UNIVERSITY OF CALIFORNIA, SAN DIEGO

Acquiring latent linguistic structure using computational models

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Philosophy

in

Linguistics

by

Gabriel R. Doyle

Committee in charge:

Professor Roger Levy, Chair
Professor Eric Bakovic
Professor David Barner
Professor Charles Elkan
Professor Andrew Kehler

2014

The Dissertation of Gabriel R. Doyle is approved and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California, San Diego

2014

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

There's a part at the end of Norton Juster's classic "The Phantom Tollbooth" where the hero has returned from a difficult quest and asks his patrons about a secret that they could not tell him before he finished the quest. His patrons, representing the realms of language and mathematics, reply off-handedly that the task was impossible – "but if we'd told you then, you might not have gone – and, as you've discovered, so many things are possible just as long as you don't know they're impossible."

That line stuck with me long before I actually understood it. I think I do now, thanks most prominently to three people. The first two are my parents, Karen & Mike, who always treated it as the most natural thing in the world that someone from a family with a spotty academic record should want to get a doctorate, and did anything they could to help get me there (or wherever else I would have hoped to end up). Their endless support of and belief in me led to this dissertation.

The other person who's hammered home Juster's point has been my advisor and committee chair, Roger Levy, who always manages to make it seem that the work you're trying to do is well within your grasp, even if it isn't, and convinces you to go a little bit further, even if that's impossible. I couldn't have ended up in a better place or with a better advisor.

I owe deep thanks to the rest of my committee as well: Eric Baković, Dave Barner, Charles Elkan, and Andy Kehler – as well as Rachel Mayberry, who was on my original committee before the topic shifted – who never failed to provide ideas, inspirations, and helpful inquisitions along a very winding research path. They were contagiously enthusiastic in discussions even when I was worn out, and their ability to remind me of the philosophical forest when I'd get stuck on trees was essential.

The members, past and present, of the Computational Psycholinguistics Lab are also a big part of this dissertation, through many, many discussions of language and math

and mass transit design, plus the occasional coffee run or mountain hike. They include: Emily Morgan, Mark Myslín, Victoria Fossum, Rebecca Colavin, Bozena Pajak, Klinton Bicknell, Albert Park, and Nathaniel Smith. This work (and a lot of my non-dissertation work) has also been helped by our excellent undergraduate lab assistants: Tiffany Rzvani, Kaylie Fernald, Mike Brooks, Melodie Yen, Cody McCormack, Bill Presant, Wednesday Bushong, Bonnie Chinh, Ksenia Kozhukhovskaya, and Agatha Ventura.

Lastly, there are the people who have contributed so much to my last few years without fitting into one of the neat little categories above. Thanks for making my time in San Diego so good. They include: R Mata, Grant Loomis, Kate Davidson, Dan Michel, Zoe Ziliak Michel, Hope Morgan, Boyoung Kim, Maria Anjarwala, Becky Martin, Holly Morgan, Aditya Menon, and Dirk Hovy.

Chapter 2, in full, is an exact copy of the material as it appears in Doyle, Bicknell, and Levy (2014) [Nonparametric learning of phonological constraints in Optimality Theory. In K. Toutanova and H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (pp. 1094-1103). Baltimore: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is an exact copy of the material as it appears in Doyle and Levy (2013) [Combining multiple information types in Bayesian word segmentation. In L. Vanderwende, H. Daumé III, and K. Kirchhoff (Eds.), Proceedings of the 2013 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (pp. 117-126). Atlanta: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper.

Chapter 5, in full, is an exact copy of the material as it appears in Doyle and Elkan (2009) [Accounting for burstiness in topic models. In L. Bottou and M. Littman (Eds.),

Proceedings of the 26th International Conference on Machine Learning (pp. 281-288). Montreal: Omnipress.] The dissertation author was the primary investigator and author of this paper.

VITA

2005	A.B. in Mathematics Princeton University
2011	M.A. in Linguistics University of California, San Diego
2014	Ph.D. in Linguistics University of California, San Diego

ABSTRACT OF THE DISSERTATION

Acquiring latent linguistic structure using computational models

by

Gabriel R. Doyle

Doctor of Philosophy in Linguistics

University of California, San Diego, 2014

Professor Roger Levy, Chair

Language contains a great deal of latent structure, which shapes the produced linguistic forms but does not directly appear in them. Identifying this latent structure is both a core goal of linguistic theories and the task confronting a child acquiring language. This dissertation investigates the acquisition of latent linguistic structure using computational models over a range of linguistic phenomena. These models share two central features. First, they rely as much as possible on observed language data to determine the latent structure. This supports emergentist accounts of acquisition, where a learner relies primarily on cognitively-general learning methods to extract the structure

from the data, as opposed to innatist accounts, which rely primarily on substantial innate foreknowledge of the linguistic structure. Second, the models combine general Bayesian analysis principles with appropriate representations of the linguistic structure for each problem to maximize the information they obtain from the language data.

Four models are proposed in this dissertation. Two examine the source of phonological constraints in Optimality Theory, and argues that they can be acquired from language data with little to no innate phonological structure, contrary to the standard, innatist position. The third model addresses early word segmentation from speech, and shows that a Bayesian model for incorporating multiple cues outperforms a single-cue model on segmentation, as well as providing a possible explanation for human segmentation behavior. The final model moves into applications, and shows that accurately representing the bursty behavior of the language data improves the fit of a topic model.

Chapter 1

Introduction

Language contains a great deal of latent, or hidden, structure – structure that shapes the form of the produced language, but that does not directly appear in that production. This includes the syntactic structure that informs the order of the words, the phonological constraints that shape their phonemes, and the phonetic categorization that drives phonemic recognition. Although these structures are not immediately visible in the language data, their presence and their form can be inferred through patterns in the data.

One of the central goals of linguistic theory is to define these latent structures, whether in the formal/applied perspectives that extract the linguistic information in an immediately interpretable way to the psycholinguistic/cognitive perspectives that represent the mental framework and processing of language. Furthermore, acquiring latent structure is the core of language acquisition, as fluency in a language amounts to fluency over its latent structure. As such, there is always a need to improve our ability to identify and represent latent structure, both from a formal and a cognitive perspective.

This dissertation investigates the acquisition of latent linguistic structure using computational models, across a variety of linguistic structures and covering both applications and psycholinguistic facets. Chapters 2 and 3 build models for the acquisition of phonological constraints from data, reducing the amount of phonological structure

that is assumed to be innate. Chapter 4 examines the problem of word segmentation and proposes a model to incorporate multiple cues (syllable identity and stress) in order to both improve segmentation performance and fidelity to human behavior. Chapter 5 looks at applications, improving the performance of topic models by more appropriately modeling burstiness in both linguistic and non-linguistic data.

The models constructed in this work address the problem of unsupervised learning from positive-only data, the kind of learning problem that confronts a human learner of a language. While the models address different linguistic learning problems and differ in their specific results, they share two major goals.

The first is to rely as much as possible on the observed language data in determining the latent structure, rather than relying primarily on innate foreknowledge of the latent structure. This is an *emergentist* or *empiricist* stance on language acquisition, where the learner relies on an outline of the linguistic structure, based in cognitively- or statistically-general methods, and relies primarily on the language data to build the latent structure. This is in comparison to an *innatist* or *nativist* stance, where the learner relies on innate knowledge, specific to language, to supply most of the linguistic structure, with the language data used only to identify the particular structure of the language being learned.

The second goal is to use more appropriate model structures for the learning problems to improve the models' ability to extract information from the observed language data. Previous approaches to these learning problems have overlooked important information because it did not fit the generative models used to represent the data. The models in this dissertation use statistically-general approaches to capture new aspects of the linguistic information, which result in both significant improvements in the quality of the learned structure and the ability to learn components of the linguistic structure

that could not be learned before.[1] In addition, these models are designed to work with multiple possible model structures (e.g., different phonological constraint grammars in Ch. 2), to open new avenues for testing the practical effects of different theoretical choices.

In this introduction, I will discuss the motivation for computational models of language, both for cognitive purposes and for applications. I will go deeper into the cognitive side and discuss the use of computational models in rational analysis, including addressing what problems can be tackled by computational models in general, and what problems are tackled by the specific models in this dissertation. The computational models here focus on problems of acquisition, and so I will also briefly discuss issues of language acquisition, especially as it relates to questions of innateness of structure and data-driven learning.

1.1 Computational Models

This dissertation covers four computational models of language – statistical models that use Bayesian analysis to infer likely latent structures for some observed language data. These are also referred to as rational models, as they implement the optimal behaviors for processing information in the presence of uncertainty (Anderson, 1990). Rationality, in this context, means behaving in normatively optimal ways. The central idea of rational analysis is that humans are approximately rational when performing our core cognitive tasks, although we labor under various restrictions on memory, perception, and computational time.

Two examples of these core tasks are visual perception and motor control (Chater & Oaksford, 2008). Visual perception was a seminal area for developing rational models

[1] Of course, there is still much useful information in the input that the models proposed in this dissertation do not capture.

(Marr, 1982), and probabilistic computational models have been successful in predicting behavior in tasks such as shape identification (Freeman, 1994; Blake, Bülthoff, & Sheinberg, 1993), object and boundary recognition (Tu, Chen, Yuille, & Zhu, 2005), and motion illusions (Weiss, Simoncelli, & Adelson, 2002). Within motor control, Bayesian probabilistic models have given insight into movement trajectories (Körding & Wolpert, 2004) and timing (Miyazaki, Nozaki, & Nakajima, 2005).

While it may not be clear that language is a core cognitive task, it too has been found to fit well with the rational analysis framework. The successes of computational linguistics and natural language processing have shown that computational models are able to extract useful structure from language data, especially through the use of Bayesian probabilistic methods. The ability of news-aggregating websites to group similar articles and of speech-recognition software to understand its users are two now-mundane examples of how computational models can identify the structure of language data. Rational analysis also fits with human behavior; difficulties in online sentence comprehension, for instance, often appear with low-probability continuations (Hale, 2001; Levy, 2008). Behavior comporting with rational analysis also appears in word segmentation acquisition (M. Frank, Goldwater, Griffiths, & Tenenbaum, 2010), eye-movements during reading (Bicknell, 2011), and many other linguistic problems.

However, there are two potential concerns in modeling human behavior as rational. First, implementing such models often relies on sophisticated and computationally-intense mathematical analyses to infer the latent structure. Second, humans often behave seemingly irrationally in their conscious decision-making, especially when probabilities and uncertainty are involved (Kahneman & Tversky, 2000; Shafir, Smith, & Osherson, 1990).

These concerns do not undermine the computational modeling approach, but rather suggest that the approach is appropriate for some goals and inappropriate for

others. Rational analysis of *conscious* probabilistic reasoning may be inappropriate although some of the irrational behaviors may result from rational behavior in one context being applied outside that context (Weiss et al., 2002; Sher & McKenzie, 2008). But, as noted above, the less conscious and more central processes do reflect rational behavior, and the language processing problems covered in this dissertation largely fall within this realm. Computational power also need not be a limiting factor, depending on the problem. Humans process speech, for example, by performing an approximation of mathematically-complex Fourier analysis through physical means (Stevens, 1998). Limitations on computational resources can also be introduced into a model, sometimes improving the model's fit to human performance (M. Frank et al., 2010; Phillips & Pearl, 2012).

The issue of identifying what aspects of an information processing problem a model addresses is often phrased in terms of Marr (1982)'s levels of analysis. Marr proposed three levels: implementational, algorithmic, and computational. These levels represent increasing abstractness of the analysis. Implementational-level analysis attempts as much as possible to mimic the exact infrastructure of the brain when trying to solve a problem. Algorithmic-level analysis abstracts slightly from this, proposing an algorithm to solve the problem that fits with the brain's commonly-accepted limitations, but without specifying how the algorithm would be implemented in the brain. Computational-level analysis abstracts even further, and focuses on how a problem could be solved mathematically, without committing to an algorithm or an implementation. Depending on the purpose of the analysis, computational-level analysis may or may not take into account the brain's limitations. The first two levels focus on the brain; the last focuses on the problem.

The models in this dissertation focus on computational-level analysis.[2] Computational-

[2]Rational analysis is usually framed at this level (in fact, Anderson (1990), introducing rational analysis,

level solutions define upper limits to the solvability of a problem, allowing a researcher to identify what information is available and extractable given a certain set of assumptions. These assumptions include the data available to be learned from, the structure that is assumed ahead of time, and the probabilistic functions that assess the likelihood of different structures. This "upper limit" comes with a caveat; for many computational models, analytic solutions are intractable, and many, including those in this dissertation, are solved through approximate means. In practice, this means that human performance may surpass model performance. This affects what conclusions are appropriate to draw from computational models, discussed in Sect. 1.3. To better understand the purpose of computational models of language, we detour briefly into the problems faced by a language learner.

1.2 The learning problem

The learning tasks being modeled in this dissertation are all unsupervised learning problems, using positive-only data. *Unsupervised* learning means that none of the data is labeled for the latent structure the model is trying to learn. *Positive-only* data means that there are no negative examples (or at least none labeled as negative). These two traits are also true of the language data that a child receives; most structures are never explicitly labeled, and if the child encounters an ungrammatical utterance, it is unlikely to be marked as such.

The unsupervised, positive-only setting is a difficult one for extracting latent structure to the point that these conditions are core components of the *poverty of the stimulus* argument in language acquisition (Chomsky, 1965). The poverty of the stimulus idea argues that limitations on both the quantity and quality of language data children

referred to this as the *rational* level). Danks (2008) argues that this definition of the analysis levels are inadequate, and that a more nuanced division is not only more appropriate, but allows for rational analysis at a larger range of levels.

encounter means that language acquisition can only proceed if humans have substantial innate knowledge of the structure of language. This innate knowledge forms the core of Universal Grammar.

The poverty of the stimulus is based on three claims:

1. The observed language is noisy, containing some level of accidental or erroneous productions that are not marked as such.

2. The data size is very small, lacking sufficient data to construct a mature grammar.

3. The data contains few or no labeled negative examples to inform a learner of what is ungrammatical.

The position that humans must have considerable innate knowledge, specific to language, in order to acquire language is an *innatist* or *nativist* position. The opposing stance, that little to no innate, language-specific knowledge is necessary, is called an *emergentist* or *empiricist* position. These are, of course, relative positions, as few people would argue for a maximally innatist or emergentist position. Instead, these arise mostly in the context of specific problems, arguing that more of the structure should be innate or learned.

A major component of the studies in this dissertation is arguing for more emergentist positions within language acquisition, by showing that more information can be extracted from the language data than the poverty of the stimulus idea suggests. The studies touch on all three facets of the poverty argument, with the first and third facets being most prominently addressed. The models use real or realistic language data in their experiments, dealing with the inherent noise and occasional inappropriate productions that real language data provides. The models also all lack negative examples, and overcome this by inferring implicit negative evidence. Implicit negative evidence comes from

the realization (motivated in this work by Bayesian analysis) that a form that appears far less often than expected under one's current beliefs about a language's structure is almost certainly being avoided (e.g., is ungrammatical).

The models in the following chapters argue that learning with cognitively-general learning principles, such as Bayesian analysis, can be sufficient to overcome the poverty of the stimulus. The most direct example of this is in the phonological constraint acquisition chapters, which argue for a strong emergentist stance, that constraints can be learned from distributional data instead of relying on innate prespecification or derivation from articulatory and perceptual difficulty (a relatively weaker emergentist stance). The multiple-cue word segmentation model also argues for an emergentist view on stress bias acquisition, as opposed to the innate Metrical Segmentation Strategy. The topic model does not directly address matters of acquisition.

1.3 Assessing Computational Models

The previous section examined the general problems of language acquisition; this section expands on this by examining how computational models can be used to gain insight into acquisition and other problems.

