is provided *before* the ephemeral speech signal rather than *after*. This is consistent with a theory by which the intermediate representation of speech consists of graded activation over discrete linguistic categories (e.g., phones, words) but does not include the acoustic-phonetic evidence which gave rise to that activation. Crucially, this is a *Markovian* process: listeners encode a state of activation but do not retain the precise sensory evidence which led to that belief. It is worth noting that even retaining extremely course-grained information about the relevant acoustic cues would have been sufficient for adaptation (i.e. tracking “high” and “low” VOT), yet this did not occur. The structure of intermediate representations thus has substantial ramifications as a bottleneck for broader theories of phonological representation. For instance, the present findings raise important questions for exemplar accounts of phonology (e.g. Bybee, 2002; Pierrehumbert, 2001; Johnson, 2006): it does not seem possible that acoustic-phonetic detail be stored in stable representations when it is not active in an

intermediate state over the span of even several seconds.

The findings in Chapter 2 are mirrored in the restrictions that the immediacy of linguistic computation places on the representation of word meanings constructed during learning. In Chapter 3 I introduced a model of word learning (NGM) grounded in category formation. The NGM outlines a mechanism by which hearing novel words invites a learner to create a new category from component “features.” Once a hypothesis about the word meaning has been generated, the learner is able to evaluate subsequent labeled objects with respect to this hypothesized meaning. However, as in the intermediate representation of speech, the learner no longer has access to the underlying distribution which generated that belief. This by-product of the immediacy of linguistic computation means that learners are sensitive properties of the input timing (whether referents are displayed in parallel or in sequence) that are orthogonal to the raw statistics. Important evidence for our understanding of psycholinguistic systems often comes from particular conditions under which human behavior is “non-optimal” or diverges from the input distribution. Simon (1996) lays this point out nicely: *“A bridge, under its usual conditions of service, behaves simply as a relatively smooth level surface on which vehicles can move. Only when it has been overloaded do we learn the physical properties of the materials from which it is built.”* Chapter 4 evaluates further predictions of the NGM through a new eye-tracking paradigm. Using eye-gaze as a measure of selective attention to component features, we can shed light on the content and time-course of intermediate representations as they emerge during word learning. I find that, consistent with the general approach, learners’ attention during learning is limited only to the features present in their current hypothesis.

Finally, Chapter 5 explores the effect of the immediacy of computation on language production and “syntactic optionality.” Under an influential existing account (Jaeger, 2010), speakers’ choices are governed by a preference to distribute information evenly and efficiently over time. While this notion of efficiency is well-defined, it does not specify a mechanism that generates linguistic output. I describe a framework of language production in which

behavior is rapid and incremental: the system outputs lexical items as soon as they are retrieved (within the bounds of the grammar). I evaluate the predictions of this “incremental generation” account in comparison to other theories by fitting a statistical model over large-scale corpus data on the English verb-particle alternation. The output of the production system is best understood, in this explicit model, as the by-product of psycholinguistic factors: whether the object or the particle is linearized first depends on the intermediate representation following output of the verb. While in many instances, “incremental generation” and Uniform Information Density make convergent predictions, by focusing on the underlying production mechanism, I identify specific cases where these two theories differ and “incremental generation” is uniquely supported. Thus functional global behavior is the by-product of a local process: the immediacy of computation reduces demands on the language producer by placing a grammatical structure in the output as quickly as possible.

Taken together these case studies represent a rich analysis of the immediacy of linguistic computation and its system-wide impact on the mental representations and cognitive algorithms of language. I take seriously the view that to make fundamental progress we need to move towards process-level causal explanations in addition to high-level descriptions of computational phenomena. That the design of language is, in part, governed by principles of computational efficiency is a broadly accepted view, but often underspecified. I believe that the immediacy of linguistic computation, as a reflex of inherent cognitive constraints, represents an explicit mechanistic account of Chomsky’s “third-factor” which, along with UG and experience, conspires to shape language. My hope for this work is that it will serve as an initial guide and a foundation for me and others interested in tackling questions in {computational,psycho}-linguistics.

## APPENDIX A

### A.1. Stimulus Lists

#### A.1.1. Exposure Target Items

<hr/>	<hr/>
/t/-item	/d/-item
<hr/>	<hr/>
tab	dab
tally	dally
tangle	dangle
tear	dare
tech	deck
teem	deem
tense	dense
tie	dye
tier	deer
time	dime
tip	dip
toll	dole
tomb	doom
tongue	dung
tour	door
tow	dough
town	down
tub	dub
tummy	dummy
tune	dune
tusk	dusk
two	do

Table 12: Target stimulus pairs used in Experiments 1, 2, and S1.

A.1.2. Exposure Filler Items

Filler-item				
acre	airline	amuse	angry	annual
assembly	average	awareness	beach	begin
brush	business	capable	carve	cheese
chemical	clinic	color	companion	conscious
cruel	curiously	emphasis	employ	energy
ever	exam	experience	feeling	fellow
female	fierce	film	firm	fish
five	flesh	fool	four	frame
freely	from	funeral	gain	governor
happily	hire	holy	impression	improve
impulse	jealous	jump	lung	magazine
march	marsh	measure	missing	mouse
muscle	myself	navy	nearly	nervous
none	normally	offence	pack	parallel
pencil	permission	pile	plunge	polish
pour	pursue	religious	rope	rough
roughly	safely	scale	shall	sheep
similar	slip	small	smoke	soak
socially	soil	somewhere	sure	suspicious
sweep	vaguely	vessel		

Table 13: Filler stimuli used in Experiments 1, 2, and S1.

A.2. Full Null Model Structures for Mixed Effects Regression Analyses

Model structures shown below are the constrained (null) models against which the more complex models with the effects of interest were tested. Descriptions of the fixed and random effects in each null model are given, followed by the syntax used in R (*lmer* package).

For models in Chapter 2 that were run on *all* data (first and last Test Half included together), the following was always the null model structure used:

- Main effects of Test Half and VOT (centered)
- For `subject.id`, a random intercept
- For `testExemplar.id`, a random intercept

```
-----
t_choice ~ 1 + testHalf + VOT_centered +
          (1 | subject.id) + (1 | testExemplar.id)
```

For all models in Chapter 2 that were run on *separate test halves* (first and last Test Half separately), the following was always the null model structure used, except for those indicated further below:

- Main effect of VOT (centered)
- For `subject.id`, a random intercept and an uncorrelated random slope of VOT (centered)
- For `testExemplar.id`, a random intercept

```
-----
t_choice ~ 1 + VOT_centered + (1 | subject.id) +
          (0 + VOT_centered | subject.id) + (1 | testExemplar.id)
```

For analysis in Experiment 1 of the effects of interest (Shifted Phone and Timing) in First Test Half only (*without* the ceiling/floor cutoff, i.e. the analysis reported in the main manuscript), the following was the null model structure used:

- Main effect of VOT (centered)
- For `subject.id`, a random intercept
- For `testExemplar.id`, a random intercept

```
-----
t_choice ~ 1 + VOT_centered + (1 | subject.id) + (1 | testExemplar.id)
```

For analysis in Experiment 2 of Timing condition on Shifted-/d/ data only, First Test Half



only (*without* the ceiling/floor cutoff, i.e. the analysis reported in the main manuscript), the following was the null model structure used:

```

- Main effect of VOT (centered)
- For subject.id, a random intercept
- For testExemplar.id, a random intercept
-----
t_choice ~ 1 + VOT_centered + (1 | subject.id) + (1 | testExemplar.id)

```

For analysis in Experiment S1 of the effects of interest (Shifted Phone and Timing) in Last Test Half only (*with* the ceiling/floor exclusion criterion that was pre-registered but ultimately not used in the main manuscript), the following was the null model structure used:

```

- Main effect of VOT (centered)
- For subject.id, a random intercept
- For testExemplar.id, a random intercept
-----
t_choice ~ 1 + VOT_centered + (1 | subject.id) + (1 | testExemplar.id)

```

### A.3. Full Regression Outputs for Best Fitting Models

All brackets in this section represent 95% confidence intervals.

#### A.3.1. Experiment 1

Predictor	Coefficient	z	p	Odds Ratio
(Intercept)	1.55 [0.31, 2.79]	2.45	.014	4.72 [1.36, 16.32]
VOT	3.08 [2.99, 3.17]	67.89	< .001	21.81 [19.95, 23.84]
Shifted Phone	-0.15 [-0.33, 0.02]	-1.73	.083	0.86 [0.72, 1.02]
Timing	-0.19 [-0.37, -0.02]	-2.15	.031	0.83 [0.69, 0.98]
Test Half	-0.12 [-0.17, -0.08]	-5.27	< .001	0.89 [0.85, 0.93]
Shifted Phone x Timing	-0.08 [-0.26, 0.09]	-0.95	.343	0.92 [0.77, 1.09]
Shifted Phone x Test Half	0.05 [0.01, 0.1]	2.27	.023	1.05 [1.01, 1.1]
Timing x Test Half	0.03 [-0.02, 0.07]	1.21	.227	1.03 [0.98, 1.08]
Phone x Timing x Test Half	0.05 [0.01, 0.1]	2.38	.017	1.06 [1.01, 1.11]

Table 14: Output of the best fitting model on all trials for Experiment 1

#### A.3.2. Experiment 2

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.84 [0.45, 3.23]	2.60	.009	6.29 [1.57, 25.17]
VOT	3.47 [3.31, 3.63]	43.13	< .001	32.2 [27.5, 37.7]
Shifted Phone	-0.12 [-0.34, 0.1]	-1.08	.282	0.89 [0.71, 1.1]
Timing	-0.23 [-0.45, -0.01]	-2.08	.038	0.79 [0.64, 0.99]
VOT x Shifted Phone	0.24 [0.1, 0.37]	3.41	< .001	1.27 [1.11, 1.45]
VOT x Timing	0.02 [-0.12, 0.15]	0.24	.811	1.02 [0.89, 1.17]
Shifted Phone x Timing	-0.26 [-0.48, -0.04]	-2.33	.02	0.77 [0.62, 0.96]
VOT x Shifted Phone x Timing	-0.23 [-0.37, -0.1]	-3.33	< .001	0.79 [0.69, 0.91]

Table 15: Output of the best fitting model on the first half of test trials for Experiment 1

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.9 [0.51, 3.29]	2.68	.007	6.69 [1.67, 26.8]
VOT	4.27 [3.89, 4.66]	21.78	< .001	71.78 [48.86, 105.45]
Shifted Phone	-0.1 [-0.36, 0.15]	-0.79	.428	0.9 [0.7, 1.17]
Timing	-0.11 [-0.37, 0.15]	-0.84	.403	0.9 [0.69, 1.16]
VOT x Shifted Phone	0.25 [-0.12, 0.61]	1.32	.186	1.28 [0.89, 1.84]
VOT x Timing	0.38 [0.02, 0.74]	2.04	.041	1.46 [1.02, 2.09]

Table 16: Output of the best fitting model on the last half of test trials for Experiment 1

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.98 [0.29, 3.68]	2.29	.022	7.26 [1.33, 39.59]
VOT	4.5 [3.99, 5.02]	17.17	< .001	90.26 [53.99, 150.92]
Shifted Phone	-0.53 [-0.92, -0.14]	-2.68	.007	0.59 [0.4, 0.87]

Table 17: Output of the best fitting model on the first half of test trials, text-before condition for Experiment 1

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	2.32 [0.88, 3.76]	3.15	.002	10.17 [2.4, 43.09]
VOT	4.02 [3.59, 4.46]	18.05	< .001	55.96 [36.14, 86.63]

Table 18: Output of the best fitting model on the first half of test trials, text-after condition for Experiment 1

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.28 [-0.35, 2.91]	1.54	.123	3.6 [0.71, 18.31]
VOT	3.44 [3.35, 3.53]	76.58	< .001	31.11 [28.49, 33.97]
Shifted Phone	-0.09 [-0.28, 0.1]	-0.95	.343	0.91 [0.75, 1.1]
Timing	-0.16 [-0.35, 0.03]	-1.62	.106	0.85 [0.7, 1.03]
Test Half	-0.18 [-0.22, -0.14]	-8.80	< .001	0.83 [0.8, 0.87]
Shifted Phone x Timing	-0.24 [-0.43, -0.05]	-2.48	.013	0.78 [0.65, 0.95]

Table 19: Output of the best fitting model on all trials for Experiment 2

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.5 [-0.58, 3.57]	1.41	.158	4.46 [0.56, 35.63]
VOT	4.68 [4.36, 4.99]	28.96	< .001	107.49 [78.32, 147.52]

Table 20: Output of the best fitting model on the last half of test trials for Experiment 2

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.58 [-0.17, 3.33]	1.77	.078	4.85 [0.84, 28.06]
VOT	4.16 [3.78, 4.54]	21.69	< .001	64.06 [43.99, 93.29]
Shifted Phone	-0.41 [-0.72, -0.09]	-2.54	.011	0.66 [0.49, 0.91]

Table 21: Output of the best fitting model on the first half of test trials, text-before condition for Experiment 2

<b>Predictor</b>	<b>Coefficient</b>	<b>z</b>	<b>p</b>	<b>Odds Ratio</b>
(Intercept)	1.77 [0.3, 3.24]	2.37	.018	5.88 [1.36, 25.48]
VOT	3.85 [3.5, 4.21]	21.25	< .001	47.18 [33.06, 67.32]

Table 22: Output of the best fitting model on the first half of test trials, text-after condition for Experiment 2

#### A.4. Bayes Factor Calculation

Bayes Factors were computed in R using the `brms` package (Bürkner, 2017) and the following parameters:

- Default priors assigned by BRMS
- Iterations: 2000 (Default)
- Chains: 4 (Default)
- Delta: 0.999 (increased to ensure accurate estimation of posteriors)
- Maximum Tree Depth: 15 (increased to ensure accurate estimation of posteriors)

`brms` calls for the analysis of the test-phase (first half of trials only) comparing a model with and without a main effect of shifted-phone (“ambigPhoneme”):

```

full_brms_bernoulli = brm(t_choice ~ 1 + VOT_centered + ambigPhoneme +
  (0 + VOT_centered | subject.id) +
  (1 | subject.id) + (1 | testExemplar.id),

  data = subset(data.test.cur, experimentName == currExp &
    blockOrderHalf == curHalf &
    contextCondition == contextConditionSetting),
  control = list(adapt_delta = 0.999, max_treedepth = 15),
  family = bernoulli, save_all_pars = TRUE,
  iter = 2000, cores = getOption("mc.cores", 4L))

null_brms_bernoulli = brm(t_choice ~ 1 + VOT_centered +
  (0 + VOT_centered | subject.id) +
  (1 | subject.id) + (1 | testExemplar.id),

  data = subset(data.test.cur, experimentName == currExp &
    blockOrderHalf == curHalf &
    contextCondition == contextConditionSetting),
  control = list(adapt_delta = 0.999, max_treedepth = 15),
  family = bernoulli, save_all_pars = TRUE,
  iter = 2000, cores = getOption("mc.cores", 4L))

BF_brms_bridge = bayes_factor(full_brms_bernoulli, null_brms_bernoulli)

```

brms calls for the analysis of exposure phase match/mismatch accuracy on ambiguous trials. Comparing a model with and without a main effect of Timing (“contextCondition”):

```
full_brms_exposure = brm(accuracy ~ 1 + contextCondition +
  (1 | subject.id) + (1 | testExemplar.id),
  data = data.train.ambigOnly,
  control = list(adapt_delta = 0.999, max_treedepth = 15),
  family = bernoulli, save_all_pars = TRUE,
  iter = 2000, cores = getOption("mc.cores", 4L))

null_brms_exposure = brm(accuracy ~ 1 + (1 | subject.id) +
  (1 | testExemplar.id),
  data = data.train.ambigOnly,
  control = list(adapt_delta = 0.999, max_treedepth = 15),
  family = bernoulli, save_all_pars = TRUE,
  iter = 2000, cores = getOption("mc.cores", 4L))

BF_brms_bridge_Exp2 = bayes_factor(full_brms_exposure, null_brms_exposure)
```

## A.5. Secondary Analyses

### A.5.1. Checking for Bimodality in Participant Responses

To check whether some participants showed the expected adaptation effect in the text-after Timing conditions and others did not, I tested for bimodality in psychometric thresholds. In particular, I fit psychometric functions for each participant using the R package *quickpsy* (Linares and López i Moliner, 2016) —first half of test data only— and extracted the 50% categorization thresholds. I excluded any subjects with poor fits (thresholds  $< 0$  ms or  $> 100$  ms; one participant in Experiment 1, one in Experiment 2, and four participants in Experiment S1).

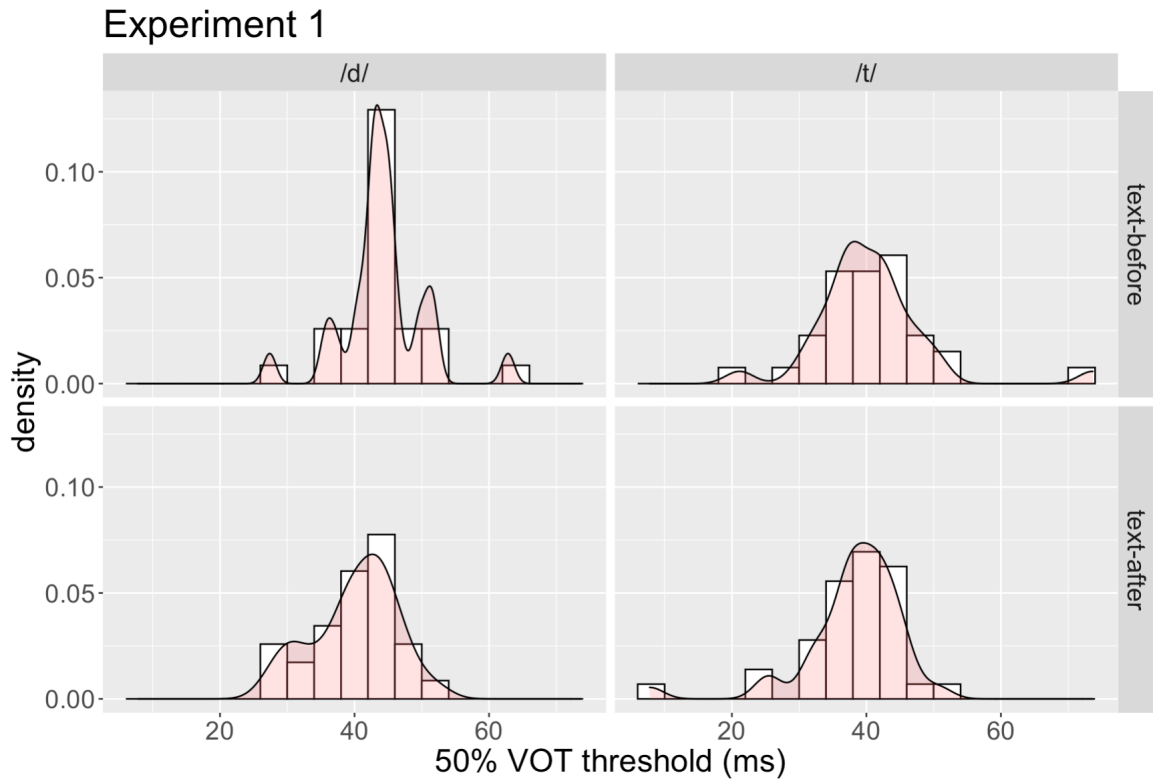


Figure 26: Distributions of the 50% categorization thresholds in Experiment 1 from the main manuscript. No evidence for bimodality was observed in any condition.

I used Hartigan’s dip test (Hartigan and Hartigan, 1985) for evidence for the alternative hypothesis of non-unimodality. No evidence for the alternative was present in any condition ( $D$ ’s  $< 0.065$ , uncorrected  $p$ ’s  $> .34$ ). This can also be seen in Figures 26-27, which depicts

the distributions of 50% categorization thresholds for each experiment.



Figure 27: Distributions of the 50% categorization thresholds in Experiment 2 from the main manuscript. No evidence for bimodality was observed in any condition.

#### A.5.2. Checking for Relationship between Exposure RTs and Test Performance

Perhaps one reason I failed to observe adaptation in the text-after conditions was due to a difference in response behavior of the participants in the two tasks (text-before vs. text-after), rather than about a difference in the intermediate representations available for adaptation. In particular, it may be that participants responded with different latency in the text-before or text-after conditions. Thus, if the sub-phonemic information requires a delay to come “on-line” for decision making, then that would result in minimal or no adaptation in the text-after condition, simply due to the difference in response times.

To test this, I conducted simple linear regressions predicting 50% categorization thresholds median response times (RTs) at training, separately for each experiment, with main effects and interactions of Shifted Phone and Timing. If the difference in RTs led to the adaptation



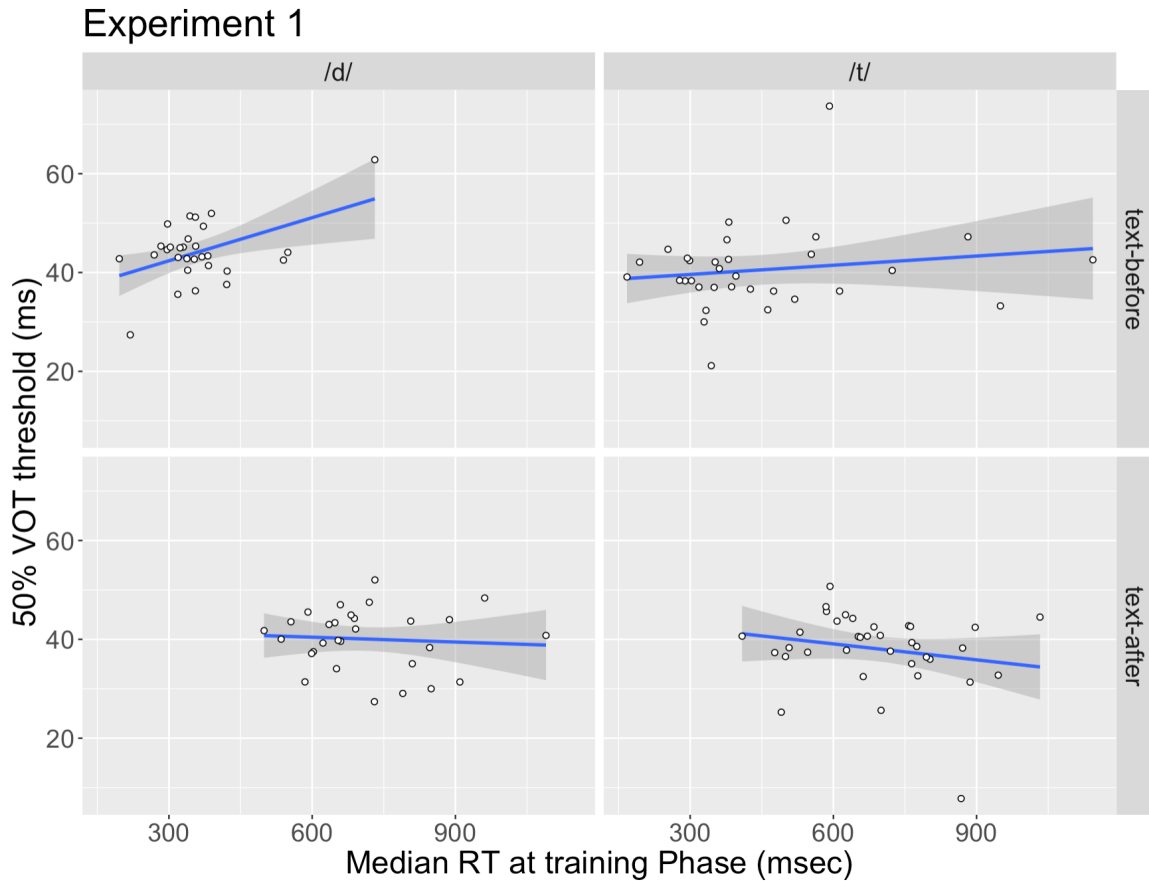


Figure 28: There is no significant relationship between RTs and categorization thresholds at test

differences that was observed, then we should see a significant relationship between median RTs and the thresholds (and possibly an interaction with condition). However, this was not the case. In Experiment 1, median RT was not a significant predictor of threshold, nor did it interact with the conditions of interest ( $p$ 's  $> .27$ ). In Experiment 2, median RT was also not a significant predictor of threshold, nor did it interact with the conditions of interest ( $p$ 's  $> .32$ ). This is visualized for Experiment 1 in Figure 28<sup>1</sup> and for Experiment 2 in Figure 29.

<sup>1</sup>Note that what appears to be a significant correlation in Experiment 1, text-before Shifted-/d/, is driven by one outlier participant: the participant with about 700ms median RT, and 62 ms categorization threshold. With this participant included, the Spearman correlation is  $\rho = 0.48$ ,  $p = 0.09$ ; but excluded, the correlation is  $\rho = 0.14$ ,  $p = 0.49$ .

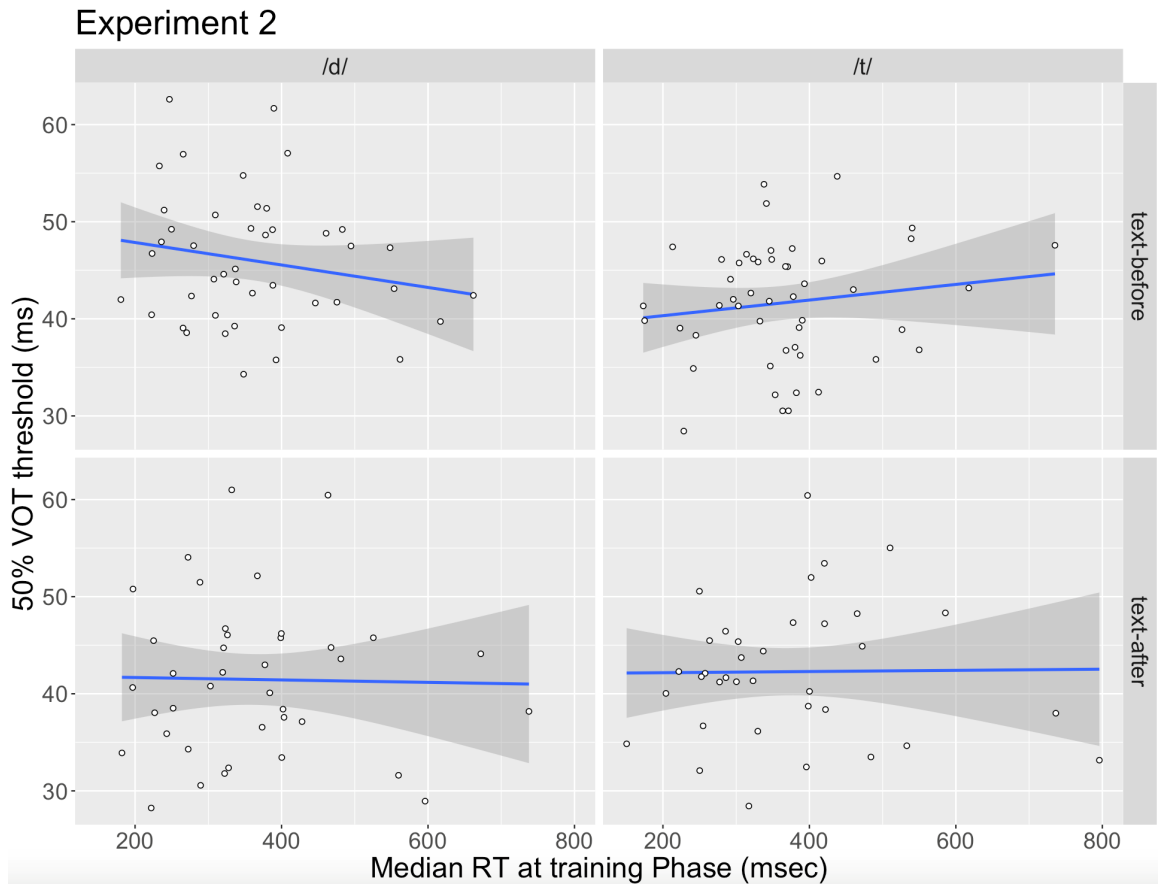


Figure 29: There is no significant relationship between RTs and categorization thresholds at test

## A.6. Effect of Pre-Registered Exclusion Criteria for All Experiments

In the pre-registered analysis plan for Experiment 2 and Experiment S1, I had an additional criterion to exclude those participants whose performance at the extrema of the VOT distributions (20ms and 80ms) was more than 0.15 away from ceiling or floor. This additional exclusion was added to the pre-registration after observing that some participants' psychometric functions in Experiment 1 did not conform to the usual "S" shape, due to deviance from floor/ceiling performance at the extrema. However, I ultimately decided to diverge from this pre-registered criterion because there was no theoretical reason to expect categorizations at our chosen extrema (e.g. 20ms VOT) to necessarily be at floor or ceiling. I note that excluding these participants did not qualitatively change the reported results in any of the experiments in the main manuscript. Below, I report the additional exclusions due to this criterion (Table 23) and results of the main analyses of Chapter 2 with these exclusions (Subsections A.6.1 - A.6.3).