The biggest advantage from computational models is that they allow a researcher to test hypotheses about what is going in the human mind by manipulating such aspects as what structure is innate versus learned, or what underlying principles the learner is bound by in its learning. The answers to such questions are of significant theoretical and practical import, but there is no way of directly manipulating these within the human mind. Using models allows the researcher test different structures to determine which are reasonable fits for the behaviors we see from humans. Model behavior can be assessed in three ways:

1. generic performance
2. human performance
3. application performance

In addition, the performance assessments can be based on a single model, to argue that a certain set of assumptions are sufficient for learning, or on a comparison of models, to argue that one is a better explanation than others. To illustrate the purpose of each of these assessments, I will look at existing studies, and show where the models in this dissertation fall in this spectrum.

Generic Performance The first type of assessment is to look at what I will loosely call *generic* performance. Generic performance is simply looking at measures such as data likelihood to determine whether the computational model is extracting sufficient information to learn the structure of the linguistic data. This assessment, when performed on a single computational model, has as its goal determining what information is available, and whether, under the stated assumptions about the learning methods and model structure, the model is able to learn what we expect it to learn. This is usually the first step in assessing a model.

Many Optimality Theory learning algorithms focus on this assessment, including Recursive Constraint Demotion (Tesar & Smolensky, 2000) and the Gradual Learning Algorithm (Boersma & Hayes, 2001). The GLA is a model for the acquisition of constraint weights (but not constraint identities, which I will examine in Ch. 2 and 3) in Optimality Theory. It uses error-driven learning; each time it makes an erroneous prediction, it increases the weights on constraints that favor the correct output and decreases the weights on constraints that favor the erroneous prediction. Boersma and Hayes focus on showing that such a method can learn appropriate constraint weights

for a range of phonological phenomena, establishing information sufficiency, as the basic data and structure provide sufficient information to learn appropriate weights. They do not compare the learning behavior against empirical human learning behavior, beyond showing that the final state of the learning model is similar to the presumed adult phonology. The crucial point to establish is that the problem can be learned, although the specifics of human learning are left open.

The work in this dissertation on acquiring phonological constraints is an example of modeling to establish information sufficiency. In Ch. 2, I show that a phonology for vowel harmony can be learned using very limited phonological information, and no phonologically-driven structure for constraints. However, the constraints learned by this method are noisy, and lack the robust and generalizable definitions that humans presumably use in their phonological analysis. Ch. 3 adds phonological structure for the constraints and shows how this allows the model to learn less noisy constraints that are generalizable and resemble the standard kinds of constraints in phonology. Both of these methods establish information sufficiency, in that they show that the learning problem can be solved with a specific set of information and assumptions, and largely match human phonology, but they do not directly attempt to match the human learning process. This is a major goal in continued work on this learning problem.

Generic performance can also be tested comparatively to determine if the structure of one model allows for improved information extraction. The word segmentation model (Ch. 4) provides an example of this, as it compares the accuracy of the segmentations and extracted lexicons for models that ignore or include stress information in their segmentation decisions.

Human performance The second assessment of a computational model of cognition examines its predictions about human behavior. Whereas generic performance

assessments look at whether structure can be learned in principle, comparing a models performance to human performance addresses whether it can be learned in practice. Because this presents a more precise goal than simply maximizing generic performance, it often requires more fully developed models to attain good performance.

M. Frank et al. (2010) provide an example of this from Bayesian single-cue word segmentation. They ran three experiments in which participants listened to sentences from an artificial language, and then were asked which of a pair of possible words was more likely to be a word in the language. The normative likelihood in these experiments was based on the frequency of each possible word in the exposure phase. The three experiments manipulated the length of the sentences, the amount of exposure, and the size of the lexicon, to determine how these variables affected human segmentation.

With the human performance established, they also ran the segmentation model on the same data as the humans received, to determine how the variables affected model segmentation. This model performs Bayesian analysis on chunking cues, as opposed to other models that used non-Bayesian methods or used bracketing cues instead. They show that the Bayesian chunking model makes predictions that are much more tightly correlated with human performance than any of the other models, providing evidence for both Bayesian probabilistic thinking and the use of chunking cues by human segmenters. Furthermore, the model's ability to explain the effect of lexicon size on human performance is poor without any restrictions of the computational power of the model, but when memory limitations are added, human and model performance are highly correlated. This provides evidence for bounded rationality in human word segmentation. This also shows the importance of having different assessment techniques; a better fit to human performance requires worse generic performance.

The models in this dissertation do not go into human performance in great depth, as the phonological constraint acquisition models are still in their early stages. The

multiple-cue word segmentation model (Ch. 4) does look at human performance to examine the nature of the stress bias in segmentation. The fit to human performance is not as tight as M. Frank et al.s, but it does show that the observed change in cue importance as children age can be a reasonable response to the data children encounter. It also provides another example where improving the fit to human performance would almost certainly lower generic performance.

Application performance Because computational models perform rational analysis, they can also be put to use for natural language processing and other non-cognitive applications. Assessing performance on applications is similar to the generic performance assessment, but in terms of the goals of the application rather than the goals of the model itself.

Rasiwasia et al. (2010) provide a multimedia example, by modeling the relationships between texts and their associated images. A topic model is used to extract the text information, but its specific performance on this portion of the task is irrelevant. The application in this work was to provide a cross-modal query system, where text queries returned related images and image queries returned related texts. Performance for this model was assessed by the proportion of returned images that were from the same category as the query text, or vice versa. In such an application, it is possible for a model that would perform worse on generic or human measures to perform better in the context of the application.

The compound-multinomial topic model in Ch. 5 provides an example of application performance when it is applied to the task of classifying companies based on their stock price fluctuations (see also Doyle & Elkan, 2009b), with performance assessed based on the known company categories. The rest of the models lack such non-cognitive applications at present.

1.4 Overview of the models

This section provides a brief overview of each of the four models proposed in this dissertation, how they relate to the concepts discussed above, and what each can tell us about language and cognition.

In general, the models in this work argue for shifting explanatory power away from innate structures and onto the data. The models of phonological constraint acquisition replace the traditional assumption of an innate set of phonological constraints with ones learned from distributional data (Ch. 2 and 3). The multiple-cue model of word segmentation (Ch. 4) argues that the observed stress-based segmentation biases can arise from early-childhood word segmentation, rather than being the driving force behind it. The compound-multinomial topic model (Ch. 5) shows that such simple information as the co-occurrence of words can provide substantial information about their semantic relationships.

The phonological acquisition models use Bayesian inference over possible constraint violation sets to identify probable constraints with varying amounts of phonological structure; previous work either treated these constraints as innate or assumed that they must be learned from phonetically-grounded explanations, rather than the data itself. The word segmentation model represents more of the phonetic information in the language data than previous models, and shows that this both improves performance and explains the acquisition of biases that had previously been proposed as innate. The compound-multinomial topic model improves the representation of a semantic topic by introducing a hierarchical topic structure, which improves the model's ability to explain the language data, which leads to improvements in a range of applications.

1.4.1 Chapter 2: Constraint Acquisition without Phonological Structure

Chapter 2 describes the Indian Buffet Process Optimality Theory (IBPOT) model, which uses distributional data to find phonological constraints. The main goal of this model is to show that there is sufficient information in the language data to learn constraint violations and the relative importance of the constraints, as opposed to the standard OT assumption that constraints are innate and only their relative importances are learned.

The IBPOT model defines a phonological constraint extensionally – in terms of the wordforms that violate that constraint. This model does not presuppose any phonological structure over the constraints, minimizing the amount of innate knowledge within the model and providing the flexibility to learn any constraint, including possible language-specific constraints (J. Smith, 2004). Tests on Wolof vowel harmony show that this model is capable of learning a set of constraints that effectively explain the observed forms, and have a similar structure to the phonologically standard constraints. However, the lack of presupposed phonological structure leads to noise in the constraint estimates and limits the predictive power of the model. These limitations motivate the Rational Rules OT (RROT) model of the next chapter. Chapter 2, in full, is an exact copy of the material as it appears in the *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Doyle, Bicknell, & Levy, 2014).

1.4.2 Chapter 3: Constraint Acquisition with Phonological Structure

Chapter 3 describes the Rational Rules Optimality Theory (RROT) model, which extends the model structure from the IBPOT model by allowing knowledge of phonological structure to influence what constraint violations are more likely than others. This phonological structure comes from a tree-grammar over possible constraint definitions,

defined in terms of phonetic features. This is only a small increase in the amount of innate structure over the IBPOT model, and it is still much less presupposed structure than the fully innate constraint set that is standardly assumed in OT. It also does not render the constraint learning problem trivial, as the grammar produces an infinite space of possible constraint definitions. Tests on the English plural show that the RROT model is capable of learning a constraint structure that not only predicts the correct forms for previously-unseen words (which the IBPOT model could not) but also largely agrees with the standard set of constraints invoked to explain the English plural.

1.4.3 Chapter 4: Multiple-Cue Word Segmentation

Chapter 4 proposes a model for integrating different types of segmentation cues in early word segmentation. Prevailing models of word segmentation have used only a single cue, usually syllable or phoneme identity, often framed in terms of transition probabilities (Saffran, Aslin, & Newport, 1996; Goldwater, Griffiths, & Johnson, 2006). However, there is extensive evidence that children use multiple different cues when segmenting words from speech – including stress patterns, phonotactics, and allophonic variation – and that the importance of these cues changes over time (Johnson & Jusczyk, 2001; Thiessen & Saffran, 2003). This chapter introduces multiple-cue modeling (syllable identity and stress) into an existing Bayesian single-cue segmentation framework (Goldwater et al., 2006). The multiple-cue framework leads to significant improvements in the segmentation performance, and shows how exposure to data can cause the relative importance of different cues to shift. Chapter 4, in full, is an exact copy of the material as it appears in the *Proceedings of the 2013 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (Doyle & Levy, 2013).

1.4.4 Chapter 5: Burstiness in Topic Models

Chapter 5 moves into applied computational modeling, and illustrates how the effectiveness of a model can depend on the appropriateness of its structure for the task at hand. This chapter proposes the Dirichlet Compound Multinomial Latent Dirichlet Allocation (DCMLDA) model, a topic model that accounts for *burstiness*, the tendency for words to re-occur within the same document (Church & Gale, 1995). Topic models are a class of models that use the co-occurrences of words across documents to uncover semantic relationships between them, motivated by the idea that words that tend to co-occur are more likely to be related. Accounting for burstiness provides a more accurate estimate of the co-occurrence between words. Tests in this chapter show that this improves the model's performance on both linguistic data and bursty non-linguistic data. Chapter 5, in full, is an exact copy of the material as it appears in the *Proceedings of the 26th International Conference on Machine Learning* (Doyle & Elkan, 2009a).

Chapter 2

Nonparametric learning of phonological constraints in Optimality Theory

Abstract We present a method to jointly learn features and weights directly from distributional data in a log-linear framework. Specifically, we propose a non-parametric Bayesian model for learning phonological markedness constraints directly from the distribution of input-output mappings in an Optimality Theory (OT) setting. The model uses an Indian Buffet Process prior to learn the feature values used in the log-linear method, and is the first algorithm for learning phonological constraints without presupposing constraint structure. The model learns a system of constraints that explains observed data as well as the phonologically-grounded constraints of a standard analysis, with a violation structure corresponding to the standard constraints. These results suggest an alternative data-driven source for constraints instead of a fully innate constraint set.

2.1 Introduction

Many aspects of human cognition involve the interaction of constraints that push a decision-maker toward different options, whether in something so trivial as choosing a movie or so important as a fight-or-flight response. These constraint-driven decisions can be modeled with a log-linear system. In these models, a set of constraints is weighted

and their violations are used to determine a probability distribution over outcomes. But where do these constraints come from?

We consider this question by examining the dominant framework in modern phonology, Optimality Theory (Prince & Smolensky, 1993, OT), implemented in a log-linear framework, MaxEnt OT (Goldwater & Johnson, 2003), with output forms' probabilities based on a weighted sum of constraint violations. OT analyses generally assume that the constraints are innate and universal, both to obviate the problem of learning constraints' identities and to limit the set of possible languages.

We propose a new approach: to learn constraints with limited innate phonological knowledge by identifying sets of constraint violations that explain the observed distributional data, instead of selecting constraints from an innate set of constraint definitions. Because the constraints are identified as sets of violations, this also permits constraints specific to a given language to be learned. This method, which we call IBPOT, uses an Indian Buffet Process (IBP) prior to define the space of possible constraint violation matrices, and uses Bayesian reasoning to identify constraint matrices likely to have generated the observed data. In identifying constraints solely by their extensional violation profiles, this method does not directly identify the intensional definitions of the identified constraints, but to the extent that the resulting violation profiles are phonologically interpretable, we may conclude that the data themselves guide constraint identification. We test IBPOT on tongue-root vowel harmony in Wolof, a West African language.

The set of constraints learned by the model satisfy two major goals: they explain the data as well as the standard phonological analysis, and their violation structures correspond to the standard constraints. This suggests an alternative data-driven genesis for constraints, rather than the traditional assumption of fully innate constraints.

2.2 Phonology and Optimality Theory

2.2.1 OT structure

Optimality Theory has been used for constraint-based analysis of many areas of language, but we focus on its most successful application: phonology. We consider an OT analysis of the mappings between underlying forms and their phonological manifestations – i.e., mappings between forms in the mental lexicon and the actual vocalized forms of the words.[1]

Stated generally, an OT system takes some input, generates a set of candidate outputs, determines what constraints each output violates, and then selects a candidate output with a relatively unobjectionable violation profile. To do this, an OT system contains four major components: a generator GEN, which generates candidate output forms for the input; a set of constraints CON, which penalize candidates; a evaluation method EVAL, which selects an winning candidate; and H, a language-particular weighting of constraints that EVAL uses to determine the winning candidate. Previous OT work has focused on identifying the appropriate formulation of EVAL and the values and acquisition of H, while taking GEN and CON as given. Here, we expand the learning task by proposing an acquisition method for CON.

To learn CON, we propose a data-driven markedness constraint learning system that avoids both innateness and tractability issues. Unlike previous OT learning methods, which assume known constraint definitions and only learn the relative strength of these constraints, the IBPOT learns constraint violation profiles and weights for them simultaneously. The constraints are derived from sets of violations that effectively explain the observed data, rather than being selected from a pre-existing set of possible constraints.

[1] Although phonology is usually framed in terms of sound, sign languages also have components that serve equivalent roles in the physical realization of signs (Stokoe, 1960).

2.2.2 OT as a weighted-constraint method

Although all OT systems share the same core structure, different choices of EVAL lead to different behaviors. In IBPOT, we use the log-linear EVAL developed by Goldwater and Johnson (2003) in their MaxEnt OT system. MEOT extends traditional OT to account for variation (cases in which multiple candidates can be the winner), as well as gradient/probabilistic productions (Anttila, 1997) and other constraint interactions (e.g., cumulativity) that traditional OT cannot handle (Keller, 2000). MEOT also is motivated by the general MaxEnt framework, whereas most other OT formulations are ad hoc constructions specific to phonology.

In MEOT, each constraint C_i is associated with a weight $w_i < 0$. (Weights are always negative in OT; a constraint violation can never make a candidate more likely to win.) For a given input-candidate pair (x,y), $f_i(y,x)$ is the number of violations of constraint C_i by the pair. As a maximum entropy model, the probability of y given x is proportional to the exponential of the weighted sum of violations, $\sum_i w_i f_i(y,x)$. If $\mathscr{Y}(x)$ is the set of all output candidates for the input x, then the probability of y as the winning output is:

$$p(y|x) = \frac{\exp\left(\sum_i w_i f_i(y,x)\right)}{\sum_{z\in\mathscr{Y}(x)} \exp\left(\sum_i w_i f_i(z,x)\right)} \tag{2.1}$$

This formulation represents a probabilistic extension of the traditional formulation of OT (Prince & Smolensky, 1993). Traditionally, constraints form a strict hierarchy, where a single violation of a high-ranked constraint is worse than any number of violations of lower-ranked constraints. Traditional OT is also deterministic, with the optimal candidate always selected. In MEOT, the constraint weights define hierarchies of varying strictness, and some probability is assigned to all candidates. If constraints' weights are close together, multiple violations of lower-weighted constraints can reduce a candidate's probability below that of a competitor with a single high-weight violation. As the distance

ete	*ɪ	Parse(rtr)	Harmony	Parse(atr)	Score
ete					0
ɛte					-24
etɛ					-24
ɛtɛ					-8

ɪte	*ɪ	Parse(rtr)	Harmony	Parse(atr)	Score
ite					-32
ɪte					-80
itɛ					-56
ɪtɛ					-72

ɛtɛ	*ɪ	Parse(rtr)	Harmony	Parse(atr)	Score
ete					-32
ɛte					-48
etɛ					-48
ɛtɛ					0

itɛ	*ɪ	Parse(rtr)	Harmony	Parse(atr)	Score
ite					-32
ɪte					-120
itɛ					-16
ɪtɛ					-72

Figure 2.1. Tableaux for the Wolof input forms ete, ɛtɛ, ɪte, and itɛ. Black indicates violation, white no violation. Scores are calculated for a MaxEnt OT system with constraint weights of -64, -32, -16, and -8, approximating a traditional hierarchical OT design. Values of grey-striped cells have negligible effects on the distribution (see Sect. 2.4.3).

between weights in MEOT increases, the probability of a suboptimal candidate being chosen approaches zero; thus the traditional formulation is a limit case of MEOT.