<b>Experiment</b>	<b>Shifted Phone</b>	<b>Timing</b>	<b>N after exclusions</b>	<b>% excluded</b>
Exp 1	/d/	text-before	27	10%
Exp 1	/d/	text-after	28	3%
Exp 1	/t/	text-before	28	20%
Exp 1	/t/	text-after	31	18%
Exp 2	/d/	text-before	40	13%
Exp 2	/d/	text-after	30	29%
Exp 2	/t/	text-before	36	18%
Exp 2	/t/	text-after	32	27%
Exp S1	/d/	text-before	39	25%
Exp S1	/d/	text-after	36	18%
Exp S1	/t/	text-before	46	15%
Exp S1	/t/	text-after	33	25%

Table 23: Exclusions with ceiling/floor cutoff for each experiment (by condition).

### *A.6.1. Experiment 1 Results with Full Exclusions*

The best-fitting model was one including a main effect of VOT and a main effect of Test Half with main effects and interactions of Shifted Phone, Timing, and Test Half. This model was a better fit than one that did not include the interaction of Shifted Phone and Timing and the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(2) = 10.13$ ,  $p = .006$ , and better than a model without the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(1) = 7.49$ ,  $p = .006$ .

In the First Half, the best-fitting model was one that included a main effect of VOT and main effects and interactions of Shifted Phone and Timing. This model was a better fit than one that did not include the interaction of Shifted Phone and Timing,  $\chi^2(1) = 4.96$ ,  $p = .026$ . In contrast, in the Last Half, the best-fitting model was one that included a main effect of VOT, Shifted Phone, and Timing, and interactions of VOT and Shifted Phone, and VOT and Timing. A model with the additional interaction of Shifted Phone and Timing was not a significant improvement,  $\chi^2(1) = 0.79$ ,  $p = .37$ , nor was one with the additional triple interaction of VOT, Shifted Phone, and Timing,  $\chi^2(2) = 2.29$ ,  $p = .32$ .

### *A.6.2. Experiment 2 Results with Full Exclusions*

The best-fitting model was one including a main effect of VOT and main effects and interactions of Shifted Phone, Timing, and Test Half. This model was a better fit than one that did not include the interaction of Shifted Phone and Timing and the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(2) = 6.85$ ,  $p = .03$ , and better than a model without the triple interaction of Shifted Phone, Timing, and Test Half,  $\chi^2(1) = 4.16$ ,  $p = .04$ .<sup>2</sup>

In the First Half, the best-fitting model was indeed one that included a main effect of VOT and main effects and interactions of Shifted Phone and Timing. This model was a better fit

---

<sup>2</sup>This best fitting model with all main effects and interactions did not meet the cutoff of tolerance ( $\max|\text{grad}| = 0.00148$ , with default tolerance of 0.001). A simpler model without an intercept for the test exemplar did converge and the results are qualitatively similar.

Likewise, the model I compare to (without the triple interaction) did not meet the cutoff of tolerance for convergence ( $\max|\text{grad}| = 0.001009$ ). As before, a simpler model without the intercept for test exemplar converged and were qualitatively similar.

than one that did not include the interaction of Shifted Phone and Timing,  $\chi^2(1) = 3.96$ ,  $p = .047$ , and marginally better than one that included only a main effect of VOT,  $\chi^2(3) = 6.95$ ,  $p = .07$ . In contrast, in the Last Half, the effects of Shifted Phone and Timing were more subtle: a model that included main effects and interactions of VOT, Shifted Phone, and Timing was significantly better than one without the triple interaction of VOT, Shifted Phone, and Timing,  $\chi^2(1) = 5.16$ ,  $p = .02$ , and better than one without the full triple interaction or the interaction of Shifted Phone and Timing,  $\chi^2(2) = 6.75$ ,  $p = .034$ ; however, this model was not better than one that included only a main effect of VOT,  $\chi^2(6) = 11.0$ ,  $p = .09$ .

#### *A.6.3. Experiment S1 Results with Full Exclusions*

The best-fitting model was one including a main effect of VOT and main effects and interactions of Shifted Phone and Test Half. This model was a better fit than one that did not include the interaction of Shifted Phone and Test Half,  $\chi^2(1) = 38.4$ ,  $p < .001$ , and better than a model without a main effect or interactions of Shifted Phone,  $\chi^2(2) = 51.3$ ,  $p < .001$ . Including Timing as a main effect or interaction with any of the other factors did not improve the fit, all  $p$ 's  $> .22$ .

Given the significant interaction of Shifted Phone and Test Half, I tested for the effects of interest in each test phase half separately. In the First Half, the best-fitting model was indeed one that included a main effect of VOT and main effect of Shifted Phone. This model was a better fit than one that did not include the main effect of Shifted Phone,  $\chi^2(1) = 20.99$ ,  $p < .001$ . A model with the additional interaction of Shifted Phone and VOT was not a significantly better fit,  $\chi^2(1) = 1.48$ ,  $p = .224$ , nor was a model with the main effect or interaction of Shifted Phone and Timing,  $\chi^2(2) = 1.28$ ,  $p = .528$ . In the Last Half the results were largely the same. The best-fitting model was indeed one that included a main effect of VOT and main effect of Shifted Phone. This model was a better fit than one that did not include the main effect of Shifted Phone,  $\chi^2(1) = 4.57$ ,  $p = .033$ . A model with the additional interaction of Shifted Phone and VOT was not a significantly better fit,  $\chi^2(1) = 0.51$ ,  $p = .474$ , nor was a model with the main effect or interaction of Shifted Phone and

Timing,  $\chi^2(2) = 1.80$ ,  $p = .407$ .

#### A.7. Distribution of Participant Exclusions

In Experiment 1, the exclusion rate of 3% is even across conditions: 2 out of 65 text-before participants were excluded because of low (below 80%) accuracy on target exposure items, and 2 out of 67 text-after participants were excluded because of low (below 80%) accuracy on target exposure items. No participants were excluded for overly fast exposure response times.

For experiment 2, the increase in exclusions was driven primarily by overly fast response times by a minority of participants. 12 out of 106 text-before participants were excluded (for a total rate of 11.3%), however, 7 of those 12 were excluded for overly fast response times (RTs less than 150ms on more than 25% of all responses). Of the 5 text-before participants excluded for match-mismatch inaccuracy, 3 were due to poor performance on just filler items, while 2 were due to poor performance on either just target items or both target and filler items. For the text-after participants in Experiment 2, there were 13 participants excluded out of 88 (total rate of 14.8%). Again, the majority of these (8 of 13) were due to overly fast response times. 2 participants were excluded based on their test-phase responses (“t-responses” lower at 80ms than 20ms indicating either non-compliance or that they had accidentally flipped the response buttons). Only 3 of 88 text-after participants were excluded due to poor performance on the match-mismatch task (1 for low accuracy on confirming target items, and 2 for low accuracy on confirming filler items)

A.8. Visualizing three-way interactions in main experiments

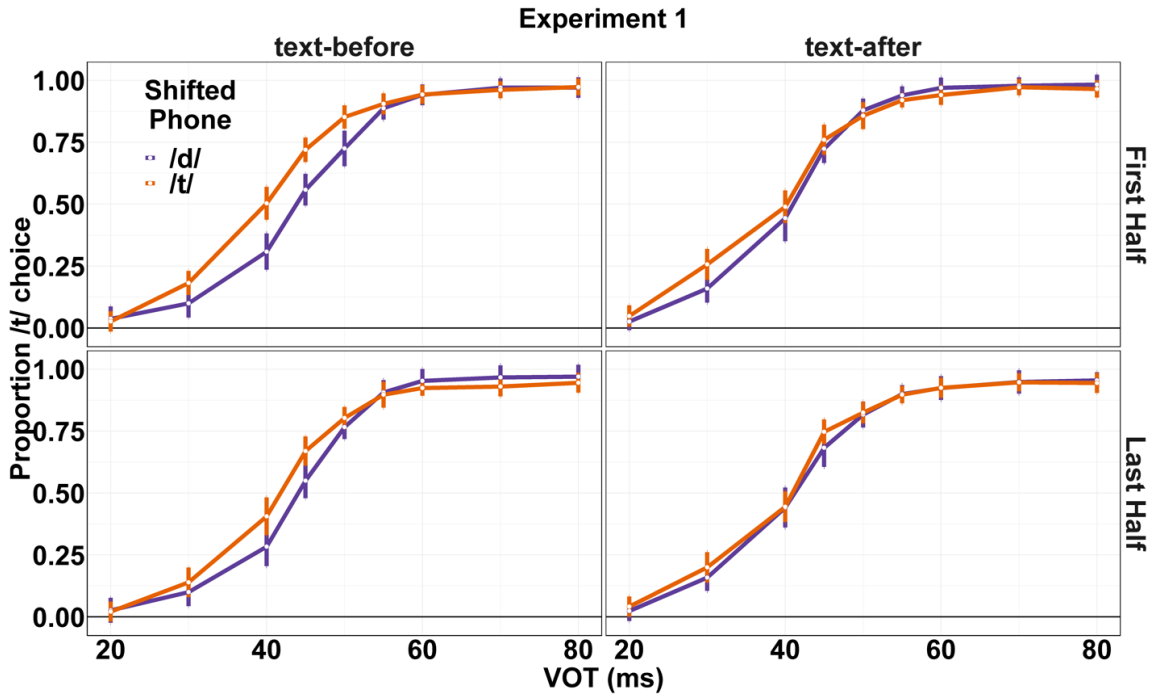


Figure 30: Psychometric functions for phoneme categorization during testing in Experiment 1. Output split by Shifted Phone (/t/ or /d/), Timing condition (text-before or text-after), and test-phase half (first four test blocks or remaining five). Adaptation occurred in the text-before but not text-after condition and faded over the course of the test phase (first vs. last half). Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008).

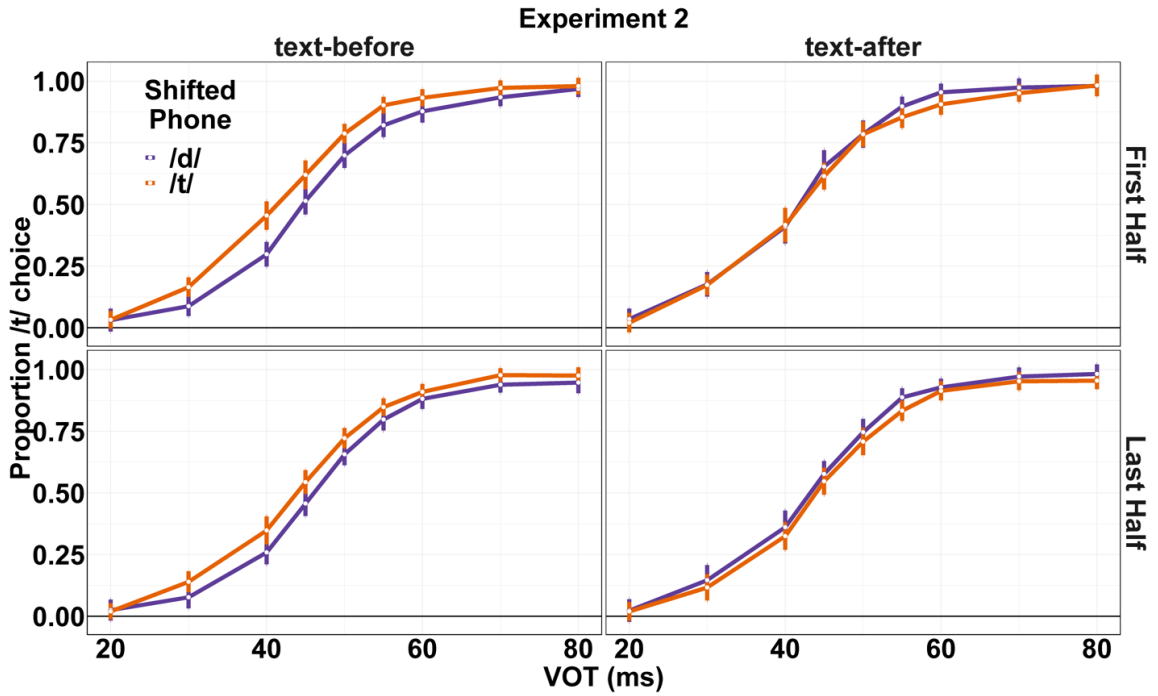


Figure 31: Psychometric functions for phoneme categorization during testing in Experiment 2. Output split by Shifted Phone (/t/ or /d/), Timing condition (text-before or text-after), and test-phase half (first four test blocks or remaining five). Adaptation occurred in the text-before but not text-after condition and faded over the course of the test phase (first vs. last half). Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008).



## A.9. Norming study

The audio for target stimuli in Experiment 1 was created by gluing different portions of “t”-word onsets onto the rime of the “d”-words. Since pitch contour ( $F_0$ ), which is a secondary cue to voicing (Dmitrieva et al., 2015; Hombert et al., 1979), is realized on the vowel, this means that while the VOT values were edited, all the target stimuli retained secondary information consistent with voicing (i.e. the “d” interpretation). To correct for this, I edited new versions of the target audio which additionally corrected for pitch ( $F_0$ ) and used a norming study to select a maximally ambiguous VOT cutoff.

The norming study consisted of only a single identical test phase for each participant. On each trial, participants were exposed to audio of either an isolated word or CV syllable with a t/d onset ranging from 10ms to 80ms VOT and asked to categorize the word as beginning with “t” or “d.”

### A.9.1. Design

The experiment consisted of 1,053 trials. On each trial, participants were exposed to audio with a “t/d” onset with a particular VOT value. After listening to the audio, participants were asked to make a categorization judgment as to whether the sound contained “t” or “d”. The 1,053 trials were divided between thirteen exemplars, nine VOT levels [20, 30, 40, 45, 50, 55, 60, 70, 80], with nine repetitions for each exemplar and level (13x9x9). Test items were randomized within a set of nine blocks, so each stimulus was heard once before any stimulus was repeated.

### A.9.2. Participants

I recruited 44 native English-speaking University of Pennsylvania undergraduates who received course credit for their participation. Participants were presented with stimuli in one of two randomly ordered lists.

### A.9.3. Stimuli

There were a total of twenty-six VOT continua: twenty-two full words and four CV syllables. The twenty-two full word continua were the target items used in Experiment 1 and the four

CV syllable continua were the same as the test stimuli used in Experiment 1.

Recordings were taken from the same speaker as the experiments in the Chapter 2. In order to remove  $F_0$ -contour as a cue to voicing, I manually extracted the pitch contours for each word-pair using PRAAT. A new  $F_0$ -onset value was chosen at 2/3rds of the gap between the d-onset and t-onset words. I used a PRAAT script to resynthesize the pitch-contours of the d-onset words with a new contour which begins at the designated 2/3rds boosted  $F_0$  value and follows a smooth cline (using pseudo-linear interpolation with a step-size of 10ms) down to the original d-word pitch at 160ms into the vocoid.

These  $F_0$ -modified stimuli were used to create continua with VOT ranging from 10 to 100ms in steps of 5ms. Each step was generated using the same splicing procedure as in Experiment 1: the onset of a t-word was glued onto the rime of the  $F_0$ -modulated d-word at the specified VOT value at the nearest zero-crossing point.

#### *A.9.4. Procedure*

The norming study was written using custom javascript code interfaced with psiTurk, as in Experiment 1. Pre-experiment questionnaire and volume check were the same as Experiment 1. Participants completed the experiment in a web browser and were encouraged to use headphones during the experiment, although use of headphones could not be verified.

Participants received instructions telling them to decide whether the audio they heard corresponds to the word starting with “t” or “d” with both choices displayed on screen (e.g. “time” or “dime”). The side for choosing t-word vs. d-word were consistent within participant but randomized between participants.

#### *A.9.5. Exclusions*

I excluded participants if they did not complete the entire study, if 20% or more of their response times were  $< 75$  ms, if their proportion /t/ responses were not significantly lower for the 3 lowest and 3 highest VOT levels, or if the proportion /t/ responses at VOT extrema were  $> .33$  away from ceiling/floor. After these exclusions, 23 participants remained for the analyses.

### A.9.6. Analysis

I fit separate psychometric functions for each continuum using maximum likelihood estimation with the R package *quickpsy*. The variable of interest was the 50% threshold values at which test stimuli were equally likely to be categorized as “t” or “d”.

### A.9.7. Results

The median 50% threshold was 46.9ms VOT. Among the stimuli used, this was closest to the 45ms VOT items, so that is the level I established to use as the ambiguity point for the exposure phase of Experiments 2 and S1. Figure 32 shows a violin plot with the median 50% threshold for each continuum.

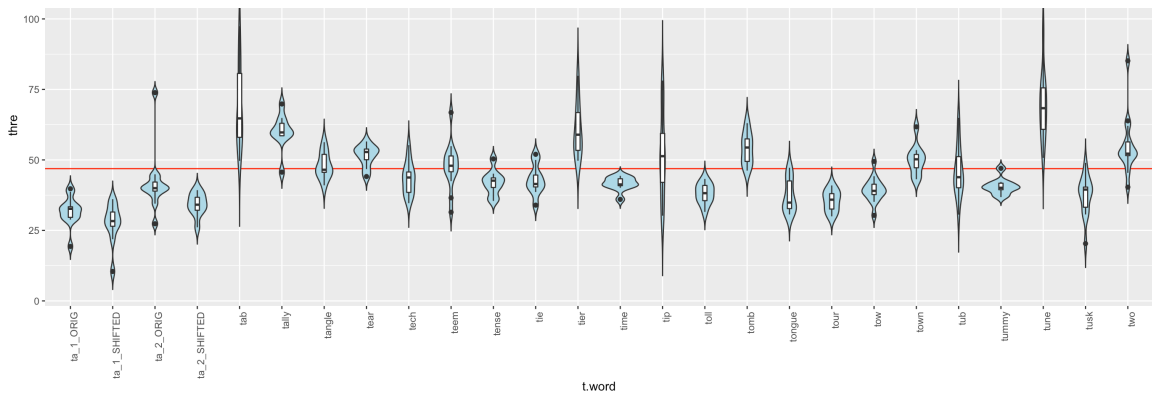


Figure 32: Violin plot of the median 50% threshold for “t” / “d” categorization for each continuum in the norming study. Red line shows the overall median 50% threshold at 46.9ms. “\_SHIFTED” and “\_ORIG” correspond to pitch-edited and original-pitch CV continua respectively.

## A.10. Experiment S1

In Experiment S1, a confound was introduced between “edited speech” and phonological category (either /t/ or /d/ depending on condition). Thus, the ambiguous target items were paired with the same audio files as in Experiment 2, but the unambiguous target words were paired with audio that had not been run through the VOT- and pitch-manipulation scripts. This assignment is illustrated in Figure 33. In Experiments 1 and 2 tokens of “edited speech” were evenly balanced during the exposure phase between ambiguous and unambiguous targets (between /t/ and /d/) and thus could not directly be recruited for learning the shifted distribution. However, in Experiment S1, the confound between acoustic-manipulation and

a speech category could be represented by participants in a number of ways (building a novel category for “edited speech”, falsely recognizing that the edited and unedited tokens belong to two different speakers, etc.). Thus participants might be able to learn a general bias, e.g. “edited speech is always a /d/”, rather than needing to directly update their phone boundaries.

#### A.10.1. Method

##### Design

The design for Experiment S1 matched the design for Experiment 2 in all aspects except for the assignment of audio to unambiguous target words. In Experiment S1, the unambiguous targets were paired with completely unedited audio as in Figure S3. All design, analyses, and exclusions were pre-registered.

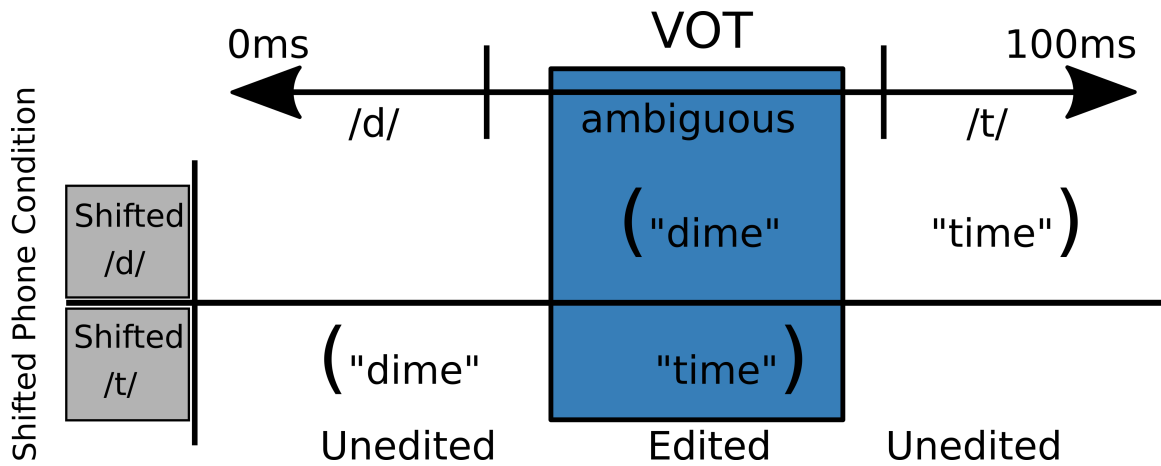


Figure 33: Assignment of audio to text in Experiment S1. There is a confound between “edited speech” and the particular phonological category under manipulation.

##### Participants

I recruited 176 participants from Amazon Mechanical Turk divided evenly between the same four exposure conditions as in Experiments 1 and 2 (text-before with shifted-/d/, text-before with shifted-/t/, text-after with shifted-/d/, and text-after with shifted-/t/). Subjects were paid \$2.41 for their participation.

## Stimuli and Procedure

The stimuli were the same VOT- and pitch-corrected items as in Experiment 2 with the exception of the change to unambiguous target items. Only the “ambiguous” target items for the exposure phase and the test items were run through the audio-editing scripts, the unambiguous target items were simply the original unedited audio with only the volume normalized for average loudness. The procedure for Experiment S1 remained unchanged in Experiment 2.

## Exclusions and Analysis

Exclusions and analyses were identical to those in Experiment 1. This resulted in 159 remaining participants for analysis (exclusion rate of 10%), divided among the conditions in the following way: 40 in text-before shifted-/t/, 44 in text-before shifted-/d/, 39 in text-after shifted-/t/, 36 in text-after shifted-/d/.

### *A.10.2. Results*

In the exposure phase, performance of the included participants was high and was comparable across conditions: accuracy in confirming the audio/subtitle match on unambiguous target items was above 98%, on ambiguous targets was above 96%, and on fillers was above 96%. This suggests that for the included participants, the matching task at exposure was not any more difficult in one condition over another.

Data from the test phase appear in Figure 34 (split by Shifted Phone and timing condition). As can be observed, adaptation was successful: the psychometric functions are different between shifted-/t/ and -/d/ ranges. In contrast to the results from both Experiments 1 and 2, this adaptation effect was observed in both timing conditions, text-before and text-after. As before, the adaptation effect (i.e. effect of Shifted Phone condition on categorization) faded over time, weakening by the second half of the test phase.

These results were confirmed in mixed-effects model comparisons. First, I compared models over all of the data. The best-fitting model was one including a main effect of VOT and

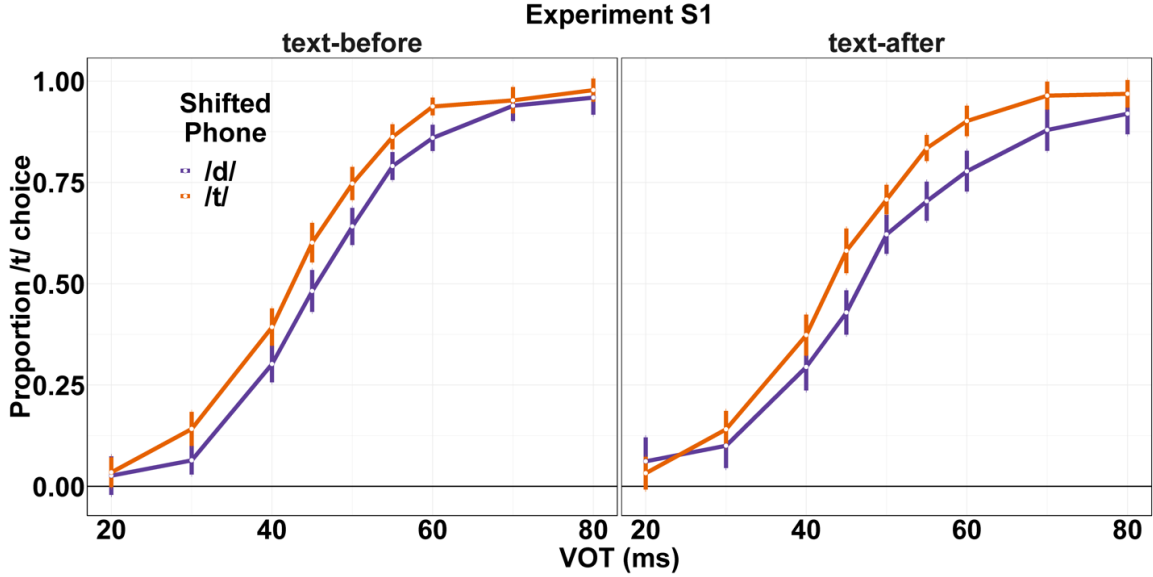


Figure 34: Main results for Experiment S1. Psychometric functions for phoneme discrimination during testing. Output split by Shifted Phone (/t/ or /d/) and Timing condition (text-before or text-after). Unlike in Experiments 1 and 2, adaptation occurred in both the text-before and text-after conditions. Data points are subject means and error bars are within-subject 95% confidence intervals (Morey, 2008).

main effects and interactions of Shifted Phone and Test Half. This model was a better fit than one that did not include the interaction of Shifted Phone and Test Half,  $\chi^2(1) = 46.72$ ,  $p < .001$ , and better than a model without a main effect of Shifted Phone or its interaction with Test Half,  $\chi^2(2) = 61.8$ ,  $p < .001$ . Including Timing as a main effect or interaction with any of the other factors did not improve the fit, all  $p$ 's  $> .19$ .