2.2.3 OT in practice

Figure 2.1 shows *tableaux*, a visualization for OT, applied in Wolof (Archangeli & Pulleyblank, 1994; Boersma, 1999). We are interested in four Wolof constraints that combine to induce vowel harmony: *ɪ, PARSE[rtr], HARMONY, and PARSE[atr]. The meaning of these constraints will be discussed in Sect. 2.4.1; for now, we will only consider their violation profiles. Each column represents a constraint, with weights decreasing left-to-right. Each tableau looks at a single input form, noted in the top-left cell: ete, ɛtɛ, ɪte, or itɛ.

Each row is a candidate output form. A black cell indicates that the candidate, or input-candidate pair, violates the constraint in that column.[2] A white cell indicates no violation. Grey stripes are overlaid on cells whose value will have a negligible impact on the distribution due to the values of higher-ranked constraint.

Constraints fall into two categories, faithfulness and markedness, which differ

[2]In general, a constraint can be violated multiple times by a given candidate, but we will be using binary constraints (violated or not) in this work. See Sect. 2.5.2 for further discussion.

in what information they use to assign violations. Faithfulness constraints penalize mismatches between the input and output, while markedness constraints consider only the output. Faithfulness violations include phoneme additions or deletions between the input and output; markedness violations include penalizing specific phonemes in the output form, regardless of whether the phoneme is present in the input.

In MaxEnt OT, each constraint has a weight, and the candidates' scores are the sums of the weights of violated constraints. In the ete tableau at top left, output ete has no violations, and therefore a score of zero. Outputs ɛte and etɛ violate both HARMONY (weight 16) and PARSE[atr] (weight 8), so their scores are 24. Output ɛtɛ violates PARSE[atr], and has score 8. Thus the log-probability of output ɛtɛ is 1/8 that of ete, and the log-probability of disharmonious ɛte and etɛ are each 1/24 that of ete. As the ratio between scores increases, the log-probability ratios can become arbitrarily close to zero, approximating the deterministic situation of traditional OT.

2.2.4 Learning Constraints

Choosing a winning candidate presumes that a set of constraints CON is available, but where do these constraints come from? The standard assumption within OT is that CON is innate and universal. But in the absence of direct evidence of innate constraints, we should prefer a method that can derive the constraints from cognitively-general learning over one that assumes they are pre-specified. Learning appropriate model features has been an important idea in the development of constraint-based models (Della Pietra, Della Pietra, & Lafferty, 1997).

The innateness assumption can induce tractability issues as well. The strictest formulation of innateness posits that virtually all constraints are shared across all languages, even when there is no evidence for the constraint in a particular language (Tesar & Smolensky, 2000). Strict universality is undermined by the extremely large set of

constraints it must weight, as well as the possible existence of language-particular constraints (J. Smith, 2004).

A looser version of universality supposes that constraints are built compositionally from a set of constraint templates or primitives or phonological features (Hayes, 1999; J. Smith, 2004; Idsardi, 2006; Riggle, 2009). This version allows language-particular constraints, but it comes with a computational cost, as the learner must be able to generate and evaluate possible constraints while learning the language's phonology. Even with relatively simple constraint templates, such as the phonological constraint learner of Hayes and Wilson (Hayes & Wilson, 2008), the number of possible constraints expands exponentially. Depending on the specific formulation of the constraints, the constraint identification problem may even be NP-hard (Idsardi, 2006; Heinz, Kobele, & Riggle, 2009). Our approach of casting the learning problem as one of identifying violation profiles is an attempt to determine the amount that can be learned about the active constraints in a paradigm without hypothesizing intensional constraint definitions. The violation profile information used by our model could then be used to narrow the search space for intensional constraints, either by performing post-hoc analysis of the constraints identified by our model or by combining intensional constraint search into the learning process. We discuss each of these possibilities in Section 2.5.2.

Innateness is less of a concern for faithfulness than markedness constraints. Faithfulness violations are determined by the changes between an input form and a candidate, yielding an independent motivation for a universal set of faithfulness constraints (McCarthy, 2008). Some markedness constraints can also be motivated in a universal manner (Hayes, 1999), but many markedness constraints lack such grounding.[3] As such, it is unclear where a universal set of markedness constraints would come from.

[3]McCarthy (2008, §4.8) gives examples of "ad hoc" intersegmental constraints. Even well-known constraint types, such as generalized alignment, can have disputed structures (Hyde, 2012).

2.3 The IBPOT Model

2.3.1 Structure

The IBPOT model defines a generative process for mappings between input and output forms based on three latent variables: the constraint violation matrices F (faithfulness) and M (markedness), and the weight vector w. The cells of the violation matrices correspond to the number of violations of a constraint by a given input-output mapping. F_{ijk} is the number of violations of faithfulness constraint F_k by input-output pair type (x_i, y_j); M_{jl} is the number of violations of markedness constraint $M_{\cdot l}$ by output candidate y_j. Note that M is shared across inputs, as M_{jl} has the same value for all input-output pairs with output y_j. The weight vector w provides weight for both F and M. Probabilities of output forms are given by a log-linear function:

$$p(y_j|x_i) = \frac{\exp\left(\sum_k w_k F_{ijk} + \sum_l w_l M_{jl}\right)}{\sum_{y_z \in \mathscr{Y}(x_i)} \exp\left(\sum_k w_k F_{izk} + \sum_l w_l M_{zl}\right)} \tag{2.2}$$

Note that this is the same structure as Eq. 2.1 but with faithfulness and markedness constraints listed separately. As discussed in Sect. 2.2.4, we assume that F is known as part of the output of GEN (Riggle, 2009). The goal of the IBPOT model is to learn the markedness matrix M and weights w for both the markedness and faithfulness constraints.

As for M, we need a non-parametric prior, as there is no inherent limit to the number of markedness constraints a language will use. We use the Indian Buffet Process (Griffiths & Ghahramani, 2005), which defines a proper probability distribution over binary feature matrices with an unbounded number of columns. The IBP can be thought of as representing the set of dishes that diners eat at an infinite buffet table. Each diner (i.e., output form) first draws dishes (i.e., constraint violations) with probability proportional to the number of previous diners who drew it: $p(M_{jl} = 1|\{M_{zl}\}_{z<j}) = n_l/j$. After choosing from the previously taken dishes, the diner can try additional dishes that

no previous diner has had. The number of new dishes that the j-th customer draws follows a Poisson(α/j) distribution. The complete specification of the model is then:

$$M \sim IBP(\alpha); \qquad \mathscr{Y}(x_i) = Gen(x_i)$$
$$w \sim -\Gamma(1,1); \;\; y|x_i \sim LogLin(M,F,w,\mathscr{Y}(x_i))$$

2.3.2 Inference

To perform inference in this model, we adopt a common Markov chain Monte Carlo estimation procedure for IBPs (Görür, Jäkel, & Rasmussen, 2006; Navarro & Griffiths, 2007). We alternate approximate Gibbs sampling over the constraint matrix M, using the IBP prior, with a Metropolis-Hastings method to sample constraint weights w.

We initialize the model with a randomly-drawn markedness violation matrix M and weight vector w. To learn, we iterate through the output forms y_j; for each, we split $M_{-j\cdot}$ into "represented" constraints (those that are violated by at least one output form other than y_j) and "non-represented" constraints (those violated only by y_j). For each represented constraint $M_{\cdot l}$, we re-sample the value for the cell M_{jl}. All non-represented constraints are removed, and we propose new constraints, violated only by y_j, to replace them. After each iteration through M, we use Metropolis-Hastings to update the weight vector w.

Represented constraint sampling We begin by resampling M_{jl} for all represented constraints $M_{\cdot l}$, conditioned on the rest of the violations ($M_{-(jl)}, F$) and the weights w. This is the sampling counterpart of drawing existing features in the IBP generative process. By Bayes' Rule, the posterior probability of a violation is proportional to product of the likelihood $p(Y|M_{jl}=1, M_{-jl}, F, w)$ from Eq. 2.2 and the IBP prior probability $p(M_{jl}=1|M_{-jl}) = n_{-jl}/n$, where n_{-jl} is the number of outputs other than y_j that violate constraint $M_{\cdot l}$.

Non-represented constraint sampling After sampling the represented constraints for y_j, we consider the addition of new constraints that are violated only by y_j. This is the sampling counterpart to the Poisson draw for new features in the IBP generative process. Ideally, this would draw new constraints from the infinite feature matrix; however, this requires marginalizing the likelihood over possible weights, and we lack an appropriate conjugate prior for doing so. We approximate the infinite matrix with a truncated Bernoulli draw over unrepresented constraints (Görür et al., 2006). We consider in each sample at most K^* new constraints, with weights based on the auxiliary vector w^*. This approximation retains the unbounded feature set of the IBP, as repeated sampling can add more and more constraints without limit.

The auxiliary vector w^* contains the weights of all the constraints that have been removed in the previous step. If the number of constraints removed is less than K^*, w^* is filled out with draws from the prior distribution over weights. We then consider adding any subset of these new constraints to M, each of which would be violated only by y_j. Let M^* represent a (possibly empty) set of constraints paired with a subset of w^*. The posterior probability of drawing M^* from the truncated Bernoulli distribution is the product of the prior probability of M^* $\left(\frac{\frac{\alpha}{K^*}}{N_Y+\frac{\alpha}{K^*}}\right)$ and the likelihood $p(Y|M^*, w^*, M, w, F)$, including the new constraints M^*.

Weight sampling After sampling through all candidates, we use Metropolis-Hastings to estimate new weights for both constraint matrices. Our proposal distribution is $Gamma({w_k}^2/\eta, \eta/w_k)$, with mean w_k and mode $w_k - \frac{\eta}{w_k}$ (for $w_k > 1$). Unlike Gibbs sampling on the constraints, which occurs only on markedness constraints, weights are sampled for both markedness and faithfulness features.

2.4 Experiment

2.4.1 Wolof vowel harmony

We test the model by learning the markedness constraints driving Wolof vowel harmony (Archangeli & Pulleyblank, 1994). Vowel harmony in general refers to a phonological phenomenon wherein the vowels of a word share certain features in the output form even if they do not share them in the input. In the case of Wolof, harmony encourages forms that have consistent tongue root positions.

The Wolof vowel system has two relevant features, tongue root position and vowel height. The tongue root can either be advanced (ATR) or retracted (RTR), and the body of the tongue can be in the high, middle, or low part of the mouth. These features define six vowels:

	high	mid	low
ATR	i	e	ə
RTR	ɪ	ɛ	a

We test IBPOT on the harmony system provided in the Praat program (Boersma, 1999), previously used as a test case by Goldwater and Johnson (2003) for MEOT learning with known constraints. This system has four constraints:[4]

- Markedness:
  - *ɪ: do not have ɪ (high RTR vowel)
  - HARMONY: do not have RTR and ATR vowels in the same word
- Faithfulness:
  - PARSE[rtr]: do not change RTR input to ATR output

[4]The version in Praat includes a fifth constraint, but its value never affects the choice of output in our data and is omitted in this analysis.

– PARSE[atr]: do not change ATR input to RTR output

These constraints define the phonological standard that we will compare IBPOT to, with a ranking from strongest to weakest of *ɪ >> PARSE[rtr] >> HARMONY >> PARSE[atr]. Under this ranking, Wolof harmony is achieved by changing a disharmonious ATR to an RTR, unless this creates an ɪ vowel. We see this in Figure 2.1, where three of the four winners are harmonic, but with input itɛ, harmony would require violating one of the two higher-ranked constraints. As in previous MEOT work, all Wolof candidates are faithful with respect to vowel height, either because height changes are not considered by GEN, or because of a high-ranked faithfulness constraint blocking height changes.[5]

The Wolof constraints provide an interesting testing ground for the model, because it is a small set of constraints to be learned, but contains the HARMONY constraint, which can be violated by non-adjacent segments. Non-adjacent constraints are difficult for string-based approaches because of the exponential number of possible relationships across non-adjacent segments. However, the Wolof results show that by learning violations directly, IBPOT does not encounter problems with non-adjacent constraints.

The Wolof data has 36 input forms, each of the form V_1tV_2, where V_1 and V_2 are vowels that agree in height. Each input form has four candidate outputs, with one output always winning. The outputs appear for multiple inputs, as shown in Figure 2.1. The candidate outputs are the four combinations of tongue-roots for the given vowel heights; the inputs and candidates are known to the learner. We generate simulated data by observing 1000 instances of the winning output for each input.[6] The model must learn the markedness constraints *ɪ and HARMONY, as well as the weights for all four

[5] In the present experiment, we assume that GEN does not generate candidates with unfaithful vowel heights. If unfaithful vowel heights were allowed by GEN, these unfaithful candidates would incur a violation approximately as strong as *ɪ, as neither unfaithful-height candidates nor ɪ candidates are attested in the Wolof data.

[6] Since data, matrix, and weight likelihoods all shape the learned constraints, there must be enough data for the model to avoid settling for a simple matrix that poorly explains the data. This represents a similar training set size to previous work (Goldwater & Johnson, 2003; Boersma & Hayes, 2001).

constraints.

We make a small modification to the constraints for the test data: all constraints are limited to binary values. For constraints that can be violated multiple times by an output (e.g., *ɪ twice by ɪtɪ), we use only a single violation. This is necessary in the current model definition because the IBP produces a prior over binary matrices. We generate the simulated data using only single violations of each constraint by each output form. Overcoming the binarity restriction is discussed in Sect. 2.5.2.

2.4.2 Experiment Design

We run the model for 10000 iterations, using deterministic annealing through the first 2500 iterations. The model is initialized with a random markedness matrix drawn from the IBP and weights from the exponential prior. We ran versions of the model with parameter settings between 0.01 and 1 for α, 0.05 and 0.5 for η, and 2 and 5 for K^*. All these produced quantitatively similar results; we report values for $\alpha = 1$, $\eta = 0.5$, and $K^* = 5$, which provides the least bias toward small constraint sets.

To establish performance for the phonological standard, we use the IBPOT learner to find constraint weights but do not update M. The resultant learner is essentially MaxEnt OT with the weights estimated through Metropolis sampling instead of gradient ascent. This is done so that the IBPOT weights and phonological standard weights are learned by the same process and can be compared. We use the same parameters for this baseline as for the IBPOT tests. The results in this section are based on nine runs each of IBPOT and MEOT; ten MEOT runs were performed but one failed to converge and was removed from analysis.

Table 2.1. Data, markedness matrix, weight vector, and joint log-probabilities for the IBPOT and the phonological standard constraints. MAP and mean estimates over the final 1000 iterations for each run. All IBPOT/PS differences are significant ($p < .005$ for MAP M; $p < .001$ for others).

	MAP		Mean	
	IBPOT	PS	IBPOT	PS
Data	-1.52	-3.94	-5.48	-9.23
M	-51.7	-53.3	-54.7	-53.3
w	-44.2	-71.1	-50.6	-78.1
Joint	-97.4	-128.4	-110.6	-140.6

2.4.3 Results

A successful set of learned constraints will satisfy two criteria: achieving good data likelihood (no worse than the phonological-standard constraints) and acquiring constraint violation profiles that are phonologically interpretable. We find that both of these criteria are met by IBPOT on Wolof.

Likelihood comparison First, we calculate the joint probability of the data and model given the priors, $p(Y, M, w|F, \alpha)$, which is proportional to the product of three terms: the data likelihood $p(Y|M, F, w)$, the markedness matrix probability $p(M|\alpha)$, and the weight probability $p(w)$. We present both the mean and MAP values for these over the final 1000 iterations of each run. Results are shown in Table 2.1.

All eight differences are significant according to t-tests over the nine runs. In all cases but mean M, the IBPOT method has a better log-probability. The most important differences are those in the data probabilities, as the matrix and weight probabilities are reflective primarily of the choice of prior. By both measures, the IBPOT constraints explain the observed data better than the phonologically standard constraints.

Interestingly, the mean M probability is lower for IBPOT than for the phonological standard. Though the phonologically standard constraints exist independently of the IBP

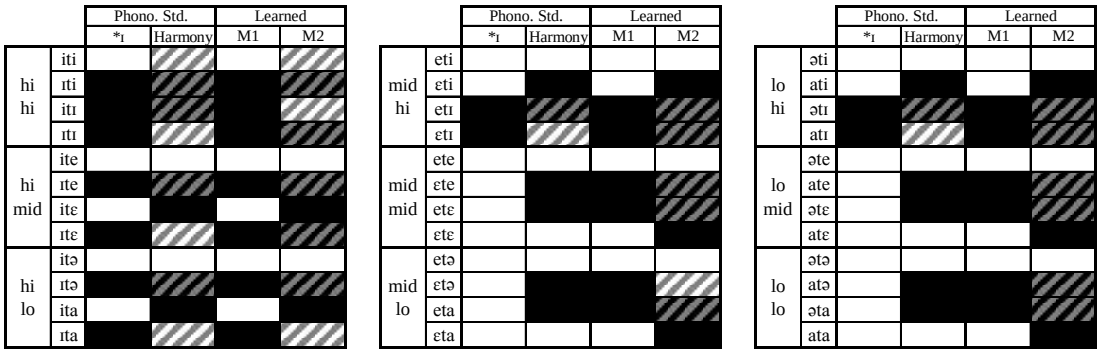


Figure 2.2. Phonologically standard (*ɪ, HARMONY) and learned (*M1,M2*) constraint violation profiles for the output forms. Learned weights for the standard constraints are -32.8 and -15.3; for *M1* and *M2*, they are -26.5 and -8.4. Black indicates violation, white no violation. Grey stripes indicate cells whose values have negligible effects on the probability distribution.

prior, they fit the prior better than the average IBPOT constraints do. This shows that the IBP's prior preferences can be overcome in order to have constraints that better explain the data.

Constraint comparison Our second criterion is the acquisition of meaningful constraints, that is, ones whose violation profiles have phonologically-grounded explanations. IBPOT learns the same number of markedness constraints as the phonological standard (two); over the final 1000 iterations of the model runs, 99.2% of the iterations had two markedness constraints, and the rest had three.

Turning to the form of these constraints, Figure 2.2 shows violation profiles from the last iteration of a representative IBPOT run.[7] Because vowel heights must be faithful between input and output, the Wolof data is divided into nine separate *paradigms*, each containing the four candidates (ATR/RTR $\times$ ATR/RTR) for the vowel heights in the input.

The violations on a given output form only affect probabilities within its paradigm. As a result, learned constraints are consistent within paradigms, but across paradigms,

[7] Specifically, from the run with the median joint posterior.

the same constraint may serve different purposes.

For instance, the strongest learned markedness constraint, shown as *M1* in Figure 2.2, has the same violations as the top-ranked constraint that actively distinguishes between candidates in each paradigm. For the five paradigms with at least one high vowel (the top row and left column), *M1* has the same violations as *ɪ, as *ɪ penalizes some but not all of the candidates. In the other four paradigms, *ɪ penalizes none of the candidates, and the IBPOT learner has no reason to learn it. Instead, it learns that *M1* has the same violations as HARMONY, which is the highest-weighted constraint that distinguishes between candidates in these paradigms. Thus in the high-vowel paradigms, *M1* serves as *ɪ, while in the low/mid-vowel paradigms, it serves as HARMONY.

The lower-weighted *M2* is defined noisily, as the higher-ranked *M1* makes some values of *M2* inconsequential. Consider the top-left paradigm of Figure 2.2, the high-high input, in which only one candidate does not violate *M1* (*ɪ). Because *M1* has a much higher weight than *M2*, a violation of *M2* has a negligible effect on a candidate's probability.[8] In such cells, the constraint's value is influenced more by the prior than by the data. These inconsequential cells are overlaid with grey stripes in Figure 2.2.

The meaning of *M2*, then, depends only on the consequential cells. In the high-vowel paradigms, *M2* matches HARMONY, and the learned and standard constraints agree on all consequential violations, despite being essentially at chance on the indistinguishable violations (58%). On the non-high paradigms, the meaning of *M2* is unclear, as HARMONY is handled by *M1* and *ɪ is unviolated. In all four paradigms, the model learns that the RTR-RTR candidate violates *M2* and the ATR-ATR candidate does not; this appears to be the model's attempt to reinforce a pattern in the lowest-ranked faithfulness constraint (PARSE[atr]), which the ATR-ATR candidate never violates.

[8] Given the learned weights in Fig. 2.2, if the losing candidate violates *M1*, its probability changes from 10^{-12} when the preferred candidate does not violate *M2* to 10^{-8} when it does.

Thus, while the IBPOT constraints are not identical to the phonologically standard ones, they reflect a version of the standard constraints that is consistent with the IBPOT framework.[9] In paradigms where each markedness constraint distinguishes candidates, the learned constraints match the standard constraints. In paradigms where only one constraint distinguishes candidates, the top learned constraint matches it and the second learned constraint exhibits a pattern consistent with a low-ranked faithfulness constraint.

2.5 Discussion and Future Work

2.5.1 Relation to phonotactic learning

Our primary finding from IBPOT is that it is possible to identify constraints that are both effective at explaining the data and representative of theorized phonologically-grounded constraints, given only input-output mappings and faithfulness violations. Furthermore, these constraints are successfully acquired without any knowledge of the phonological structure of the data beyond the faithfulness violation profiles. The model's ability to infer constraint violation profiles without theoretical constraint structure provides an alternative solution to the problems of the traditionally innate and universal OT constraint set.

As it jointly learns constraints and weights, the IBPOT model calls to mind Hayes and Wilson's (Hayes & Wilson, 2008) joint phonotactic learner. Their learner also jointly learns weights and constraints, but directly selects its constraints from a compositional grammar of constraint definitions. This limits their learner in practice by the rapid explosion in the number of constraints as the maximum constraint definition size grows. By directly learning violation profiles, the IBPOT model avoids this explosion, and the violation profiles can be automatically parsed to identify the constraint definitions that are

[9] In fact, it appears this constraint organization is favored by IBPOT as it allows for lower weights, hence the large difference in w log-probability in Table 2.1.

consistent with the learned profile. The inference method of the two models is different as well; the phonotactic learner selects constraints greedily, whereas the sampling on M in IBPOT asymptotically approaches the posterior.

The two learners also address related but different phonological problems. The phonotactic learner considers phonotactic problems, in which only output matters. The constraints learned by Hayes and Wilson's learner are essentially OT markedness constraints, but their learner does not have to account for varied inputs or effects of faithfulness constraints.

2.5.2 Extending the learning model

IBPOT, as proposed here, learns constraints based on binary violation profiles, defined extensionally. A complete model of constraint acquisition should provide intensional definitions that are phonologically grounded and cover potentially non-binary constraints. We discuss how to extend the model toward these goals.

IBPOT currently learns extensional constraints, defined by which candidates do or do not violate the constraint. Intensional definitions are needed to extend constraints to unseen forms. Post hoc violation profile analysis, as in Sect. 2.4.3, provides a first step toward this goal. Such analysis can be integrated into the learning process using the Rational Rules model (Goodman, Tenebaum, Feldman, & Griffiths, 2008) to identify likely constraint definitions compositionally. Alternately, phonological knowledge could be integrated into a joint constraint learning process in the form of a naturalness bias on the constraint weights or a phonologically-motivated replacement for the IBP prior.

The results presented here use binary constraints, where each candidate violates each constraint only once, a result of the IBP's restriction to binary matrices. Non-binarity can be handled by using the binary matrix M to indicate whether a candidate violates a constraint, with a second distribution determining the number of violations. Alternately,

a binary matrix can directly capture non-binary constraints; R. Frank and Satta (1998) converted existing non-binary constraints into a binary OT system by representing non-binary constraints as a set of equally-weighted overlapping constraints, each accounting for one violation. The non-binary harmony constraint, for instance, becomes a set {*(at least one disharmony), *(at least two disharmonies), etc.}.

Lastly, the Wolof vowel harmony problem provides a test case with overlaps in the candidate sets for different inputs. This candidate overlap helps the model find appropriate constraint structures. Analyzing other phenomena may require the identification of appropriate abstractions to find this same structural overlap. English regular plurals, for instance, fall into broad categories depending on the features of the stem-final phoneme. IBPOT learning in such settings may require learning an appropriate abstraction as well.

2.6 Conclusion

A central assumption of Optimality Theory has been the existence of a fixed inventory of universal markedness constraints innately available to the learner, an assumption by arguments regarding the computational complexity of constraint identification. However, our results show for the first time that nonparametric, data-driven learning can identify sparse constraint inventories that both accurately predict the data and are phonologically meaningful, providing a serious alternative to the strong nativist view of the OT constraint inventory.

2.7 Acknowledgments

We wish to thank Eric Baković, Emily Morgan, Mark Myslín, the UCSD Computational Psycholinguistics Lab, the Phon Company, and the ACL reviewers for their discussions and feedback on this work. This research was supported by NSF award IIS-0830535 and an Alfred P. Sloan Foundation Research Fellowship to RL.

This chapter, in full, is an exact copy of the material as it appears in Doyle, Bicknell, and Levy (2014) [Nonparametric learning of phonological constraints in Optimality Theory. In K. Toutanova and H. Wu (Eds.), Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (pp. 1094-1103). Baltimore: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper.

Chapter 3

Data-driven acquisition of phonological constraints with underlying phonological structure

Abstract We present a method for learning phonological constraints in an Optimality Theory (OT) framework. Specifically, we propose a non-parametric Bayesian model for learning phonological markedness constraints directly from the distribution of input-output mappings. The model introduces basic phonological structure through an infinite grammar over phonetic features, which provides a prior over the space of constraints. The model, tested on English regular plurals, learns a system of constraints that explains observed forms and extends to new forms as well as the innate constraints of a standard OT analysis. The constraints learned by the model also reflect the structure of the standard innate constraints, and the differences between the learned and innate constraints suggest that learning across phonological patterns is critical to finding appropriate constraints. This provides an emergentist alternative to the common assumption that phonological constraints are innately known.

3.1 Introduction

The fact that children can learn the phonology of a particular language is obvious, but it remains unclear exactly what that learning entails. How much of the phonological learning task is innate? Are humans innately aware of which constraints or rules are available, or is acquiring the constraints or rules part of the learning task?

The standard view, especially within Optimality Theory (OT), is that virtually all of the phonological structure is innate and universal; the differences between languages arise only from the language-particular ordering of the universal constraints. Phonological acquisition is then limited to determining the relative importances of the constraints for the language being learned. This can be called the *innatist* position. Alternatively, phonological acquisition could include acquiring the constraints; this is an *emergentist* position.

One of the major arguments undergirding the innatist position is that there is no adequate learning method for acquiring the full range of constraints (Kager, Pater, & Zonneveld, 2004). Emergentist proposals are able to acquire "phonetically grounded" constraints, those that can be induced from articulatory or perceptual experience (Archangeli & Pulleyblank, 1994; Boersma, 1998; Hayes, 1999; Flack, 2007), but these proposals are unable to learn "formal" constraints, such as stress alignment constraints, which cannot be so induced.

This chapter strengthens the emergentist position by proposing and testing a non-parametric Bayesian model for acquiring phonological constraints without knowledge of articulation or perception. The model uses distributional data and basic phonological structure (in this case, a set of phonological features) to acquire its constraints. The model defines a flexible framework that can be implemented with various proposals for the underlying phonological structure, including formal structures that existing emergentist

proposals would not learn.

We test the model on the morphophonology of the English plural. We examine the model's performance on three measures: how well its learned structure explains the observed plural forms, how well it predicts the plural forms of previously-unseen words, and how well the learned constraint structure corresponds to the standard phonological constraints. We show that the performance on observed forms, as measured by data probabilities and joint data/structure likelihood, is on par with that of the standard innate constraints. Predictive performance on novel forms is also equal to that of the innate constraints, selecting appropriate plurals for the novel words. Lastly, the structure of the constraints is similar to the phonological standard constraints, and the deviations between the innate and learned constraints suggest an important role to be played by general phonotactic knowledge or cross-paradigm learning.

Overall, we show that the learning model finds appropriate constraints in an infinite space of possible constraint definitions, and that these constraints have similar generality to innately-defined constraints. Because this learning method is driven by the set of observed phonological forms rather than directly coming from articulatory or perceptual motivations, it provides a stronger alternative to innate constraints than previous emergentist proposals.

3.2 Phonological Acquisition

3.2.1 Constraint-Based Phonology

Constraint-based phonology, in the form of Optimality Theory (OT; Prince & Smolensky, 1993), uses a system of violable constraints that are used to choose between multiple candidate output forms for an input form. Stated generally, an OT analysis works as follows. Given an underlying form, a set of candidate outputs is generated,

and each is compared against the set of constraints. The candidate output with the least objectionable constraint violations (whether violating fewer or less important constraints) is chosen as the output form.[1]

We work in the MaxEnt OT (MEOT) framework (Goldwater & Johnson, 2003), which is a log-linear extension of the traditional OT framework (Prince & Smolensky, 1993). Unlike the traditional framework, where constraints are categorically ranked, MEOT weights the features.[2] Weighted features extend traditional OT, as they can capture a categorical structure when the weights are far apart (like traditional OT) but can also account for gradient outcomes and variation when the weights are closer together (unlike traditional OT). MEOT also has the advantage of connecting conceptually to the general class of log-linear models, which have well-studied convergence and learning behavior (Della Pietra et al., 1997; Lafferty, McCallum, & Pereira, 2001; N. Smith & Eisner, 2005) and have proven effective in a range of language and human cognition studies (Görür et al., 2006; Poon, Cherry, & Toutanova, 2009). Of particular importance to this work, Hayes and Wilson (2008) used a log-linear framework in their model of phonotactic constraint acquisition.

3.2.2 Constraint structures and their acquisition

OT constraints divide into two general types: faithfulness and markedness, depending on whether the constraint considers the underlying input form. Faithfulness constraints penalize output candidates for deviating from their inputs in specific ways, such as the deletion of a phoneme or the change of a phonetic feature. Identifying and assessing violations of faithfulness constraints are straightforward, and Riggle (2009)

[1] This can also be a gradient evaluation, with less objectionable candidates being more likely, which can account for variation in the output. (e.g., (Anttila, 1997; Boersma, 1999))

[2] In its use of weighted features, MEOT is similar to one of OT's progenitors, Harmonic Grammar (Legendre, Miyata, & Smolensky, 1990). Harmonic Grammar differs in that it allows for positive or negative weights. Thus "violations" of an HG constraint could increase a candidate's acceptability, unlike in traditional or MaxEnt OT.

has shown how they can be represented as part of the output of a finite-state transducer that generates the output candidates from the input form. As such, we will treat these constraints as known a priori in our model, although the model must still learn appropriate weights for them. In principle, faithfulness constraints could be learned similarly to the markedness constraints, as discussed in Sect. 3.6.3.