Given the significant interaction of Shifted Phone and Test Half, I next tested for the effects of interest (Shifted Phone and Timing) in each test phase half separately. In the First Half, the best-fitting model was indeed one that included a main effect of VOT and main effect of Shifted Phone. This model was a better fit than one that did not include the main effect of Shifted Phone,  $\chi^2(1) = 25.3$ ,  $p < .001$ . A model with the additional interaction of Shifted Phone and VOT was not a significantly better fit,  $\chi^2(1) = 1.66$ ,  $p = .198$ , nor was a model with the main effect of Timing and its interaction with Shifted Phone,  $\chi^2(2) = 1.90$ ,  $p = .39$ . In the Last Half, the results were largely the same. The best-fitting model was indeed

one that included a main effect of VOT and main effect of Shifted Phone. This model was a better fit than one that did not include the main effect of Shifted Phone,  $\chi^2(1) = 4.95$ ,  $p = .03$ . A model with the additional interaction of Shifted Phone and VOT was not a significantly better fit,  $\chi^2(1) = 0.54$ ,  $p = .46$ , nor was a model with the main effect of Timing and its interaction with Shifted Phone,  $\chi^2(2) = 2.24$ ,  $p = .33$ .

Finally, I directly compared the effect of Shifted Phone separately in the two Timing conditions, text-before and text-after, to confirm that the effect was indeed present in both the text-before and text-after conditions (First Half of test phase only). For text-before, the best-fitting model was one that included main effects of VOT and Shifted Phone. This model was a better fit than one that did not include the effect of Shifted Phone,  $\chi^2(1) = 13.0$ ,  $p = < .001$ . Similarly, an effect of Shifted Phone was also observed in the text-after condition. The best-fitting model was one that included only main effects of VOT and Shifted Phone, as well as their interaction. This model was a better fit than one that did not include the main effect of Shifted Phone or its interaction with VOT,  $\chi^2(2) = 17.0$ ,  $p < .001$ , and better than one that did not also include the interaction of Shifted Phone and VOT,  $\chi^2(1) = 4.26$ ,  $p = .04$ . These modeling results demonstrate that the adaptation was present and detectable in both the text-before and text-after conditions.

### *A.10.3. Discussion*

Due to the confound between edited audio and a phonological category, participants were able to learn a general mapping between edited tokens (whether represented directly, or represented as speech from two different speakers, etc.) and a phonological category (either /t/ or /d/) in both the text-before and text-after conditions. While I believe the primary takeaway from Experiment S1 should be that participants are capable of rapidly learning a large range of possible correlations when stimuli are not properly controlled, I also note that these results do not entail a signal-retention interpretation. Under AOC participants in Experiment S1 were able to learn a general bias between phonemes and “editedness” (or interpreting the edited and unedited tokens as coming from two different speakers, etc.). This is possible since symbolic/category representations, whether lexical, phonemic, or speaker-

status persist over time in a way that the signal does not.



## APPENDIX B

### B.1. Parameter Sensitivity

Parameter	Tuned Value
Prominent salience mean	0.5
Non-prominent salience mean	0.4
Salience standard deviation	0.25
Distance threshold	0.4
Semantic incompatibility parameter	0.1

Table 24: Tuned parameter values from Section 3.4.3

### B.2. Gradient analysis of PSE in Lewis and Frank (2018)

The main analysis for generalization-level was scored using the binary outcome of broad (basic) vs. narrow (subordinate) meanings. This excluded the minority of trials on which a participant selected some, but not all, of the basic-level matches. To make sure that my analyses of SCE and PSE are not being disrupted by removing these potentially “uncertain” participants, I additionally ran analyses over the whole set of trials (including the 6.7%, 104 out of 1560, of trials with mixed test selections). Generalization level outcomes were coded as a gradient measure and fit with mixed-effects linear regressions. This alternative coding scheme did not have a significant effect on the presence of either SCE or PSE (See Table 25 for all trials) nor on the shape of the three-way interaction between Presentation-Style, Training-Number, and Block-Order (see Table 26 for second-block trials and Table 27 for first-block trials).

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	0.303	0.043	7.073	.015
Presentation-Style (PSE)	-0.048	0.023	-2.113	.035
Training-Number (SCE)	0.197	0.012	16.414	<.001
Block-Order	-0.181	0.023	-8.051	<.001
Presentation x Number (NTI)	0.020	0.024	0.833	.405
Presentation x Block	-0.015	0.045	-0.336	.737
Number x Block	-0.535	0.024	-22.242	<.001
Presentation x Number x Block	0.161	0.048	3.339	<.001

Table 25: Data from Lewis and Frank (2018). Dependent variable is the generalization-level outcome on all trials. Linear mixed model predicting generalization based on listed effects as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects along with a three-way interaction between Presentation-Style, Training-Number, and Block-Order.

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	0.170	0.044	3.827	.052
Presentation-Style (PSE)	-0.007	0.025	-0.298	.766
Training-Number (SCE)	0.016	0.025	0.647	.518
Presentation x Number (NTI)	0.005	0.050	0.099	.921

Table 26: Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization proportion on second-block trials. Linear mixed model predicting generalization based on presentation-style, training-number, the presentation-number interaction, as well as random slopes for subject and stimulus class. Neither SCE nor PSE manifest on second-block trials.

<b>Predictor</b>	<b>Coefficient</b>	<b>Std. Error</b>	<b>z</b>	<b>P(&gt; z )</b>
(Intercept)	0.437	0.043	10.121	.004
Presentation-Style (PSE)	-0.088	0.032	-2.769	.006
Training-Number (SCE)	0.379	0.032	11.948	<.001
Presentation x Number (NTI)	0.035	0.063	0.554	.579

Table 27: Data from Lewis and Frank (2018). Dependent variable is the outcome of broad vs. narrow generalization on first-block trials. Linear mixed model predicting generalization based on presentation-style, training-number, the presentation-number interaction, as well as random slopes for subject and stimulus class. PSE and SCE emerge as significant main effects.

## APPENDIX C

### C.1. Nonce word labels

Blicket	Bugorn	Forbo
Gaka	Gronan	Lopus
Mipa	Pipit	Ralex
Ratat	Sipot	Talet
Torun	Vatrus	Wagnum
Zened		

Table 28: Disyllabic nonce word labels used in Experiment 1 (Chapter 4)

### C.2. Possible Feature Alternations

The stimuli used in Experiment 1 (Chapter 4). With five binary features, each object is one of 32 possible instantiations per domain.

Location	Feature	Alternation
West	Head	Pointy / Round
Center	Body	Striped / Spotted
South	Legs	Six / Four
North	Back	Thin / Wide
East	Tail	Stinger / Curly

(a) “Bug” alternations

Location	Feature	Alternation
West	Tail	Straight / Wavy
Center	Body	Feathers / Diamonds
South	Legs	Toes / Webbed
North	Back	Big-Wing / Leaf-Wings
East	Head	Mohawk / Beanie

(b) “Bird” alternations

Location	Feature	Alternation
West	Awning	Straight / Curved
Center	Windows	Circles / Balcony
South	Path	Striped / Bricks
North	Roof	Chimneys / Slanted
East	Yard	Deciduous / Evergreens

(c) “House” alternations

Location	Feature	Alternation
West	Front	Round / Pointed
Center	Window	Single / Circles
South	Arm	Net / Claw
North	Top	Antenna / Telescope
East	Motor	Fan / Turbine

(d) “Submarine” alternations

Table 29: Potential feature alternations for each domain.

### C.3. Gaze Heatmaps

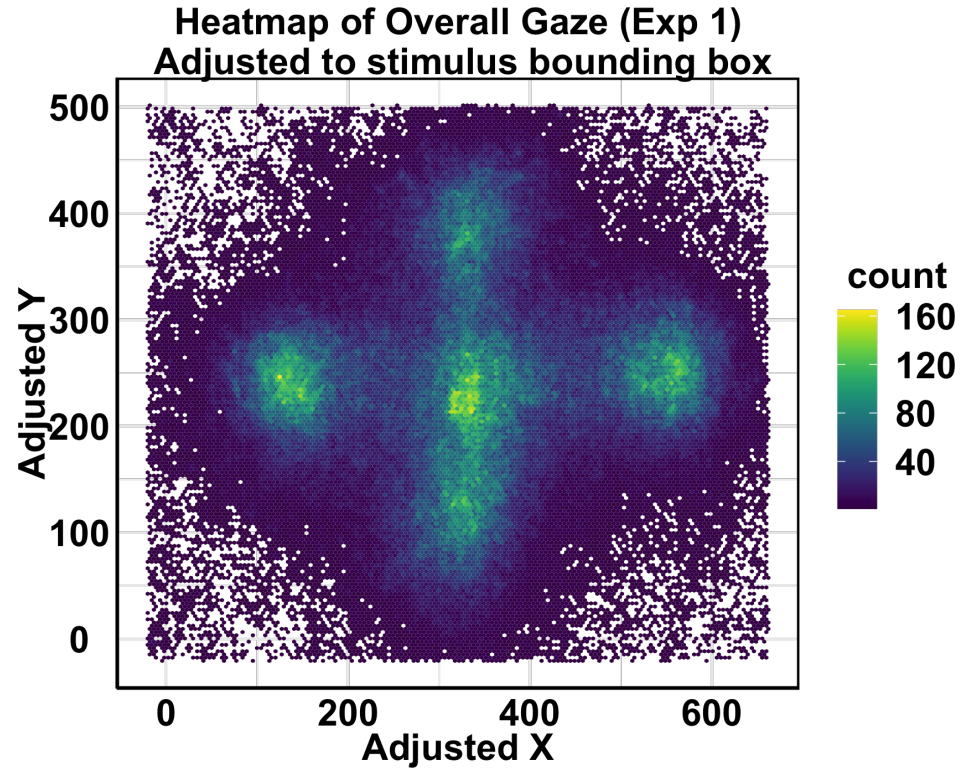
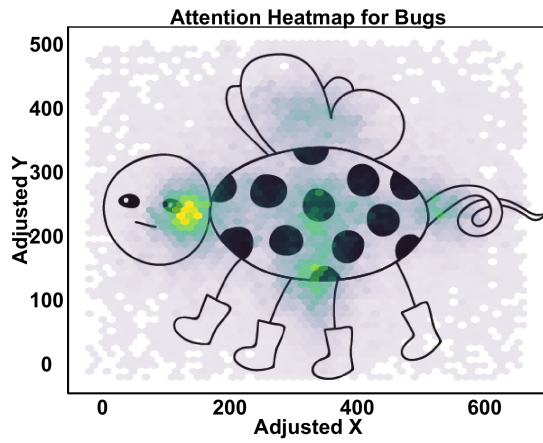
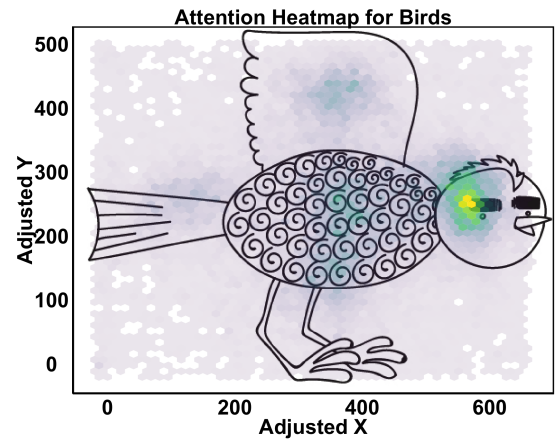


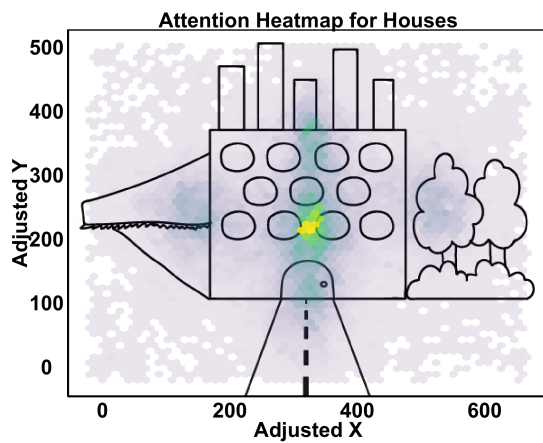
Figure 35: Heatmap of gaze (within stimulus bounding box) throughout all training trials and all stimulus domains.