The model focuses on learning markedness constraints, which penalize output candidates independent of the input form. Many are phonetically or psycholinguistically grounded and penalize features that are articulatorily or perceptually difficult, such as long consonant sequences (Archangeli & Pulleyblank, 1994). It is possible that some markedness constraints are "formally grounded" constraints, which rely on formal linguistic elements beyond the phonetically-grounded motivations (Flack, 2007), although the existence of formally grounded constraints is debated (Kager et al., 2004). Alignment constraints are one example of a debated formal constraint; McCarthy and Prince (1993) treat alignment constraints as psycholinguistically motivated via edge prominence, while Flack (2007) argues that alignment constraints are formally grounded due to the difficulties in learning the wide range of alignment constraints from limited data. Even constraints with inarguable perceptual grounding could be formal constraints, if the constraint structure does not match the phonetic grounding; Parker (2002) and Flack (2007) argue that the sonority scale merely correlates with perceptual intensity, and requires some formal grounding to produce appropriate constraints.

The difficulty of learning the constraint set appears not only in models but as a lack of agreement among phonologists as to what the exact "right" set of markedness constraints is (McCarthy, 2008). As such, a model for learning markedness constraints is important not only as an explanation of how human learners do so but as potential evidence to adjudicate between proposed constraint definitions at a theoretical level.

The traditional OT view has been that constraints are innate and universal (Tesar &

Smolensky, 2000; Kager et al., 2004). Having innate constraints mean that differences in languages' phonologies are due solely to differences in the importance of the constraints in those languages, which simplifies the learning problem (Heinz et al., 2009). The universal constraint set also explains observed universals in phonological typology by arguing that unobserved possibilities are ruled out by the set of constraints or restrictions on their application (e.g., de Lacy, 2004).

The alternative position is the emergentist stance – that constraints are acquired in the course of learning. Constraints in this system are "phonetically grounded" (Archangeli & Pulleyblank, 1994), and arise from the formalization of articulatory or perceptual concerns, such as avoiding phoneme sequences that are difficult to articulate. Hayes (1999) provides a partial structure for such a system with a constraint-generation mechanism that relies on thresholds in the level of articulatory difficulty, and showed that most of these threshold-derived constraints make phonological sense.

There are two major theoretical points on which the stances differ. First, innatist proposals predict that the range of child phonologies is no larger than the range of adult phonologies, since they both use the same constraints. Second, existing strictly emergentist accounts predict that formal constraints must not exist (Kager et al., 2004). Although these are clear predictions, empirical evidence has been equivocal. Evaluating the first claim is difficult due to the high variation between child phonologies even of the same language (Hayes, 1999), although possible evidence of consonant harmony in child but not adult phonologies argue against it (Pater, 1997; Pater & Werle, 2003). As for the second prediction, phenomena that suggest a need for formal constraints can often be explained through clever grounded constraints (Kager et al., 2004).

Multiple learning models (Tesar & Smolensky, 2000; Boersma & Hayes, 2001; Goldwater & Johnson, 2003) have been successful in learning appropriate language-specific constraint weights with an innate constraint set, but no emergentist models are

yet able to learn a full set of markedness constraint definitions in the presence of underlying forms and faithfulness constraints. The present work replaces the fully-specified constraint set with a grammar over the building blocks of the constraints. The learned constraints are then shaped by distributional data as well as the phonological structure provided by the constraint grammar. Unlike existing strictly emergentist accounts, this model's grammar provides an avenue for the integrating formal linguistic elements necessary for potential formally grounded constraints into data-driven constraint acquisition.

3.2.3 Previous emergentist models

There are multiple frameworks for emergentist constraints, but we will limit the discussion here to two that provide fully-fleshed computational models for constraint acquisition. Hayes and Wilson (2008)'s model learns phonotactic constraints, which assess the goodness of a phoneme sequence as a word in a language. Phonotactic constraints are essentially markedness constraints, as they depend only on the output form. Phonotactic knowledge emerges early in human language acquisition, and Hayes and Wilson note that phonotactic constraints represent a potential starting point for acquiring markedness constraints.

Hayes and Wilson's model uses sequences of feature bundles to define constraints, as does our model. Feature selection in their model follows (Della Pietra et al., 1997), using greedy selection based on the ratio between the violations of a constraint by the data to the expected violations given the set of previously selected constraints. This requires a finite constraint space, imposed by a hard limit on the length of the constraint definitions. Their model obtains phonotactic constraints that largely agree with the standard theoretical constraint definitions, and make accurate acceptability predictions on both attested and novel wordforms across a variety of phonotactic phenomena.

The phonotactic learning problem presents a useful springboard to phonological

learning. It is a slightly less complex problem than the phonological learning problem, as there is no underlying form for the outputs, no alternative candidates, and no faithfulness constraints, although it is hardly an easy problem, as there is no explicit negative data a learner can use. Phonotactic learning also, as a result of its early appearance in human acquisition, may provide an initial estimate of the markedness constraints of a language.

Moving to phonological constraint learning, (Doyle et al., 2014) propose a model that learns markedness constraints in the presence of faithfulness constraints. In this model, constraint learning focuses on learning a matrix of markedness violations, with each row being a candidate output and each column a constraint. Bayesian inference is performed on the matrix, using an Indian Buffet Process prior (Griffiths & Ghahramani, 2005), which provides no phonological knowledge during constraint acquisition.

As a result, their model includes no phonological structure. Lacking phonological structure has the advantage of allowing any constraint justified by the data to be learned, including the formal constraints that have proved difficult for previous emergentist proposals. However, it also limits the model to learning extensional constraint definitions – the set of observed forms that do and do not violate each constraint. This prevents the model from making predictions about the winning candidates for novel underlying forms, and can lead to noisy estimation of low-ranked constraints. They also find that the model can learn constraints that have inconsistent meanings across different types of inputs, impeding attempts to use phonological structure to identify abstract constraint definitions post hoc. The lack of structure also means that the IBPOT model would not be easily extended to learn faithfulness constraints, as the model could reduce all the faithfulness and markedness constraints into a single faithfulness constraint that penalizes all unattested unfaithful input-output pairs, and a single markedness constraint that penalizes the output forms in unattested faithful input-output pairs.

Our model extends beyond each of these previous models. It learns markedness

constraints in the presence of faithfulness constraints, unlike the phonotactic learner. It also learns over an infinite space of constraint definitions. By incorporating a grammar of constraints, its constraints correspond to real phonological structure, unlike those of the phonologically-unmotivated IBP. This grammar also means that the constraints are intensionally defined, and can be used to predict the likely output forms for novel inputs.

3.3 Model design

3.3.1 General structure

The phonological model consists of two constraint matrices, the observed F (faithfulness) and the unobserved M (markedness), as well as a vector of constraint weights w. The cells of the violation matrices correspond to the number of violations of a constraint by a given input-output mapping. F_{ijk} is the number of violations of faithfulness constraint F_k by input-output pair type (x_i, y_j); M_{jl} is the number of violations of markedness constraint $M_{\cdot l}$ by output candidate y_j (independent of the input x). For each input x_i, some subset of the output forms $\{y_j\}$ are generated by GEN as candidates; this subset will be denoted $\mathscr{Y}(x_i)$. The weight vector w provides weights for both F and M, and probabilities of the output candidates are given by a log-linear function:

$$p(y_j|x_i) = \frac{\exp\left(\sum_k w_k F_{ijk} + \sum_l w_l M_{jl}\right)}{\sum\limits_{y_z \in \mathscr{Y}(x_i)} \exp\left(\sum_k w_k F_{izk} + \sum_l w_l M_{zl}\right)} \tag{3.1}$$

F is assumed to be known, as part of the process of generating candidates (Riggle, 2009). M is a non-parametric matrix with a known number of rows (candidates) but an unknown number of columns (constraints). The number of columns is generated by a Poisson prior with parameter α, and the violations within each column are generated from a grammar G of constraints, using a Rational Rules model with parameter b, which will be discussed in detail in the next section. w is a vector of exponentially distributed

weights with parameter η_G for both constraint types, and is also learned. Within OT, constraint violations always penalize the violator, so $w < 0$ for all constraints.

Equation 3.1 defines the probability of a candidate y_j being chosen as the output for a single instance of an underlying form x_i, given the current estimates of the constraints and weights. To get the probability of the observed data as a whole, we take the product of the probabilities across all observed input-output pairs. Letting n_{ij} be the number of times that y_j is observed as the output form for x_i, and $n_{i\cdot}$ be the number of times the underlying form x_i occurs, the probability of the whole dataset Y is:

$$p(Y|M,F,W) \propto \prod_i \frac{\exp\left(\sum_{jk} w_{Fk} n_{ij} f_{ijk} + \sum_{jl} w_{Ml} n_{ij} m_{jl}\right)}{\left(\sum_{y_z \in \mathscr{Y}(x_i)} \exp\left(\sum_k w_{Fk} f_{izk} + \sum_l w_{Ml} m_{zl}\right)\right)^{n_{i\cdot}}} \tag{3.2}$$

The probability in Equation 3.2 also indicates how well the learned structure explains the data Y, and this value will be one metric by which the model's learning is assessed in Sect. 3.5.1. The last important probability in this model is the joint probability of the learned structures (M and w) given the priors and known variables (Y and F):

$$p(M,w|Y,F,\alpha,b,\eta_G,G) \propto p(Y|M,F,w)p(M|b,G)p(w|\eta_G) \tag{3.3}$$

This joint probability is proportional to the product of the probabilities of the data (Eqn. 3.2), weights (exponential prior), and matrix M (defined in the following secitons). The learning method described below approximates draws from this joint distribution, meaning that the values of M and w learned by the model will be based on trade-offs between improved data, matrix, and weight probabilities. This will also be used as a metric to assess the quality of the learned structure in Sect. 3.5.1. The complete specification of the model is then:

$$M \sim RR(G,b); \qquad \mathscr{Y}(x_i) = Gen(x_i)$$
$$w \sim -\exp(\eta_G); \quad y|x_i \sim LogLin(M,F,w,\mathscr{Y}(x_i))$$

3.3.2 Constraint grammar and violation profiles

Phonological knowledge in this model comes from the constraint grammar. Different constraint grammars may be chosen to test their learnability and typological consequences; this will be examined in more depth in Sect. 3.6. At present, we use a simple probabilistic context-free grammar (PCFG) to generate sequences of feature-value bundles, similar to the phonotactic constraints of Hayes and Wilson (2008). The PCFG has the benefit of imposing a bias toward simpler constraint definitions (i.e., those spanning fewer phonemes and specifying fewer feature-value pairs). The specific grammar used in the experiment is discussed in Sect. 3.4.

The markedness matrix M is not defined directly by constraint definitions, but by violation profiles, which are binary vectors whose values are zero for candidates that do not violate the constraint and one for those that do. The violation profiles are generated from constraint definitions using the Rational Rules framework (Goodman et al., 2008). Binary violations are assessed from the constraint definition (the binarization method is discussed in Sect. 3.4), but candidates can be exceptions (switching the constraint value from zero to one or vice versa), with the number of exceptions exponentially distributed. Given a constraint definition d, the probability of it producing a violation profile m is given by:

$$p(m|d) \propto exp(-bQ(m;d)) \tag{3.4}$$

where $Q(m;d)$ is the number of exceptions in m given d, and b is a model parameter, with larger b penalizing exceptions more strongly. Usually, phonological constraints

are assumed to have no exceptions; but there are two related reasons to allow them within the model. First, allowing exceptions improves learnability; the learner does not have to jump directly to the right definition for a profile, but can gradually improve its estimate. Second, exceptions allow the learner to learn a constraint that does not exactly match anything in the constraint grammar. This presents a potential avenue for learning language-particular constraints, if they exist (J. Smith, 2004).

We marginalize over possible constraint definitions to estimate the probability of a violation profile; the estimation method is discussed in Sect. 3.3.4.

3.3.3 Inference on M and w

For the model to find good phonological systems, we perform Markov Chain Monte Carlo (MCMC) inference over the space of markedness matrices M and weight vectors w. Five MCMC sampling moves are used to explore this space.

Adding columns to M The model is initialized with randomly-drawn markedness violation matrix M and weight vector w.[3] Potential new columns for M are theoretically drawn from the infinite set of possible constraint profiles; to make this tractable, we follow Görür et al. (2006) and perform a truncated Bernoulli draw over at most K^* new constraints from an auxiliary matrix M^* with weight vector w^*. The columns of this auxiliary matrix are drawn in the same way as the initial values for M, by drawing constraint definitions with exceptions. This approximation retains the unbounded nature of the possible constraint space, as repeated sampling iterations can add constraints without limit.

The model chooses what columns M_+, if any, to add from M^* as a multinomial draw over the joint probabilities $p(M_+ \cup M, w_+ \cup w | Y, F, \alpha, b, \eta_G, G)$ for each subset

[3] M is drawn by a Poisson on the number of constraints, and each profile is a Rational Rules draw from a constraint definition drawn from the constraint grammar.

$M_+ \subseteq M^*$. Because the null subset is an option, the model can decide to add no new columns.

Gibbs sampling on M The values of the cells M_{jl} are resampled conditioned on the rest of the constraint violations $(M_{-(jl)}, F)$ and the weights w. The probability that $M_{jl} = z$ (where z is 0 or 1) is proportional to the product of the conditional probability $p(M_{jl} = z | m_{-jl})$ and the likelihood of the data with the new value of M. The conditional probability is calculated by marginalizing over all possible constraint definitions, as discussed in Sect. 3.3.4.

Removing columns from M After Gibbs sampling, each column of M is considered for removal; this is essentially the reverse of the first MCMC sampler, which tried to add columns. The probability of removing a column from M to yield the smaller matrix M_- is a Bernoulli draw with probability:

$$\frac{p(M_-, W_- | Y, F, \alpha, b, \eta_G, G)}{p(M_-, W_- | Y, F, \alpha, b, \eta_G, G) + p(M, W | Y, F, \alpha, b, \eta_G, G)} \tag{3.5}$$

Before any columns are removed, the auxiliary matrix M^* is deleted. Any columns that are removed is placed into M^*, along with their weights, and new columns/weights are drawn from the constraint grammar to fill the auxiliary to K^* columns.

Splitting columns within M The final sampling step on M is to consider moving a set of violations from one column into another, either an existing column or a new column that has no other violations. This move allows for violations that substantially improve the data likelihood but are exceptions to the constraint definitions that the rest of the column's violations favor to move into a column where they fit better. Without this split move, the violations would have to first be removed via Gibbs sampling, then re-added,

again via Gibbs sampling, in the new column. The first Gibbs sampling change may be very unlikely due to the loss in data likelihood, slowing down convergence.

The proposed split and its acceptance probability are drawn as follows. The likelihood of a violation M_{JL} being an exception within its profile $M_{J\cdot}$ is estimated from the proportion of samples from $p(d|M_{J\cdot})$ that mark the violation as exceptional. The set of violations V to be moved is drawn as a sequence of independent Bernoulli draws based on each violation's likelihood of being an exception. More commonly exceptional violations are more likely to be moved. The exception likelihood is smoothed using a Beta-binomial distribution with parameter β, by taking the maximum a posteriori estimate of the likelihood:

$$p(m_{JL} \in V) = \frac{N_E + \beta - 1}{N_E + N_N + 2\beta - 2} \tag{3.6}$$

The number of Metropolis samples in which M_{JL} was an exceptional violation is N_E and a non-exceptional violation is N_N. The β parameter functions as a pseudocount; the model estimates the exception likelihood as if it had seen β additional examples of the violation being exceptional and β additional examples of it being unexceptional, moderating the probability estimates. Increasing β increases the smoothing, and the effect of the smoothing decreases as more Metropolis samples are drawn.

Weight sampling Between each of the matrix samplers, we use Metropolis-Hastings to estimate new weights for both constraint matrices. Our proposal distribution is $Gamma({w_k}^2/\eta_M, \eta_M/w_k)$, with mean w_k and mode $w_k - \frac{\eta_M}{w_k}$ (for $w_k > 1$). This adds one free parameter, η_M, which increases the variance in proposal weights as it increases.

3.3.4 Inference over the constraint definitions

Many of the inference steps on M require knowledge of the prior over violation profiles (i.e., columns of M), but this is a sum over the infinite set of constraint definitions. To estimate this, we use importance sampling over constraint definitions. The underlying idea is that for most definitions d, at least one of the probabilities $p(m|d)$ and $p(d)$ is negligible. We use Metropolis sampling to move through the region of constraint definitions d where both probabilities are relatively high.