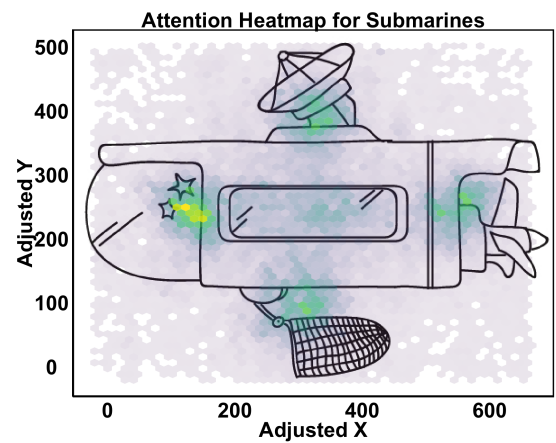
(a) Bugs



(b) Birds



(c) Houses



(d) Submarines

Figure 36: Heatmaps of overall gaze split by domain and overlaid on example stimulus

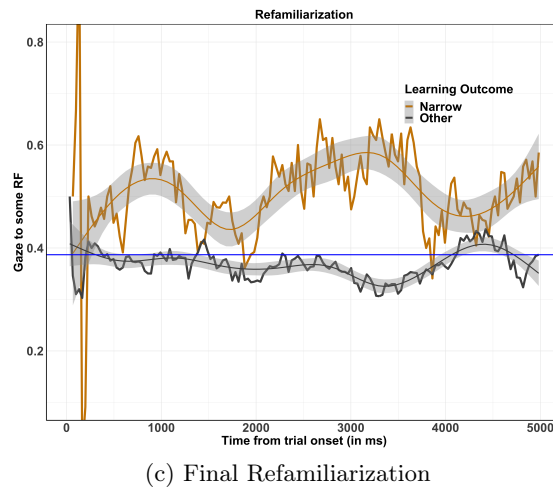
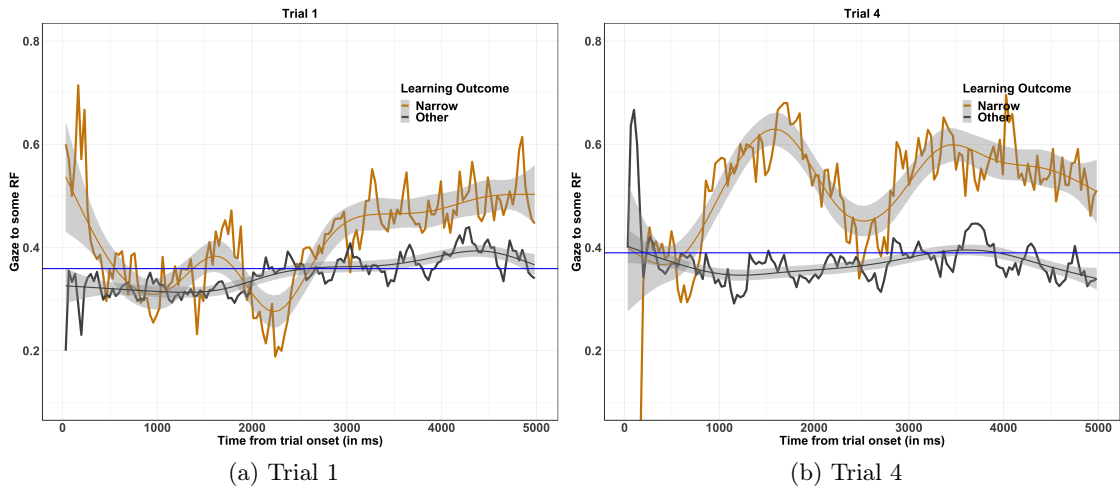


Figure 37: Plots showing timecourse of gaze-time to the RF-set as a function of learning outcome

#### C.4. Timecourse Plots

## BIBLIOGRAPHY

- B. Aarts. Verb-preposition constructions and small clauses in english. *Journal of Linguistics*, 25(2):277–290, 1989. doi: 10.1017/s0022226700014109.
- J. R. Anderson. *The Adaptive Character of Thought*. Psychology Press, 1990. doi: 10.4324/9780203771730.
- D. Ariely. Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2):157–162, 2001. doi: 10.1111/1467-9280.00327.
- J. E. Arnold, T. Wasow, A. Asudeh, and P. Alrenga. Avoiding attachment ambiguities: The role of constituent ordering. *Journal of Memory and Language*, 51(1):55–70, 2004. doi: 10.1016/j.jml.2004.03.006.
- E. L. Axelsson, L. K. Perry, E. J. Scott, and J. S. Horst. Near or far: The effect of spatial distance and vocabulary knowledge on word learning. *Acta Psychologica*, 163:81–87, 2016. doi: 10.1016/j.actpsy.2015.11.006.
- R. H. Baayen, R. Piepenbrock, and L. Gulikers. The celex lexical database (release 2). *Distributed by the Linguistic Data Consortium, University of Pennsylvania*, 1995.
- M. T. Balaban and S. R. Waxman. Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology*, 64(1):3–26, 1997. doi: 10.1006/jecp.1996.2332.
- D. A. Baldwin. Infants’ contribution to the achievement of joint reference. *Child development*, 62(5):874–890, 1991.
- D. J. Barr. Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in psychology*, 4:328, 2013.
- A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *The Journal of the Acoustical Society of America*, 113(2):1001–1024, 2003. doi: 10.1121/1.1534836.
- P. Bertelson, J. Vroomen, and B. De Gelder. Visual recalibration of auditory speech identification: A mcgurk aftereffect. *Psychological Science*, 14(6):592–597, 2003.
- D. Biber. A register perspective on grammar and discourse: Variability in the form and use of english complement clauses. *Discourse Studies*, 1(2):131–150, 1999. doi: 10.1177/1461445699001002001.
- K. Bicknell, T. F. Jaeger, and M. K. Tanenhaus. Now or... later: perceptual data are not immediately forgotten during language processing. *Behavioral and Brain Sciences*, 39, 2016.



- P. Bloom. *How children learn the meanings of words*, volume 377. MIT Press, 2000.
- K. Bock. An effect of the accessibility of word forms on sentence structures. *Journal of Memory and Language*, 26(2):119–137, 1987. doi: 10.1016/0749-596x(87)90120-3.
- K. Bock and W. Levelt. Language production. *Psycholinguistics: Critical concepts in psychology*, 5:405, 2002.
- K. Bock and W. J. Levelt. *Language production: Grammatical encoding*. Academic Press, 1994.
- E. Bonawitz, S. Denison, A. Gopnik, and T. L. Griffiths. Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, 74:35–65, 2014. doi: 10.1016/j.cogpsych.2014.06.003.
- M. Bowerman. The “no negative evidence” problem: How do children avoid constructing an overly general grammar? In *Explaining language universals*, pages 73–101. Basil Blackwell, 1988.
- A. R. Bradlow and T. Bent. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729, 2008.
- M. D. Braine et al. On two types of models of the internalization of grammars. *The ontogenesis of grammar*, 1971:153–186, 1971.
- J. Bresnan, A. Cueni, T. Nikitina, and R. H. Baayen. Predicting the dative alternation. *Cognitive foundations of interpretation*, pages 69–94, 2007.
- L. R. Brooks. Nonanalytic concept formation and memory for instances. *Cognition and Categorization*, 1978.
- R. Brown and C. Hanlon. Derivational complexity and order of acquisition in child speech. In J. R. Hayes, editor, *Cognition and the Development of Language*, pages 11–53. New York: Wiley, 1970.
- S. Brown-Schmidt and A. E. Konopka. Processes of incremental message planning during conversation. *Psychonomic Bulletin & Review*, 22(3):833–843, 2014. doi: 10.3758/s13423-014-0714-2.
- S. Brown-Schmidt and J. C. Toscano. Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience*, 32(10):1211–1228, 2017.
- A. Bungler, A. Papafragou, and J. C. Trueswell. Event structure influences language production: Evidence from structural priming in motion event description. *Journal of Memory and Language*, 69(3):299–323, 2013. doi: 10.1016/j.jml.2013.04.002.

- Z. Burchill, L. Liu, and T. F. Jaeger. Maintaining information about speech input during accent adaptation. *PloS one*, 13(8):e0199358, 2018.
- P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80(1):1–28, 2017.
- W. Bushong and T. F. Jaeger. Maintenance of perceptual information in speech perception. In *CogSci*, 2017.
- J. Bybee. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language variation and change*, 14(3):261–290, 2002.
- P.-C. Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software, Articles*, 80(1):1–28, 2017. doi: 10.18637/jss.v080.i01. URL <https://www.jstatsoft.org/v080/i01>.
- S. Caplan and K. Djärv. What usage can tell us about grammar: Embedded verb second in scandinavian. *Glossa: a journal of general linguistics*, 4(1), 2019.
- S. Caplan, J. Kodner, and C. Yang. Miller’s monkey updated: Communicative efficiency and the statistics of words in natural language. *Cognition*, 205:104466, 2020.
- S. Caplan, A. Hafri, and J. C. Trueswell. Now you hear me, later you don’t: The immediacy of linguistic computation and the representation of speech. *Psychological Science*, 32(3):410–423, 2021.
- S. Carey. The child as word learner. In M. Halle, J. Bresnan, and G. Miller, editors, *Linguistic theory and psychological reality*, pages 264–293. Cambridge, MA: MIT Press, 1978.
- P. F. Carvalho and R. L. Goldstone. Putting category learning in order: Category structure and temporal arrangement affect the benefit of interleaved over blocked study. *Memory & cognition*, 42(3):481–495, 2014.
- P. F. Carvalho and R. L. Goldstone. What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.00505.
- N. Chomsky. Syntactic structures. *Language*, 33(3 Part 1):375–408, 1957.
- M. H. Christiansen and N. Chater. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and brain sciences*, 39, 2016.
- H. H. Clark and J. E. F. Tree. Using uh and um in spontaneous speaking. *Cognition*, 84(1):73–111, 2002. doi: 10.1016/s0010-0277(02)00017-3.
- M. Clayards, M. K. Tanenhaus, R. N. Aslin, and R. A. Jacobs. Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, 108(3):804–809, 2008.

- P. Collins. The indirect object construction in english: An informational approach. *Linguistics*, 33(1):35–50, 1995.
- E. Colunga and L. B. Smith. From the lexicon to expectations about kinds: A role for associative learning. *Psychological review*, 112(2):347, 2005.
- C. M. Connine, D. G. Blasko, and M. Hall. Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraint. *Journal of Memory and Language*, 30(2):234–250, 1991.
- A. Cristia, A. Seidl, C. Vaughn, R. Schmale, A. Bradlow, and C. Floccia. Linguistic processing of accented speech across the lifespan. *Frontiers in psychology*, 3:479, 2012.
- R. G. Crowder and J. Morton. Precategorical acoustic storage (pas). *Perception & Psychophysics*, 5(6):365–373, 1969.
- C. J. Darwin and A. D. Baddeley. Acoustic memory and the perception of speech. *Cognitive psychology*, 6(1):41–60, 1974.
- I. Dautriche and E. Chemla. What homophones say about words. *PLoS One*, 11(9):e0162176, 2016.
- M. Davies. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190, 2009.
- K. De Smedt and G. Kempen. Segment grammar: A formalism for incremental sentence generation. In *Natural language generation in artificial intelligence and computational linguistics*, pages 329–349. Springer, 1991.
- N. Dehé. *Particle verbs in English: Syntax, information structure and intonation*, volume 59. John Benjamins Publishing, 2002.
- G. S. Dell. Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science*, 9(1):3–23, 1985.
- H. Deubel and W. X. Schneider. Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision research*, 36(12):1827–1837, 1996.
- O. Dmitrieva, F. Llanos, A. A. Shultz, and A. L. Francis. Phonological status, not voice onset time, determines the acoustic realization of onset f0 as a secondary voicing cue in spanish and english. *Journal of Phonetics*, 49:77–95, 2015.
- J. R. Drouin and R. M. Theodore. Lexically guided perceptual learning is robust to task-based changes in listening strategy. *The Journal of the Acoustical Society of America*, 144(2):1089–1099, 2018.