For a given profile m, we start by drawing a constraint definition d from the PCFG, then Metropolis-Hastings sample through the space of constraint definitions, with three possible move types, of equal probability. The first move is subtree-replacement, where a node in the tree is re-drawn, along with all its children. This is the Metropolis move used by Goodman et al. (2008). To improve mixing, we add two more moves: excision and insertion. Excision chooses a node N and one of its grandchildren G, and removes the intervening parent node (if and only if the G is a valid child of N according to the CFG). Insertion performs the opposite task, inserting a node between N and one of its children. In terms of the constraint definitions, subtree-replacement changes a feature value, a bundle, or a sequence of bundles; excision removes a feature, bundle or sequence of bundles; and insertion adds a feature, bundle, or sequence of bundles.

For subtree-replacement, the node being re-drawn is selected uniformly randomly from all non-terminals in the current tree; this is always possible for at least one node. That node's label is retained, but all of its children are removed and re-drawn according to the PCFG rules. If a subtree replacement is to be made, the probability of moving from tree T to T' by redrawing the subtree S_X at node X is:

$$J_R(T';T) = \frac{1}{N_R} \cdot \prod_{r \in S'_X} p(r), \tag{3.7}$$

where N_R is the number of non-terminal nodes in T, S'_X is the new subtree with root X, and r ranges over the rules in S'_X.

Node selection for excision is slightly more complicated, as it selects a node X uniformly randomly from the set of nodes that can be excised (nodes with at least one grandchild Z that is also a valid child of X under the CFG). If no excisable nodes exist in the tree, the model attempts a different move type (replacement or insertion) instead. Excision removes a node Y – the child of X and parent of Z – from the tree, as well as the current sibling of Z (with its subtree).

If an excision is to be made, the probability of choosing to excise between X and Z in tree T to yield tree T' is:

$$J_E(T';T) = \frac{1}{N_E} \cdot \frac{1}{N_{E;X}}, \tag{3.8}$$

where N_E is the number of nodes in T that have at least one excisable grandchild, and $N_{E;X}$ is the number of excisable grandchildren of X in T.

Insertion is the most complicated move. Node selection for insertion works similarly to excision; a node is drawn uniformly randomly from the set of insertable nodes. An insertable node is one that has at least one child that could also be its grandchild. As with excision, if no insertable nodes exist, a different move type is attempted. Once an insertable node X is chosen, the model chooses a child node Z uniformly among its children that could be a grandchild of X. That node becomes a grandchild of X, and the model draws a new node Y, according to the PCFG, such that Y is a valid child of X, a valid sibling of A, the remaining child node of X, and a valid parent of Z. Finally, Z needs a sibling B in its new lower position, and this node (along with its subtree) is drawn according to the PCFG. Given that an insertion is to be made to the tree T, the

probability of that insertion being node Y between X and Z is:[4]

$$J_I(T';T) = \frac{1}{N_I} \cdot \frac{1}{N_{I;X}} \cdot \frac{p(X \to AY)}{p(X \to A*)} \cdot \frac{p(Y \to ZB)}{p(Y \to (Z* \mid *Z))} \cdot \prod_{r \in S_B} p(r), \tag{3.9}$$

where N_I is the number of nodes in T that have at least one insertable child, and $N_{I;X}$ is the number of insertable children of X in T. The third fraction is the probability of choosing Y as the new child in T', and the fourth fraction is the probability of choosing B as the new sibling of Z, as well as whether Z is the left- or right-hand child of Y. The final term is the probability of the subtree S_B.

Using the jump probabilities between trees given by the above equations, we can calculate the acceptance probability of a possible Metropolis move from T to T'. This is the product of the ratio of the forward and backward jump probabilities and the ratio of the trees given the current violation profile m:

$$\frac{p(m|T')p(T')J(T;T')}{p(m|T)p(T)J(T';T)} \tag{3.10}$$

The Metropolis method samples constraint definitions $\{d^{(1)}, \cdots, d^{(n)}\}$ from the posterior distribution $p(d|m)$. These samples can be used to estimate the probability of the violation profile m given the constraint grammar G by taking the harmonic mean of $p(m|d^{(t)})$ over all samples (Newton & Raftery, 1994).[5] This provides a prior for the columns of the matrix; coupled with the Poisson prior on the number of columns, we have a prior over matrices with an indefinite number of columns.

[4]This equation assumes that Z is the right-hand child of X and the left-hand child of Y. If Z is the left-hand child of X or the right-hand child of Y or both, the probability is calculated similarly, but the third or fourth fraction changes to reflect the actual structure.

[5]Because harmonic mean estimation can be noisy and take a large number of iterations to converge (Neal, 1994), we tested a range of violation profiles and found consistent convergence to the expected constraint definitions and profile probabilities within a few thousand samples.

3.4 Experiment

We test this model by learning the constraints that produce the English regular plural system. This requires the model to identify two markedness constraints, plus weights for three additional faithfulness constraints.

3.4.1 English regular plural morphophonology

The English regular plural has one underlying form (/z/) with three attested output realizations: [z], [s], or [əz] (as in *hugs, huts*, and *hushes*, respectively). Two markedness constraints drive this alternation in the standard phonological analysis: AGREE[VOI] and *[+STR][+STR]. The former penalizes outputs where consecutive consonants do not agree in voicing, and the latter penalizes outputs where consecutive consonants are both strident.[6] These are coupled with three faithfulness constraints: MAX, DEP, and IDENT[VOI]. These penalize removing a phoneme, adding a phoneme, or changing the voicing of a phoneme.

For this experiment, we consider four candidate outputs for each input: the bare singular form, plus forms with each of the three attested allophones of the regular plural suffix. The candidates for plural *hug* (underlying /hʌgz/), for instance, are [hʌg], [hʌgz], [hʌgs], or [hʌgəz]. In general, the [əz] candidate wins only when the singular ends in a strident, the [s] candidate wins only when the singular ends in a voiceless non-strident, and the [z] candidate wins the rest of the time.

Traditional OT predicts four necessary pairwise rankings in the constraint hierarchy. First, *[+STR][+STR] and MAX must both outrank DEP, so that consecutive stridents are resolved by adding a phoneme (ə) to separate them, rather than deleting the suffix. Second, AGREE[VOI] and DEP must both outrank IDENT[VOI], so that voicing

[6] Voiced English consonants are b,d,g,v,ð,ʒ,dʒ,m,n,ŋ,l,r,w; voiceless consonants are p,t,k,f,θ,ʃ,tʃ. English stridents are s,z,ʃ,ʒ,tʃ,dʒ.

disagreements are resolved by changing the voicing of the suffix, rather than adding a intervening phoneme. (Since rankings are transitive, this implies *[+STR][+STR] and MAX both outrank IDENT[VOI] as well.)

In terms of constraint weights, the necessary rankings imply large weight differences between constraints. The estimated MaxEnt OT weights for the constraints are approximately 15 for *[+STR][+STR] and MAX, 10 for AGREE[VOI] and DEP, and 5 for IDENT[VOI], showing large gaps corresponding to the critical rankings.

For a training set, the model is exposed to the plural forms of 26 common count nouns. These nouns were chosen using norming data from Dale and Fenson (1996), with all of the nouns being reported as understood by at least 89.4% of tested 18-month-old English learners.[7] This suggests these words are common in child-directed speech, and would likely be used in acquiring the English plural. The model is given 100 examples of each plural, always using the standard pluralization.

3.4.2 The constraint grammar

Constraint definitions are similar to the phonotactic constraints of Hayes and Wilson (2008), and represent sequences of phonological feature bundles. The phonological features mark different characteristics of phonemes; for instance, the phoneme [s] has phonological features including [+consonantal, +coronal, +strident, –voiced], while the similar phoneme [z] has features including [+consonantal, +coronal, +strident, +voiced]. A feature bundle within a constraint definition matches all phonemes with all of the bundle's features. Thus [+consonantal, +strident] matches [s] and [z] but [+consonantal, +voiced] matches only [z]. Phoneme-to-feature mappings are based on Riggle (2012).[8]

[7] The words are: *baby, ball, balloon, banana, bath, bird, blanket, book, car, chair, daddy, diaper, door, drink, eye, hug, key, kiss, kitty, mommy, nap, nose, phone, shoe, spoon, toothbrush*

[8] There is one exception to this: voicing is not specified on sonorants, because sonorants do not have voiceless versions and do not trigger AGREE[VOI] violations. Sect. 3.6 discusses how such decisions about the core structure of the model can inform our view of linguistic structure.

Table 3.1. Rules within the phonological context-free grammar. C is the root symbol, and expands into one or more feature bundles B (a set of feature-value pairs that must be satisfied by a phoneme). B obligatorily splits into the bundle b and S, which can contain a Kleene star. b expands into one or more pairs P, each of which contains a value ($+/-$) and a phonological feature (one of the 23 phonological features). ε indicates an empty node.

Non-terminal expansions			Terminal expansions		
C	$\rightarrow$	$BB \mid BC \mid B$	S	$\rightarrow$	$* \mid \varepsilon$
B	$\rightarrow$	bS	V	$\rightarrow$	$+ \mid -$
b	$\rightarrow$	$PP \mid Pb \mid P$	F	$\rightarrow$	$voice \mid syllabic \mid strident \mid \cdots$
P	$\rightarrow$	VF			

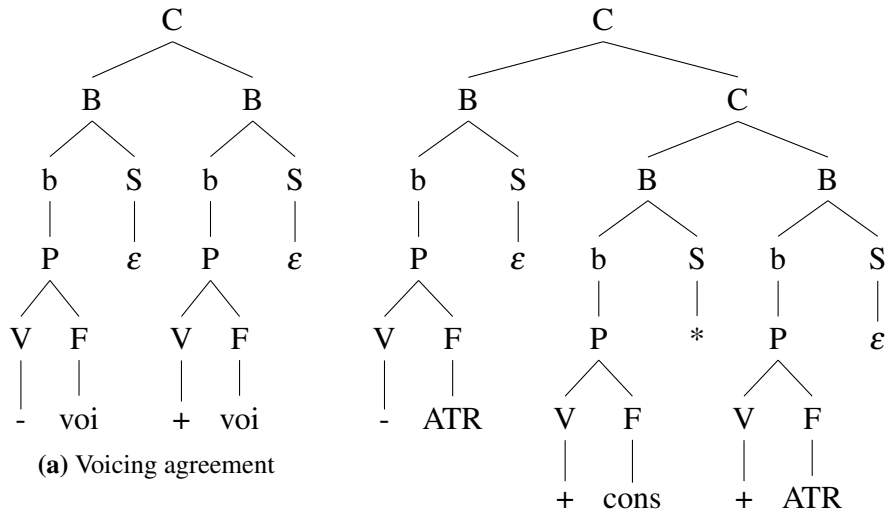
Lastly, this grammar also allows the generation of a Kleene star, which matches zero or more occurrences of a feature bundle. None of the English plural constraints require this, but it is necessary for future work generalizing the grammar to capture long-distance constraints, such as vowel harmony.

The bundle sequences are generated by a probabilistic context-free grammar with uniform rule probabilities, shown in Table 3.1, with example constraint trees in Figure 3.1. This defines an infinite space of constraint definitions, with no limit on the number of features per bundle nor the number of bundles per sequence. Of these, three definitions correspond to the phonologically standard constraints. The definition for *[+STR][+STR] is self-evident, but AGREE[VOI] is actually an umbrella for two constraints in the PCFG: *[-VOI][+VOI] and *[+VOI][-VOI]. As we will see, only the first of these agreement constraints is necessary to explain the English plural, though the second could be learned if other paradigms were added (see Sect. 3.6).

Note that the constraints, defined as sequences of feature bundles, are not binary; a candidate can contain multiple sequences that violate a constraint. However, the Rational Rules method requires binary constraints. To obtain binary versions of the constraints, we subtract the number of violations incurred by the singular form from the number of violations incurred by the candidate. Since all the candidates from a given input share the

Table 3.2. The 23 phonological features, and their values over English phonemes. Some phonemes are unspecified for a feature (neither + nor −), and will not match feature bundles that specify a value for that feature. Phoneme feature values are based on Riggle 2012, with the exception that sonorants, which are obligatorily voiced, are unspecified for voicing.

Feature	+ Phonemes	− Phonemes
ATR	i u e o	a ə ɛ ɑ æ ɪ ɔ ʊ ʌ
anterior	d l ð n s r θ z t	ʃ ʒ tʃ dʒ
back	ɑ ə g k ɔ ŋ ʊ o w ʌ u	a tʃ dʒ e æ i j ɪ ɛ
closed glottis	ʔ	æ ɑ ə tʃ ɛ ð ɪ dʒ ɔ ŋ ʃ ʊ θ ʌ ʒ a b e d g f i h k j m l o n p s r u t w v z
coronal	tʃ z d ʃ dʒ l n s r θ t ʒ ð	ɑ ə j ɛ ɪ ɔ ŋ ʊ ʌ a b e g f i h k æ m o p u w v ʔ
consonantal	tʃ ð dʒ ŋ ʃ θ ʒ b d g f k m l n p s r t v z	ɑ ə ɛ ɪ ɔ ʊ ʌ a e æ i h j o u w ʔ
continuant	ɑ ə j ɛ ð ɪ ɔ ʃ ʊ θ ʌ ʒ a e f i h æ l o s r u w v z	b d g k m n p t ŋ ʔ
distributed	ʃ ʒ tʃ dʒ	d l ð n s r θ z t
dorsal	ɑ ə tʃ ɛ ɪ dʒ ɔ ŋ ʊ ʌ a e g æ i k j o u w	ð ʃ θ ʒ b d f h m l n p s r t v z ʔ
delayed release	tʃ dʒ	æ ʔ ɑ ə ɛ ð ɪ ɔ ŋ ʃ ʊ θ ʌ ʒ a b e d g f i h k j m l o n p s r u t w v z
high	tʃ dʒ g i k j ŋ ɪ u w ʊ	a ə ɛ æ ɔ ɑ o ʌ e
labial	ɑ ə j ɛ ɪ ɔ ʊ ʌ a b e f i æ m o p u w v	tʃ ð dʒ ŋ ʃ θ ʒ d g h k l n s r t z ʔ
lateral	l	æ ʔ ɑ ə tʃ ɛ ð ɪ dʒ ɔ ŋ ʃ ʊ θ ʌ ʒ a b e d g f i h k j m o n p s r u t w v z
low	a ɑ æ	ə tʃ ɛ ɪ dʒ ɔ ŋ ʊ ʌ e g i k j o u w
nasal	ŋ m n	æ ʔ ɑ ə tʃ ɛ ð ɪ dʒ ɔ ʃ ʊ θ ʌ ʒ a b e d g f i h k j l o p s r u t w v z
pharyngeal	a ə ɛ ɑ æ i ɔ ɪ u o ʌ ʊ e	tʃ ð dʒ ŋ ʃ θ ʒ b d g f h k j m l n p s r t w v z ʔ
round	ɑ ə ʊ ɔ w u j o	a ʌ b e f i æ m p v ɪ ɛ
spread glottis	h	æ ʔ ɑ ə tʃ ɛ ð ɪ dʒ ɔ ŋ ʃ ʊ θ ʌ ʒ a b e d g f i k j m l o n p s r u t w v z
sonorant	ɑ ə ɛ ɪ ɔ ŋ ʊ ʌ a e æ i j m l o n r u w	tʃ ð dʒ ʃ θ ʒ b d g f h k p s t v z ʔ
strident	tʃ ʒ dʒ s ʃ z	ɑ ə j ɛ ð ɪ ɔ ŋ ʊ θ ʌ a b e d g f i h k æ m l o n p r u t w v ʔ
syllabic	a ə ɛ ɑ æ i ɔ ɪ u o ʌ ʊ e	tʃ ð dʒ ŋ ʃ θ ʒ b d g f h k j m l n p s r t w v z ʔ
tense	i e u o æ	a ə tʃ dʒ ɛ ɑ g ɪ k j ɔ ŋ ʊ w ʌ
voice	ɑ ə ɛ ð ɪ dʒ ɔ ŋ ʊ ʌ ʒ a b e d g æ i j m l o n r u w v z	tʃ f h k s p ʃ t ʔ θ



(a) Voicing agreement

(b) Assamese vowel harmony

Figure 3.1. Example tree-structures within the constraint CFG. Tree (a) defines a voicing agreement constraint, [-VOICE][+VOICE], that is important for the English plural. Tree (b) gives an example of a more complex constraint, a vowel harmony constraint from Assamese (Mahanta 2012), [-ATR][+CONS]*[+ATR], which uses the Kleene star to ignore any intervening consonants.

singular form as a stem, the same number of violations are subtracted from all of them, and the candidate probabilities within the log-linear model are unchanged.[9]

3.4.3 Model parameters

The model has three free parameters with theoretical importance, which we set as follows. The Poisson parameter α and the weight parameter η_G are set to 1, to encourage learning smaller matrices and weights. b, the exception parameter in the Rational Rules distribution, is set to 10, to discourage exceptions.

The sampling methods have their own free parameters, but these lack theoretical importance. The Metropolis weight parameter η_M is 1 as a default. K^*, the number of

[9]For a very small number of possible constraint definitions (e.g., *[+continuant]), the [əz] allomorph incurs multiple violations, but none of these constraints are important to the English plural, nor identified as likely definitions for violation profiles learned by the model in practice. Alternative implementations for non-binary constraint representation are discussed in Sect. 3.6.3.

auxiliary parameters, is 7, large enough to encourage movement through the space of M while remaining easily calculable. β, the smoothing parameter for estimating the exceptionality of violations, is 100. In the future, these will be tested over a wider range, but changing them should only affect convergence rates.

We run the model for 200 iterations in three trial runs, with deterministic annealing on the first half of the iterations to encourage mixing. For estimating $p(m)$, 1000 burn-in samples are taken and discarded, and 250 additional samples (every second sample out of 500 to reduce autocorrelation) of d are averaged. Each time $p(m)$ is re-calculated for a given m, half as many additional samples are drawn (e.g., 125, 62, ..., down to a minimum of 25) and incorporated into the average. This allows continued improvement of the $p(m)$ estimates for even common violation profiles without overly costly runtimes.

3.5 Results

We test the model's performance in three areas: the ability of the learned constraints and weights to explain observed data, their ability to predict the plural forms of unobserved words, and the interpretability of the constraints. In all cases, we compare the learned constraints to the baseline of the standard constraint definitions that an innate constraint set would provide.

To assess the performance of an innate constraint set, we set M to be the two standard English markedness constraints, *[+STR][+STR] and *[-VOI][+VOI] (with no exceptions), and learn weights for them using the same weight learning mechanism as in the acquisition model. The same mechanism is used so that the learned-constraint weights and innate-constraint weights (and their consequences) can be directly compared.[10] We use the same parameters to estimate this baseline as in the RROT learning model.

[10]Conceptually, this is equivalent to learning weights by MaxEnt OT, except that our model uses a Metropolis sampling rather than gradient ascent to obtain its weight estimates.

Table 3.3. Data, markedness matrix, weight vector, and joint log-probabilities for the learned and innate constraints. MAP and mean estimates over the final 100 iterations for the three model runs. Values closer to zero indicate better model fits.

	MAP		Mean	
	Learned	Innate	Learned	Innate
Data	-2.94	-5.16	-7.91	-10.6
M	-71.0	-41.3	-77.8	-41.8
w	-45.9	-56.5	-52.6	-64.7
Joint	-129.5	-114.2	-138.4	-117.1

3.5.1 Observed forms

The first goal of the model is to find constraint structures that can explain the observed plural forms within the log-linear model, establishing that the learning method is able to identify useful structures. The model satisfies this goal if it assigns high probability to the observed forms, especially as compared to the innate constraints.

We quantify this by calculating the joint probability of the data and model given the faithfulness constraints and all model parameters, $p(Y,M,w|F)$, which is proportional to the product of three terms: the data likelihood $p(Y|M,F,w)$, the markedness matrix probability $p(M|\alpha)$, and the weight probability $p(w)$.[11] We calculate both the mean and MAP (maximum probability) values of these probabilities over the final 100 iterations of each of the three model runs, and report the across-run means in Table 3.3.

Overall, the probabilities are comparable for the learned and innate structures. The data probability in the learned model is especially high. These results indicate that the model is capable of learning constraints that provide a reasonable structure for the observed plural forms, and we will look at the exact form of the learned constraints in Sect. 3.5.3. The lower matrix probabilities of the learned M are reflective of the difficulty of the search problem, which must sift through an infinite space of constraints.

[11] Though the innate runs have known constraint definitions, the reported matrix probability is marginalized over possible constraint definitions to make the distributions comparable between the innate and learned runs.

None of the differences are significant by t-tests, although the statistical power is weak with relatively few test runs. Even if some of these differences are statistically significant with more runs, the crucial result is that learning constraints results in at worst a small drop in explanatory power. That loss, if it exists, is driven by the decrease in matrix probability. Such a probability decrease would not be surprising, given the difficulty of efficiently searching through the space of possible matrices, and the cost of possible exceptions (which the innate constraints do not have).

3.5.2 Predictive behavior

The second test of the model's acquisition is how well it predicts plurals for newly-encountered words. This is a crucial feature for human acquisition; children quickly learn to generalize morphophonological systems. It also represents an important modeling step, as the existing IBPOT learner is incapable of making such predictions due to its strictly-extensional constraint definitions.

We test the model's prediction abilities by pluralizing common words that the model has not already seen. The test set is the 25 most frequent countable nouns in the Corpus of Contemporary American English (COCA; Davies, 2008) that take regular plurals.[12] These range in frequency from 400 to 1700 occurrences per million COCA words, so they are common for adult speech, but most are uncommon in child-directed speech. Two take the [əz] allomorph, seven take [s] and the rest are [z]. Five of these nouns end with phonemes that did not occur word-finally in the observed data.

To assess the predictive power of the learned constraints, we need constraint definitions, but the learning model marginalizes over definitions. We obtain definitions by using the $p(d|m)$ Metropolis sampler to generate a distribution over definitions for each violation profile. For a new candidate y, the probability that y violates a constraint

[12]These words are: *time, year, way, day, thing, world, school, state, family, student, group, country, problem, hand, part, place, case, week, company, system, program, question, government, number, night*

Table 3.4. Mean and lowest probabilities for the correct plural forms over 25 previously unobserved words, averaged over three runs. In all runs, the correct plurals had the highest probabilities of any candidates.

Probabilities	Learned	Innate
Mean	.995	.991
Lowest	.986	.974

with profile m is estimated as the proportion of constraints d drawn from the $p(d|m)$ Metropolis sampler where y is violated. The score of the candidate is then calculated as the weighted sum of these probabilities – so if y had a probability of 0.2 of violating a constraint with weight -10, this would contribute -2 to its score. Candidate probabilities for novel forms are calculated as for the observed forms, aside from this probabilistic scoring. Violation profiles are taken from the final iteration of each model run.

We assess the predictive quality of the learned structures in three ways. The first is the proportion of novel inputs where the correct candidate is predicted to be the most likely winner. For all 25 inputs, every run of the learned and innate models choose the correct output form.

The second and third assessments are based on the log-probabilities of the correct output form; we consider the mean probability of the correct candidate for each input, and lowest predicted probability of any correct candidate. Again, both versions are very successful at identifying the correct output candidate predictively (Table 3.4), averaging above 99% of the probability mass being placed on the correct outputs, and never dropping below 97%.

Overall, we find that the model's learned constraint structure is extremely accurate at predicting the plural forms of previously unseen nouns, and at least as accurate as the innate constraints. Thus the constraint acquisition model not only finds an explanatory structure for observed forms, but finds an appropriately abstract structure that it can extend to correctly predict novel plurals. This represents a major extension from the

extensionally defined constraints of Doyle et al. (2014).

3.5.3 Violation Profiles and Constraint Definitions

Having found that the learned constraints can explain observed data and extend to novel forms, we now want to know if the constraints themselves are as interpretable as the innate constraints. To do this, we compare the violation profiles and constraint definitions that the model learns to the innate constraints.

The first step is to compare the number of learned constraints to the number of innate constraints to see if the model explains the data as efficiently as the innate proposal. This is generally the case; two of the model runs have two markedness constraints consistently over the final 100 iterations. The third model run consistently uses four markedness constraints over its final 100 iterations. These extra two constraints appear to be superseding two of the faithfulness constraints (DEP and IDENT[VOI]), whose weights drop to near zero in this run. This means that all three runs have five active constraints, just as the innate structure has.

The second step is to compare the violation profiles of the learned and innate constraints.[13] We begin by assessing what proportion of candidates the learned and innate profiles agree on the presence or absence of a violation. Over the final 100 iterations, the learned violation profiles agree with their corresponding innate profiles on an average of 98.9% of all candidates.[14]

The final step is to compare the distributions over constraint definitions, as

[13] We limit this and the next analysis to the solutions with two markedness constraints. This is done in part because the four-constraint solution lacks immediate correspondences between the innate and learned constraints (one of its constraints corresponds to AGREE[VOI], but the rest are conflated with faithfulness). But it is also done because the four-constraint solution represents a local optimum, and more effective searching of the space of markedness matrices, such as a parallel-chain method, would move away from the four-constraint soltuion to the two-constraint solutions.

[14] To determine which learned profile corresponds to which innate profile, we compare all possible links and choose the system with the least overall disagreement where each learned profile corresponds to a different innate profile.

Table 3.5. The three most probable definitions (based on $p(d|m)$) for the innate violation profiles and the last iteration of each run of the learned profiles, based on 10000 Metropolis samples. The first row shows likely definitions for the agreement-like constraint; the second row shows likely definitions for the double-strident constraint. The top definition in each row is the most likely.

Innate	Learning Run 1	Learning Run 2
[-VOI][+VOI]	[-VOI][+VOI]	[-VOI,-STR][-SG,-HI]*[-NAS,+VOI]
[-VOI][+VOI,+ANT]	[-VOI][+VOI,-CG]	[-VOI,-STR][-STR,+RND]*[+VOI]
[-VOI][+VOI,+COT]	[-VOI][+VOI,+ANT]	[-VOI,-STR][+B,+PHR]*[+COR,+VOI]
[+STR][-SYL]	[+STR][-SYL]	[+STR][-SYL]
[+STR][+STR]	[+STR][-LAB]	[+STR][-DOR]
[+STR][-LAB]	[+STR][-DOR]	[+STR][+ANT]
Key: voi=voicing, str=strident, sg=spread glottis, hi=high, nas=nasal, ant=anterior,cg=closed glottis, rnd=round, cot=continuant, b=back, phr=phrasal, cor=coronal, dor=dorsal, lab=labial, *=Kleene star.		

estimated by the Metropolis sampler for $p(d|m)$ for the learned and innate profiles. Table 3.5 shows the three most likely constraint definitions for each violation profile, for both the innate violation profiles and the two two-constraint runs of the acquisition model. The top row in Table 3.5 corresponds to the AGREE[VOI] constraint, and the bottom corresponds to *[+STR][+STR].

The innate markedness matrix contains two constraints, which were required to match the *[-VOI][+VOI] and *[+STR][+STR] constraint violations exactly. However, in addition to these standard definitions, other definitions can have the same profile, due to the large number of phonological features. The Metropolis estimation of $p(d|m)$ for the innate violation profiles reflects this. In the top-left cell of Table 3.5, we see that the standard *[-VOI][+VOI] definition is the most likely definition for its violation profile, but that slightly more complicated variants are likely as well. All three of the possible definitions are consistent with the violation profile of *[-VOI][+VOI], and the grammar defines a prior favoring less complex constraints, so *[-VOI][+VOI] is the most likely definition.

The bottom-left cell of Table 3.5 shows that the standard definition may not be the most likely. All three definitions in this cell match the violation profile of the standard *[+STR][+STR] definition on the observed forms, and because they are all equally likely in the constraint grammar, the model is unable to distinguish between them. This problem of multiple solutions arises because of the limited set of possible affixes; a stem ending in a strident violates this constraint with the [z] or [s] affix, but not the [əz] affix. Because there are only the three affixes, the data does not tell the model whether it is the non-stridency of [ə] or, say, its sonorance or non-coronality that makes it acceptable.

This uncertainty over definitions is an interesting result, as it encourages simultaneous learning of multiple phonological phenomena; the *[+STR][-SYL] definition, for instance, could be ruled out by observing the faithful manifestation of s-initial onset clusters in English, as in *stop* or *spin*. Cross-phenomenon learning in this case would ease the constraint-recognition problem rather than making it harder.

The likely constraint definitions for the innate markedness matrix represent an upper bound on expected constraint recognition in the learning model. Turning to the constraints learned in the first run of the learning model, we see an essentially identical set of likely constraint definitions. This reflects the close agreement of the learned and innate markedness matrices in this first run; over the final 100 iterations of this run, the two matrices agreed on over 99.5% of their violations. This run considers the standard definition for the voicing-agreement constraint the most likely, and encounters the same uncertainty about the exact identity of the consecutive-stridents constraint as the innate model.

The second run's likely definitions for the voicing-agreement constraint are noticeably more complex, although its consecutive-stridents definitions are similar to the other runs. This constraint's violation profile is close, but not identical, to the innate

violation profile; over the final 100 iterations, it matches on 92.3% of the violations. This difference causes the model to propose a constraint definition using the Kleene star, which matches any number (including zero) of instances of a feature bundle, which is important for constraints like vowel harmony that accept any number of interceding consonants.

In the English plural, the Kleene star seems out of place, but in fact it shows a clever solution. Consider the most likely definition in the top-right cell. This constraint definition penalizes stem-final voiceless non-stridents followed by either the [əz] allomorph (with [ə] matching the [-SG,-HI] bundle and [z] matching [-NAS,+VOI]) of the [z] suffix (with the Kleene star vacuously satisfied, as [z] fits the [-NAS,+VOI] condition). Essentially, this constraint functions both as an AGREE[VOI] constraint and a faithfulness constraint (DEP), by penalizing both voice-disagreeing [z] and unnecessary [əz] forms.

This strange constraint definition is entirely appropriate given the limited data the model has received, and it is even valid for novel plural forms, as shown by the model's highly accurate predictive abilities across all runs. As with the previously-discussed definitions, bringing in additional examples from English, such as faithful realizations of non-harmonious *kid* or *peg*, would illustrate that this constraint definition is not appropriate for English and encourage the model to find a more appropriate definition.

Perhaps the most interesting part of this learned constraint is that it is essentially a consonant harmony constraint, with voicing agreement ignoring interceding vowels. Consonant harmony is a phenomenon that has motivated emergentist accounts in the past, as it appears only in child phonologies (Pater, 1997; Pater & Werle, 2003), although its appearance here is likely an accident of the limited allomorphs of the English plural.

Examining the learned constraint definitions shows that the model does well at identifying reasonably-defined constraints, subject to the limitations of its data. The learned constraints are generally quite similar to the innate constraints, both in terms

of their violation profiles and their likely definitions. However, there are some idiosyncrasies in the learned constraint definitions that encourage learning across phonological phenomena jointly to better identify the true constraint definition.

3.5.4 Experiment Summary

We proposed three ways that the constraint learner could succeed, and we find evidence for all of them. The model learns a set of constraints and weights that can both explain observed data and effectively generalize to unobserved forms. In addition, we find that the constraint definitions it learns correspond reasonably well with the definitions that come from an innate set of constraints, although additional information is needed to identify the exact same constraints as the innate set.

3.6 Discussion and Future Directions

3.6.1 Expansion of the emergentist view

One of the core problems for the emergentist view of phonological constraints has been the difficulty in explaining purely formal constraints (Kager et al., 2004). This difficulty stems from the presumed sources for emergent constraints, which have generally been held to be articulatory or perceptual concerns (Bernhardt & Stemberger, 1998; Boersma, 1998; Hayes, 1999). If purely formal constraints, such as those proposed by Goad and Rose (2004) in child phonologies, do exist, then the articulatorily/perceptually motivated emergentist view is insufficient.