- S. F. Ehrlich and K. Rayner. Contextual effects on word perception and eye movements during reading. *Journal of verbal learning and verbal behavior*, 20(6):641–655, 1981.
- J. Elness. That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies*, 65(6):519–533, 1984. doi: 10.1080/00138388408598357.
- L. L. Emberson, N. Loncar, C. Mazzei, I. Traves, and A. E. Goldberg. The blow-fish effect: children and adults use atypical exemplars to infer more narrow categories during word learning. *Journal of Child Language*, 46(05):938–954, 2019. doi: 10.1017/s0305000919000266.
- I. Erev and G. Barron. On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological review*, 112(4):912, 2005.
- W. K. Estes. *Classification and cognition*. Oxford University Press, 1994.
- J. B. Falandays, S. Brown-Schmidt, and J. C. Toscano. Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language*, 112: 104088, 2020.
- P. Farrell. English verb-preposition constructions: Constituency and order. *Language*, 81(1):96–137, 2005.
- A. Fazly, A. Alishahi, and S. Stevenson. A probabilistic computational model of cross-situational word learning. *Cognitive Science*, 34(6):1017–1063, 2010.
- M. Fedzechkina, T. F. Jaeger, and E. L. Newport. Language learners restructure their input to facilitate efficient communication. *Proceedings of the National Academy of Sciences*, 109(44):17897–17902, 2012.
- B. Ferguson and S. Waxman. Linking language and categorization in infancy. *Journal of child language*, 44(3):527–552, 2017.
- F. Ferreira and J. Henderson. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language*, 30(6):725–745, 1991.
- F. Ferreira and B. Swets. How incremental is language production? evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language*, 46(1):57–84, 2002.
- V. Ferreira. Is it better to give than to donate? syntactic flexibility in language production. *Journal of memory and language*, 35(5):724–755, 1996.
- V. Ferreira and G. Dell. Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive psychology*, 40(4):296–340, 2000.
- V. Ferreira and M. Hudson. Saying “that” in dialogue: The influence of accessibility and

- social factors on syntactic production. *Language and cognitive processes*, 26(10):1736–1762, 2011.
- A. B. Fine, T. F. Jaeger, T. A. Farmer, and T. Qian. Rapid expectation adaptation during syntactic comprehension. *PloS one*, 8(10):e77661, 2013.
- J. A. Fodor. *The modularity of mind: An essay on faculty psychology*. MIT press, 1983.
- K. I. Forster and S. M. Chambers. Lexical access and naming time. *Journal of Memory and Language*, 12(6):627, 1973.
- K. I. Forster and C. Davis. Repetition priming and frequency attenuation in lexical access. *Journal of experimental psychology: Learning, Memory, and Cognition*, 10(4):680, 1984.
- A. F. Frank and T. F. Jaeger. Speaking rationally: Uniform information density as an optimal strategy for language production. In *Proceedings of the annual meeting of the cognitive science society*, volume 30, 2008.
- S. L. Frank, L. J. Otten, G. Galli, and G. Vigliocco. The erp response to the amount of information conveyed by words in sentences. *Brain and language*, 140:1–11, 2015.
- C. Frankish. Precategorical acoustic storage and the perception of speech. *Journal of Memory and Language*, 58(3):815–836, 2008.
- V. A. Fromkin. The non-anomalous nature of anomalous utterances. *Language*, pages 27–52, 1971.
- M. E. Galle, J. Klein-Packard, K. Schreiber, and B. McMurray. What are you waiting for? real-time integration of cues for fricatives suggests encapsulated auditory memory. *Cognitive science*, 43(1):e12700, 2019.
- C. R. Gallistel. *The organization of learning*. The MIT Press, 1990.
- W. F. Ganong. Phonetic categorization in auditory word perception. *Journal of experimental psychology: Human perception and performance*, 6(1):110, 1980.
- S. A. Gelman and E. M. Markman. Categories and induction in young children. *Cognition*, 23(3):183–209, 1986.
- D. Gentner and L. L. Namy. Comparison in the development of categories. *Cognitive development*, 14(4):487–513, 1999.
- D. Gentner and L. L. Namy. Analogical processes in language learning. *Current Directions in Psychological Science*, 15(6):297–301, 2006.
- J. Gervain, M. Nespors, R. Mazuka, R. Horie, and J. Mehler. Bootstrapping word order in prelexical infants: A japanese–italian cross-linguistic study. *Cognitive psychology*, 57(1):56–74, 2008.

- L. M. Getz and J. C. Toscano. Electrophysiological evidence for top-down lexical influences on early speech perception. *Psychological science*, 30(6):830–841, 2019.
- M. L. Gick and K. J. Holyoak. Schema induction and analogical transfer. *Cognitive psychology*, 15(1):1–38, 1983.
- J. Gillette, H. Gleitman, L. Gleitman, and A. Lederer. Human simulations of vocabulary learning. *Cognition*, 73(2):135–176, 1999.
- T. Givón. The grammar of referential coherence as mental processing instructions. *Linguistics*, 30(1):5–56, 1992.
- L. Gleitman. The structural sources of verb meanings. *Language acquisition*, 1(1):3–55, 1990.
- L. R. Gleitman, D. January, R. Nappa, and J. C. Trueswell. On the give and take between event apprehension and utterance formulation. *Journal of memory and language*, 57(4):544–569, 2007.
- J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE, 1992.
- E. M. Gold. Language identification in the limit. *Information and control*, 10(5):447–474, 1967.
- S. D. Goldinger. Echoes of echoes? an episodic theory of lexical access. *Psychological review*, 105(2):251, 1998.
- S. T. Gries. *Multifactorial analysis in corpus linguistics: A study of particle placement*. A&C Black, 2003.
- T. M. Gureckis, J. Martin, J. McDonnell, A. S. Rich, D. Markant, A. Coenen, D. Halpern, J. B. Hamrick, and P. Chan. psiturk: An open-source framework for conducting replicable behavioral experiments online. *Behavior research methods*, 48(3):829–842, 2016.
- L. Gwilliams, T. Linzen, D. Poeppel, and A. Marantz. In spoken word recognition, the future predicts the past. *Journal of Neuroscience*, 38(35):7585–7599, 2018.
- J. T. Hale. *Automaton theories of human sentence comprehension*. CSLI Publications Stanford, CA, 2014.
- J. Hankamer. Unacceptable ambiguity. *Linguistic inquiry*, 4(1):17–68, 1973.
- J. A. Hartigan and P. M. Hartigan. The dip test of unimodality. *The annals of Statistics*, 13(1):70–84, 1985.

- J. A. Hawkins. *A performance theory of order and constituency*, volume 73. Cambridge University Press, 1994.
- J. A. Hawkins. *Efficiency and complexity in grammars*. Oxford University Press on Demand, 2004.
- J. M. Henderson. Visual attention and eye movement control during reading and picture viewing. In *Eye movements and visual cognition*, pages 260–283. Springer, 1992.
- J. M. Henderson, W. Choi, M. W. Lowder, and F. Ferreira. Language structure in the brain: A fixation-related fmri study of syntactic surprisal in reading. *Neuroimage*, 132:293–300, 2016.
- D. L. Hintzman. “schema abstraction” in a multiple-trace memory model. *Psychological review*, 93(4):411, 1986.
- J.-M. Hombert, J. J. Ohala, and W. G. Ewan. Phonetic explanations for the development of tones. *Language*, pages 37–58, 1979.
- S. Hoppe-Graff, T. Herrmann, P. Winterhoff-Spurk, and R. Mangold. Speech and situation: A general model for the process of speech production. In *Language and social situations*, pages 81–95. Springer, 1985.
- A. W. Inhoff and K. Rayner. Parafoveal word processing during eye fixations in reading: Effects of word frequency. *Perception & Psychophysics*, 40(6):431–439, 1986.
- T. F. Jaeger. Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology*, 61(1):23–62, 2010. doi: 10.1016/j.cogpsych.2010.02.002.
- W. James. *The principles of psychology, Vol I*. Henry Holt and Co, 1890. doi: 10.1037/10538-000.
- G. W. Jenkins, L. K. Samuelson, J. R. Smith, and J. P. Spencer. Non-bayesian noun generalization in 3-to 5-year-old children: Probing the role of prior knowledge in the suspicious coincidence effect. *Cognitive science*, 39(2):268–306, 2015.
- G. W. Jenkins, L. K. Samuelson, W. Penny, and J. P. Spencer. Learning words in space and time: Contrasting models of the suspicious coincidence effect. *Cognition*, 210:104576, 2021.
- A. Jesse. Sentence context guides phonetic retuning to speaker idiosyncrasies. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(1):184, 2021.
- A. Jesse and J. M. McQueen. Positional effects in the lexical retuning of speech perception. *Psychonomic bulletin & review*, 18(5):943–950, 2011.
- K. Johnson. Resonance in an exemplar-based lexicon: The emergence of social identity and phonology. *Journal of phonetics*, 34(4):485–499, 2006.

- K. E. Johnson and C. B. Mervis. Effects of varying levels of expertise on the basic level of categorization. *Journal of Experimental Psychology: General*, 126(3):248, 1997.
- G. Kempen and E. Hoenkamp. An incremental procedural grammar for sentence formulation. *Cognitive science*, 11(2):201–258, 1987.
- J. Kimball. Seven principles of surface structure parsing in natural language. *Cognition*, 2(1):15–47, 1973.
- T. Kraljic and A. G. Samuel. Generalization in perceptual learning for speech. *Psychonomic bulletin & review*, 13(2):262–268, 2006.
- A. Kroch and C. Small. Grammatical ideology and its effect on speech. *Linguistic variation: Models and methods*, 45755, 1978.
- J. K. Kruschke. Alcové: an exemplar-based connectionist model of category learning. *Psychological review*, 99(1):22, 1992.
- J. K. Kruschke. Models of categorization. *The Cambridge handbook of computational psychology*, pages 267–301, 2008.
- K. Lamberts. Information-accumulation theory of speeded categorization. *Psychological review*, 107(2):227, 2000.
- B. Landau, L. B. Smith, and S. S. Jones. The importance of shape in early lexical learning. *Cognitive development*, 3(3):299–321, 1988.
- J. S. Lappin and H. H. Bell. Perceptual differentiation of sequential visual patterns. *Attention, Perception, & Psychophysics*, 12(2):129–134, 1972.
- K. S. Lashley. *The problem of serial order in behavior*, volume 21. Bobbs-Merrill, 1951.
- A. LaTourrette and S. R. Waxman. A little labeling goes a long way: Semi-supervised learning in infancy. *Developmental science*, 22(1):e12736, 2019.
- C. A. Lawson. When diverse evidence is (and isn't) inductively privileged: The influence of evidence presentation on children's and adults' generalization. In *Proceedings of the Annual Conference of the Cognitive Science Society*, volume 36, pages 2537–2542, 2014a.
- C. A. Lawson. Three-year-olds obey the sample size principle of induction: The influence of evidence presentation and sample size disparity on young children's generalizations. *Journal of experimental child psychology*, 123:147–154, 2014b.
- C. A. Lawson. The influence of task dynamics on inductive generalizations: How sequential and simultaneous presentation of evidence impacts the strength and scope of property projections. *Journal of Cognition and Development*, 18(4):493–513, 2017.



- W. J. Levelt. Accessing words in speech production: Stages, processes and representations. *Cognition*, 42(1-3):1–22, 1992.
- W. J. Levelt. *Speaking: From intention to articulation*, volume 1. MIT press, 1993.
- W. J. Levelt, H. Schriefers, D. Vorberg, A. S. Meyer, and et al. The time course of lexical access in speech production: A study of picture naming. *Psychological Review*, 98(1): 122–142, 1991. doi: 10.1037/0033-295x.98.1.122.
- W. J. Levelt, A. Roelofs, and A. S. Meyer. A theory of lexical access in speech production. *Behavioral and brain sciences*, 22(1):1–38, 1999.
- R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177, 2008.
- R. P. Levy and T. F. Jaeger. Speakers optimize information density through syntactic reduction. In *Advances in neural information processing systems*, pages 849–856, 2007.
- M. L. Lewis and M. C. Frank. Still suspicious: the suspicious-coincidence effect revisited. *Psychological science*, 29(12):2039–2047, 2018.
- A. M. Liberman, K. S. Harris, H. S. Hoffman, and B. C. Griffith. The discrimination of speech sounds within and across phoneme boundaries. *Journal of experimental psychology*, 54(5):358, 1957.
- C. Lignos. *Modeling words in the mind*. PhD thesis, University of Pennsylvania, 2013.
- D. Linares and J. López i Moliner. quickpsy: An r package to fit psychometric functions for multiple groups. *The R Journal*, 2016, vol. 8, num. 1, p. 122-131, 2016.
- L. P. Lipsitt. Simultaneous and successive discrimination learning in children. *Child Development*, 32(2):337, 1961. doi: 10.2307/1125948.
- L. Liu and T. F. Jaeger. Inferring causes during speech perception. *Cognition*, 174:55–70, 2018.
- B. Lohse, J. A. Hawkins, and T. Wasow. Domain minimization in english verb-particle constructions. *Language*, pages 238–261, 2004.
- G. Lupyan, S. L. Thompson-Schill, and D. Swingley. Conceptual penetration of visual processing. *Psychological Science*, 21(5):682–691, 2010. doi: 10.1177/0956797610366099.
- B. MacWhinney. *The CHILDES project: The database*, volume 2. Psychology Press, 2000.
- K. Mahowald, E. Fedorenko, S. T. Piantadosi, and E. Gibson. Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, 126(2):313–318, 2013.
- C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational*

- Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL <http://www.aclweb.org/anthology/P/P14/P14-5010>.
- G. F. Marcus. Negative evidence in language acquisition. *Cognition*, 46(1):53–85, 1993.
- M. P. Marcus, M. A. Marcinkiewicz, and B. Santorini. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330, 1993.
- E. M. Markman. *Categorization and naming in children: Problems of induction*. Mit Press, 1989.
- E. M. Markman. Constraints children place on word meanings. *Cognitive Science*, 14(1):57–77, 1990.
- E. M. Markman and G. F. Wachtel. Children’s use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, 20(2):121–157, 1988.
- D. Marr. *Vision: A computational investigation into the human representation and processing of visual information*. The MIT Press, 1982.
- W. Marslen-Wilson and L. K. Tyler. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71, 1980.
- W. D. Marslen-Wilson and A. Welsh. Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, 10(1):29–63, 1978.
- A. Matsuo, S. Kita, G. C. Wood, and L. Naigles. Children’s use of morphosyntax and argument structure to infer the meaning of novel transitive and intransitive verbs. In *Transitivity and Valency Alternations*, pages 341–356. De Gruyter Mouton, 2016.
- H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, and D. Bates. Balancing type i error and power in linear mixed models. *Journal of memory and language*, 94:305–315, 2017.
- L. Maurits, D. Navarro, and A. Perfors. Why are some word orders more common than others? a uniform information density account. In *Advances in neural information processing systems*, pages 1585–1593, 2010.
- J. L. McClelland and J. L. Elman. The trace model of speech perception. *Cognitive psychology*, 18(1):1–86, 1986.
- H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
- B. McMurray, M. K. Tanenhaus, and R. N. Aslin. Within-category voT affects recovery from “lexical” garden-paths: Evidence against phoneme-level inhibition. *Journal of memory and language*, 60(1):65–91, 2009.

- J. M. McQueen, D. Norris, and A. Cutler. The dynamic nature of speech perception. *Language and speech*, 49(1):101–112, 2006.
- B. J. Meagher, P. F. Carvalho, R. L. Goldstone, and R. M. Nosofsky. Organized simultaneous displays facilitate learning of complex natural science categories. *Psychonomic Bulletin & Review*, 24(6):1987–1994, 2017.
- D. L. Medin and M. M. Schaffer. Context theory of classification learning. *Psychological review*, 85(3):207, 1978.
- D. L. Medin, W. D. Wattenmaker, and S. E. Hampson. Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, 19(2):242–279, 1987.
- T. N. Medina, J. Snedeker, J. C. Trueswell, and L. R. Gleitman. How words can and cannot be learned by observation. *Proceedings of the National Academy of Sciences*, 108(22):9014–9019, 2011.
- W. E. Merriman, L. L. Bowman, and B. MacWhinney. The mutual exclusivity bias in children’s word learning. *Monographs of the society for research in child development*, pages i–129, 1989.
- C. B. Mervis. Early lexical development: The contributions of mother and child. *Origins of cognitive skills*, pages 339–370, 1984.
- R. D. Morey. Confidence intervals from normalized data: A correction to Cousineau (2005). *reason*, 4(2):61–64, 2008.
- C. M. Munson. *Perceptual learning in speech reveals pathways of processing*. PhD thesis, The University of Iowa, 2011.
- G. Murphy. *The big book of concepts*. MIT press, 2004.
- W. S. Murray and K. I. Forster. Serial mechanisms in lexical access: the rank hypothesis. *Psychological Review*, 111(3):721, 2004.
- L. Naigles. Children use syntax to learn verb meanings. *Journal of child language*, 17(2):357–374, 1990.
- L. R. Naigles and P. Terrazas. Motion-verb generalizations in english and spanish: Influences of language and syntax. *Psychological Science*, 9(5):363–369, 1998.
- D. Norris, J. M. McQueen, and A. Cutler. Perceptual learning in speech. *Cognitive psychology*, 47(2):204–238, 2003.
- T. Omata and K. Holyoak. The role of comparison processes in the induction of schemas for design styles. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 34, 2012.

- K. R. Paap, J. E. McDonald, R. W. Schvaneveldt, and R. W. Noel. Frequency and pronounceability in visually presented naming and lexical decision tasks. In *Attention and performance 12: The psychology of reading*, pages 221–243. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc, 1987.
- T. Pechmann. Incremental speech production and referential overspecification. *Linguistics*, 27(1):89–110, 1989.
- F. Pellegrino, C. Coupé, and E. Marsico. Across-language perspective on speech information rate. *Language*, 87(3):539–558, 2011.
- S. T. Piantadosi, H. Tily, and E. Gibson. Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, 108(9):3526–3529, 2011.
- J. Pierrehumbert. Lenition and contrast. *Frequency and the emergence of linguistic structure*, 45:137, 2001.
- S. Pinker. *Learnability and cognition: The acquisition of argument structure*. MIT press, 1989.
- B. Pomiechowska and T. Gliga. Lexical acquisition through category matching: 12-month-old infants associate words to visual categories. *Psychological science*, 30(2):288–299, 2019.
- B. R. Postle. The cognitive neuroscience of visual short-term memory. *Current opinion in behavioral sciences*, 1:40–46, 2015.
- W. V. O. Quine. *Word and object*. MIT press, 1960.
- E. N. Ransom. Definiteness, animacy, and np ordering. In *Annual Meeting of the Berkeley Linguistics Society*, volume 3, pages 418–429, 1977.
- K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998.
- K. Rayner and S. A. Duffy. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & cognition*, 14(3):191–201, 1986.
- K. Rayner, J. Ashby, A. Pollatsek, and E. D. Reichle. The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model. *Journal of Experimental Psychology: Human Perception and Performance*, 30(4):720, 2004.
- T. Regier. The emergence of words: Attentional learning in form and meaning. *Cognitive science*, 29(6):819–865, 2005.
- B. Rehder and A. B. Hoffman. Eyetracking and selective attention in category learning. *Cognitive psychology*, 51(1):1–41, 2005a.

- B. Rehder and A. B. Hoffman. Thirty-something categorization results explained: selective attention, eyetracking, and models of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(5):811, 2005b.
- E. Reinisch and L. L. Holt. Lexically guided phonetic retuning of foreign-accented speech and its generalization. *Journal of Experimental Psychology: Human Perception and Performance*, 40(2):539, 2014.
- R. A. Rescorla. Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 6(3):207, 1980.
- A. Roelofs. The weaver model of word-form encoding in speech production. *Cognition*, 64(3):249–284, 1997.
- A. Roelofs. Rightward incrementality in encoding simple phrasal forms in speech production: Verb–particle combinations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(4):904, 1998.
- A. Roelofs. Weaver++ and other computational models of lemma retrieval and word-form encoding. In *Aspects of language production*, pages 83–126. Psychology Press, 2013.
- A. P. A. Roelofs. *Lemma retrieval in speaking: A theory, computer simulations, and empirical data*. Nijmegen: NICI, Nijmeegs Instituut voor Cognitie en Informatie, 1992.
- D. Roland, J. L. Elman, and V. S. Ferreira. Why is that? structural prediction and ambiguity resolution in a very large corpus of english sentences. *Cognition*, 98(3):245–272, 2006.
- E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive psychology*, 8(3):382–439, 1976.
- D. K. Roy and A. P. Pentland. Learning words from sights and sounds: A computational model. *Cognitive science*, 26(1):113–146, 2002.
- D. E. Rumelhart and J. L. McClelland. On learning the past tenses of english verbs. Technical report, CALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR COGNITIVE SCIENCE, 1985.
- J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928, 1996.
- J. R. Saffran, E. K. Johnson, R. N. Aslin, and E. L. Newport. Statistical learning of tone sequences by human infants and adults. *Cognition*, 70(1):27–52, 1999.
- D. Sagi. Perceptual learning in vision research. *Vision research*, 51(13):1552–1566, 2011.
- A. G. Samuel and T. Kraljic. Perceptual learning for speech. *Attention, Perception, & Psychophysics*, 71(6):1207–1218, 2009.

- H. Schriefers, A. S. Meyer, and W. J. Levelt. Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of memory and language*, 29(1):86–102, 1990.
- K. D. Schuler, J. Kodner, and S. Caplan. Abstractions are good for brains and machines: A commentary on ambridge (2020). *First Language*, 40(5-6):631–635, 2020.
- P. G. Schyns and F. Gosselin. A natural bias for basic-level object categorizations. *Journal of Vision*, 2(7):407–407, 2002.
- C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 1948.
- R. N. Shepard, C. I. Hovland, and H. M. Jenkins. Learning and memorization of classifications. *Psychological monographs: General and applied*, 75(13):1, 1961.
- R. M. Shifflin, G. T. Gardner, and D. H. Allmeyer. On the degree of attention and capacity limitations in visual processing. *Attention, Perception, & Psychophysics*, 14(2):231–236, 1973.
- H. A. Simon. *The sciences of the artificial*. MIT press, 1996.
- J. M. Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61(1):39–91, 1996.
- D. Smedt. *Incremental sentence generation: A computer model of grammatical encoding*. Nijmegen: Nijmeegs Institute for Cognition Research and Information Technology, 1990.
- D. Smedt. Parallelism in incremental sentence generation. *Parallel natural language processing*, pages 421–447, 1994.
- D. J. Smith and J. P. Minda. Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(1):3, 2000.
- E. E. Smith and D. L. Medin. *Categories and concepts*. Harvard University Press Cambridge, MA, 1981.
- E. E. Smith and D. L. Medin. *Categories and concepts*. Harvard University Press, 2013.
- L. Smith and C. Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008.
- L. B. Smith. Perceptual development and category generalization. *Child Development*, pages 705–715, 1979.
- L. B. Smith, C. Yu, H. Yoshida, and C. M. Fausey. Contributions of head-mounted cameras to studying the visual environments of infants and young children. *Journal of Cognition and Development*, 16(3):407–419, 2015.

- N. J. Smith and R. Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319, 2013.
- J. Snedeker and L. Gleitman. Why it is hard to label our concepts. *Weaving a lexicon*, 257294, 2004.
- J. Y. Son, L. B. Smith, and R. L. Goldstone. Connecting instances to promote children’s relational reasoning. *Journal of experimental child psychology*, 108(2):260–277, 2011.
- J. P. Spencer, S. Perone, L. B. Smith, and L. K. Samuelson. Learning words in space and time probing the mechanisms behind the suspicious-coincidence effect. *Psychological science*, 22(8):1049–1057, 2011.
- L. M. Stallings, M. C. MacDonald, and P. G. O’Seaghdha. Phrasal ordering constraints in sentence production: Phrase length and verb disposition in heavy-np shift. *Journal of Memory and Language*, 39(3):392–417, 1998.
- A. Staub. The effect of lexical predictability on distributions of eye fixation durations. *Psychonomic Bulletin & Review*, 18(2):371–376, 2011.
- L. Steels and T. Belpaeme. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and Brain Sciences*, 28(4):469–489, 2005. doi: 10.1017/s0140525x05000087.
- J. S. Stevens, L. R. Gleitman, J. C. Trueswell, and C. Yang. The pursuit of word meanings. *Cognitive Science*, 41:638–676, 2016. doi: 10.1111/cogs.12416.
- J. Stuhlman. *Georgia O’Keeffe: Circling Around Abstraction*. Hudson Hills Press, 2007.
- D. Temperley. Ambiguity avoidance in english relative clauses. *Language*, 79(3):464–484, 2003.
- S. Thim. *Phrasal verbs: The English verb-particle construction and its history*, volume 78. Walter de Gruyter, 2012.
- J. C. Toscano, B. McMurray, J. Dennhardt, and S. J. Luck. Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological science*, 21(10):1532–1540, 2010.
- J. C. Toscano, N. D. Anderson, M. Fabiani, G. Gratton, and S. M. Garnsey. The time-course of cortical responses to speech revealed by fast optical imaging. *Brain and Language*, 184: 32–42, 2018.
- M. J. Traxler and M. J. Pickering. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35(3):454–475, 1996.

- J. C. Trueswell, M. K. Tanenhaus, and C. Kello. Verb-specific constraints in sentence processing: Separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19(3):528, 1993.
- J. C. Trueswell, I. Sekerina, N. M. Hill, and M. L. Logrip. The kindergarten-path effect: Studying on-line sentence processing in young children. *Cognition*, 73(2):89–134, 1999.
- J. C. Trueswell, T. N. Medina, A. Hafri, and L. R. Gleitman. Propose but verify: Fast mapping meets cross-situational word learning. *Cognitive psychology*, 66(1):126–156, 2013.
- A. Tyler and W. Nagy. The acquisition of english derivational morphology. *Journal of memory and language*, 28(6):649–667, 1989.
- M. van Heugten and E. K. Johnson. Learning to contend with accents in infancy: Benefits of brief speaker exposure. *Journal of Experimental Psychology: General*, 143(1):340, 2014.
- T. von der Malsburg, R. Kliegl, and S. Vasishth. Determinants of scanpath regularity in reading. *Cognitive science*, 39(7):1675–1703, 2015.
- F. H. Wang and J. Trueswell. Being suspicious of suspicious coincidences: The case of learning subordinate word meanings, 2017. Paper presented at 42nd Boston University Conference on Language Development, Boston, MA.
- F. H. Wang and J. C. Trueswell. Spotting dalmatians: Children’s ability to discover subordinate-level word meanings cross-situationally. *Cognitive psychology*, 114:101226, 2019.
- T. Wasow and J. Arnold. Post-verbal constituent ordering in english. *Determinants of grammatical variation in English*, pages 119–54, 2003.
- T. Wasow, A. Perfors, and D. Beaver. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282, 2005.
- S. R. Waxman. Links between object categorization and naming. *Early category and concept development: Making sense of the blooming, buzzing confusion*, pages 213–241, 2003.
- S. R. Waxman and D. B. Markow. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302, 1995.
- J. B. Wells, M. H. Christiansen, D. S. Race, D. J. Acheson, and M. C. MacDonald. Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, 58(2):250–271, 2009.
- D. Whitney and A. Yamanashi Leib. Ensemble perception. *Annual review of psychology*, 69:105–129, 2018.
- R. M. Willems, S. L. Frank, A. D. Nijhof, P. Hagoort, and A. Van den Bosch. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516, 2015.



- J. P. Williams and M. D. Ackerman. Simultaneous and successive discrimination of similar letters. *Journal of Educational Psychology*, 62(2):132, 1971.
- K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott. Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, 79(7):2064–2072, 2017.
- W. Wundt. The psychology of the sentence. *Language and psychology: Historical aspects of psycholinguistics*, pages 9–32, 1904.
- F. Xu. The role of language in acquiring object kind concepts in infancy. *Cognition*, 85(3): 223–250, 2002. doi: 10.1016/s0010-0277(02)00109-9.
- F. Xu and J. B. Tenenbaum. Sensitivity to sampling in bayesian word learning. *Developmental science*, 10(3):288–297, 2007a.
- F. Xu and J. B. Tenenbaum. Word learning as bayesian inference. *Psychological Review*, 114(2):245–272, 2007b. doi: 10.1037/0033-295x.114.2.245.
- C. Yang. *Knowledge and learning in natural language*. Oxford University Press on Demand, 2002.
- C. Yang. *The price of linguistic productivity: How children learn to break the rules of language*. MIT press, 2016.
- C. Yu. A statistical associative account of vocabulary growth in early word learning. *Language learning and Development*, 4(1):32–62, 2008.
- C. Yu and L. B. Smith. Rapid word learning under uncertainty via cross-situational statistics. *Psychological Science*, 18(5):414–420, 2007. doi: 10.1111/j.1467-9280.2007.01915.x.
- G. Zellou and D. Dahan. Listeners maintain phonological uncertainty over time and across words: The case of vowel nasality in english. *Journal of Phonetics*, 76:100910, 2019.