While articulatory and perceptual concerns certainly motivate some constraints, we show that some emergent constraints can be acquired without relying on them. Instead, the patterns in the observed data can supply sufficient information to acquire appropriate constraints from a very general constraint grammar. Articulatory/perceptual concerns could influence (and sharpen) constraint learning in a variety of ways. One is to

allow the articulatory or perceptual difficulty of a sequence to make it more likely in the constraint grammar, using a quantification system like Hayes (1999)'s phonetically-driven phonology.

In addition, this model represents the first emergentist computational model for acquiring intensionally-defined markedness constraints while influenced by faithfulness constraints, showing that there is enough information in the language data to support a general alternative to innate constraints.

3.6.2 The nature of the underlying representation

One of the main purposes of a computational model for human language acquisition is to investigate the plausibility of different potential underlying structures for human language representation. The feature-driven context-free grammar works well for the current task, and has the potential to handle other common constraints, such as harmony (using the Kleene star). But this is not the only possible grammar, and it is unlikely that all markedness constraints could be captured by its present form. An advantage of the current framework is that versions of the model can be implemented with different constraint grammars to gain information about the types of markedness constraints that are possible in human language.

For example, the model can be run with sonorants marked for voice (recall that the implementation discussed in Sect. 3.4 has sonorants unmarked for voice) but this makes a constraint like AGREE[VOI] (which would be *[-VOI][+VOI,-SON] in this new grammar) a less a priori likely constraint than a *[-VOI][+VOI] constraint that penalizes a voiceless consonant followed by a vowel. A test run of the model with this grammar did learn some constraints like this, unnatural for humans but natural in the constraint grammar. This suggests that the acquisition model can provide new evidence for deciding between different phonological structures.

The flexible constraint grammar could also be used to incorporate constraint schemas (J. Smith, 2004) into an emergentist system. Schemas were proposed as a way of creating a compositional structure for the general constraint set; each schema defined a type of constraint, with variable slots that could be used to define families of constraints, like Generalized Alignment. If there are formal constraints that are not expressible by the context-free grammar, schemas could be an appealing balance between emergent and innate constraints, providing a general structure that may be innate, but learning the specific constraint identity rather than innately specifying it.

3.6.3 Extending the model

The present work establishes and tests a basic framework for learning phonological constraints. It focuses on relaxing the amount of innately-specified structure needed for learning constraints by replacing a completely-specified innate constraint set with a more general grammar over phonological features. A few steps remain before this supplies a fully emergentist account for constraint acquisition – and it is of course possible that human acquisition is not as emergentist as this theoretically fully-emergentist approach.

One important remaining step is to incorporate non-binary constraints into the learning model. Non-binarity introduces two potential difficulties into the learning problem: first, that the space of possible matrices is substantially different and larger, and second that there could be more possible matrices that adequately explain the data.

Non-binary constraints can be introduced in multiple ways. The first is to follow a tradition in some OT learning methods of stacked binary constraints (R. Frank & Satta, 1998). A non-binary constraint C that has at most v violations can be represented as a series of binary constraints: $\{$*(more than 1 violation of C), *(more than 2 violations of C), ... , *(more than $v-1$ violations of C)$\}$. This solution is not perfect: it suggests

that non-binary constraints could be bounded, so that any number of violations beyond a certain number is equivalent, and there is no evidence for this in language. Furthermore, this makes the matrix inference much more complex, as a single constraint definition could generate multiple columns of violations.

The second and more promising avenue is to allow non-binary violation profiles. One way of implementing this is to use the existing binary matrix to indicate whether a candidate violates a constraint at all, and to introduce a second matrix, with positive integer values, corresponding to the number of violations that are expected.[15] This allows much of the generative model and the inference structure to remain unchanged. All that must be added is a prior over the positive-integer-valued matrix, which can also be marginalized over constraint definitions. The probability that a given candidate would violate a constraint d a certain number of times could be based on a Poisson distribution whose mean and mode are the number of violations that d predicts.

The model also could be expanded to learn faithfulness constraints, instead of its current assumption that they arise out of the candidate-generation process GEN. In the IBPOT model, the lack of phonological structure meant that faithfulness constraints were unlearnable. A trivial IBPOT solution would be to learn that every losing input-output pair violates the same high-weight faithfulness constraint, and no markedness constraints would be needed. By introducing a grammar of faithfulness constraints, our current Rational Rules OT model would find such a specific constraint unlikely, and appropriate faithfulness constraints should be learnable. The space of possible faithfulness constraints may be simpler than that of markedness constraints (McCarthy, 2008), suggesting that the faithfulness constraints may even be easier to learn than the markedness constraints.

Another, and more daunting question, common to much OT learning work, is

[15] A similar model structure is used by Griffiths and Ghahramani (2011) to overcome the binary nature of an Indian Buffet Process for recognizing objects in an image.

how learners know the underlying forms for the observed outputs. Some work has been done on these problems in the context of traditional OT. Prince and Smolensky (1993) propose using OT in reverse, selecting the most optimal underlying form for an output. Merchant and Tesar (2005) use minimal pairs in the output to identify the range of different underlying forms. Riggle (2006) uses the principle of maximum entropy to infer likely underlying forms given their predicted constraint violations. However, all of these methods require a ranking-based OT framework with known constraints, neither of which our model assumes is available at the start of learning.

Learning the underlying forms may best be represented as another part of a joint phonological learning framework. N. Smith (1973) argues that children start by assuming that the underlying forms are essentially equivalent to the observed adult output forms. As children acquire the phonology of a language, the underlying forms differentiate themselves from their outputs where appropriate. This could be incorporated into a joint learning MCMC framework like the present model by alternating between using the current estimate of the constraint structure to infer underlying forms and the current estimate of the underlying forms to infer constraints, potentially implemented in a Bi-directional OT framework (Dekker & van Rooy, 2000).

3.7 Conclusion

The standard assumption in Optimality Theory has long been that constraints are fully innate, and while there has been work on alternative sources for constraints, this has been grounded in articulatory and perceptual considerations, which limit their application. We presented a model for acquiring emergentist constraints primarily from distributional data, with help from a basic phonological structure. The constraints acquired by this model can be used to predict novel plural forms, and closely resemble the constraint definitions that innatist accounts propose. This shows that the emergentist approach is

more powerful than previously argued, and that learned constraints need not be directly phonetically grounded.

Chapter 4

Combining multiple information types in Bayesian word segmentation

Abstract Humans identify word boundaries in continuous speech by combining multiple cues; existing state-of-the-art models, though, look at a single cue. We extend the generative model of Goldwater et al (2006) to segment using syllable stress as well as phonemic form. Our new model treats identification of word boundaries and prevalent stress patterns in the language as a joint inference task. We show that this model improves segmentation accuracy over purely segmental input representations, and recovers the dominant stress pattern of the data. Additionally, our model retains high performance even without single-word utterances. We also demonstrate a discrepancy in the performance of our model and human infants on an artificial-language task in which stress cues and transition-probability information are pitted against one another. We argue that this discrepancy indicates a bound on rationality in the mechanisms of human segmentation.

4.1 Introduction

For an adult speaker of a language, word segmentation from fluid speech may seem so easy that it barely needed to be learned. However, pauses in speech and word boundaries are not well correlated (Cole & Jakimik, 1980), word boundaries are marked

by a conspiracy of partially-informative cues (Johnson & Jusczyk, 2001), and different languages mark their boundaries differently (Cutler & Carter, 1987). This makes the problem of unsupervised word segmentation acquisition, whether by a computational model or an infant, a daunting task.

Effective segmentation relies on the flexible integration of multiple types of segmentation cues, among them statistical regularities in phonemes and prosody, coarticulation, and allophonic variation. Infants begin using multiple segmentation cues within their first year of life (Johnson & Jusczyk, 2001). Despite this, many state-of-the-art models look at only one type of information: phonemes.

In this study, we expand an existing model to incorporate multiple cues, leading to an improvement in segmentation performance and opening new ways of investigating human segmentation acquisition. On the latter point, we show that rational learners can learn to segment without encountering words in isolation, and that human learners deviate from rationality in certain segmentation tasks.

4.2 Previous work

The prevailing unsupervised word segmentation systems (e.g., Brent, 1999; Goldwater et al., 2006; Blanchard & Heinz, 2008) use only phonemic information to segment speech. However, human segmenters use additional information types, notably stress information, in their segmentation. We present an overview of these phonemic models here before discussing the prosodic model expansion. A more complete review is available in Goldwater (2007).

4.2.1 Goldwater et al (2006)

The Goldwater et al model is related to Brent (1999)'s model, both of which use strictly phonemic information to segment. The model assumes that the corpus is

generated by a Dirichlet process over word bigrams.[1] We present a basic overview here, based on Sect. 5.5 of Goldwater, 2007. To generate the word w_i given the preceding word w_{i-1}:

1. Decide if bigram $b_i = \langle w_{i-1}, w_i \rangle$ is novel
2. If b_i non-novel, draw b_i from bigram lexicon
3. If b_i novel, decide whether w_i is novel
   a. If w_i non-novel, draw w_i from word lexicon
   b. If w_i novel, draw w_i from word-generating distribution P_0.

The Dirichlet process first decides whether to draw a non-novel ("nn") bigram, with probability proportional to the number of times the previous word has appeared in the corpus:

$$p(\langle w_{i-1}, w_i \rangle \text{ nn} | w_{i-1}) = \frac{n_{\langle w_{i-1}, \cdot \rangle}}{n_{\langle w_{i-1}, \cdot \rangle} + \alpha_1}, \tag{4.1}$$

where $n_{\langle x,y \rangle}$ is the token count for bigram $\langle x, y \rangle$. If the bigram is non-novel, word w_i is drawn in proportion to the number of times it has appeared after w_{i-1} in the corpus:

$$p(w_i = x | \langle w_{i-1}, w_i \rangle \text{ nn}) = \frac{n_{\langle w_{i-1}, x \rangle}}{n_{\langle w_{i-1}, \cdot \rangle}} \tag{4.2}$$

If the bigram is novel, this could either be due to w_i being a novel word or due to w_i being an existing word that had not appeared with w_{i-1} before. The probability of w_i being a non-novel word x is

$$p(w_i = x, w_i \text{ nn} | \langle w_{i-1}, w_i \rangle \text{ novel}) = \frac{b_{\langle \cdot, w_i \rangle}}{(b_{\langle \cdot, \cdot \rangle} + \alpha_0)}, \tag{4.3}$$

[1] We will only discuss the bigram model here because it is more appropriate from both a cognitive perspective (it posits latent hierarchical structure) and engineering perspective (it segments more accurately) than the unigram model.

where $b_{\langle .,. \rangle}$ is the count of word bigram types. Finally, if w_i is a new word, its phonemic form is generated from a distribution P_0. In the Goldwater et al model, this distribution is simply the product of the unigram probabilities of the phonemes, $P(\sigma_j)$, times the probability of a word boundary, $p_\#$, to end the word:

$$p(w_i = \sigma_1 \cdots \sigma_M | w_i \text{ novel}) = p_\#(1-p_\#)^{M-1} \prod P(\sigma_j) \tag{4.4}$$

To segment an observed corpus, the model Gibbs samples over the possible word boundaries (utterance boundaries are assumed to be word boundaries).[2] The exchangability of draws from a Dirichlet process allows for Gibbs sampling of each possible boundary given all the others.

4.2.2 A cognitively-plausible variant

Phillips and Pearl (2012) make these Bayesian segmentation models more cognitively plausible in two ways. The first is to move from phonemes to syllables as the base representational unit from which words are constructed, as infants learn to categorize syllables before phonemes (Eimas, 1999). The second is to add memory and processing constraints on the learner. They find that syllable-based segmentation is better than phoneme-based segmentation in the bigram model (though worse in the unigram model), and that, counter-intuitively, the constrained learner outperforms the unconstrained learner. This improvement appears to be driven by better performance in segmenting more common words. In this work, we adopt the syllabified representation but retain the unconstrained rational learner assumption.

[2]The model assumes that utterance boundaries are generated just like other words, and includes an adjustable parameter $p_\$$ to account for their frequency.

4.2.3 Other multiple-cue models

Some previous models have incorporated multiple cues, specifically the phonemic and stress information that our model will use. Two prominent examples are Christiansen, Allen, and Seidenberg (1998)'s connectionist model and Gambell and Yang (2006)'s algebraic model. The connectionist model places word boundaries where the combination of phonemic and stress information predict likely utterance boundaries, but does not include an explicit sense of "word", and performs only modestly on the segmentation task (boundary F-scores of .40-.45). The algebraic model also underperforms the Bayesian model (Phillips & Pearl, 2012) unless it includes the heuristic that there is a word boundary between any two stressed syllables. Our model presents a more general and completely unsupervised approach to segmentation with multiple cue-types.

In general, joint inference is becoming more common in language acquisition problems and has been shown to improve performance over single-feature inference. Examples include joint inference of a lexicon and phonetic categories (Feldman, Griffiths, & Morgan, 2009), joint inference of syntactic word order and word reference (Maurits, Perfors, & Navarro, 2009), and joint inference of word meanings and speaker intentions in child-directed speech (M. Frank, Goodman, & Tenenbaum, 2009).

4.3 Model design

Our model changes P_0 from a single-cue distribution, generating only phonemes, to a multiple-cue distribution that generates a stress form as well. This can improve segmentation performance and allows the investigation of rational segmentation behavior in a multiple-cue world.

In the original model, $P_0(w_i = \sigma_1 \cdots \sigma_M) \propto \prod_j P(\sigma_j)$, where $P(\sigma_j)$ is the frequency of the phoneme σ_j. In the multiple-cue model, we first generate a phonemic form

w_i, then assign a stress pattern s_i to it.

$$\begin{aligned} P_0(w_i, s_i) &= P_W(w_i) P_S(s_i|M) \\ &= p_\#(1-p_\#)^{M-1} \prod_j^M P(\sigma_j) P_S(s_i|M) \end{aligned} \tag{4.5}$$

The phonemic form w_i has the same product-of-segments probability as the Goldwater et al model, but σ_j are now syllables instead of phonemes. We discuss the rationale behind this change in the next section.

The phonemic form is generated first, and the stress form is then drawn as a multinomial over all possible stress patterns with the same number of syllables as w_i. The stress distribution P_S is a multinomial distribution over word-length stress templates. P_S can be learned by the model based on a Dirichlet prior, but for simplicity in the present implementation, we estimate P_S as the plus-one-smoothed frequency of the stress patterns in the current segmentation. There are two stress levels (stressed or unstressed), and 2^M possible stress templates for a word of length M.[3]

Unlike phonemic forms, stress patterns are drawn as a whole word. This allows the model to capture a wide range of stress biases, although it prevents the model from generalizing biases across different word lengths. A potential future change to P_S that would allow for better generalization is discussed in Section 4.6.

4.3.1 On syllabification and stress

We change from segmenting on phonemes to segmenting on syllables in order to more easily implement stress information, which is a supersegmental feature most

[3]We do not assume that each word has one and only one stressed syllable, which would reduce the number of possible stress templates to M, for two reasons. First, in the current corpus, some words have citation forms with multiple stressed syllables. Second, in actual speech this assumption will not hold (e.g., many function words go unstressed).

appropriately located on syllables. Syllabified data has been used in some previous models of segmentation, especially those using stress information or syllable-level transition probabilities (Christiansen et al., 1998; Swingley, 2005; Gambell & Yang, 2006; Phillips & Pearl, 2012).

For studying human word segmentation, Phillips and Pearl argue syllabified speech may be a more cognitively plausible testing ground. 3-month-old infants appear to have categorical representations of syllables (Eimas, 1999), three months before word segmentation appears (Borfeld, Morgan, Golinkoff, & Rathbun, 2005), and seven months before phoneme categorization (Werker & Tees, 1984). In addition, syllabification is assumed in much work on human word segmentation, especially in artificial-language studies (e.g., Thiessen & Saffran, 2003), which calculate statistical cues at the syllable level.

The assumption that syllable boundaries are known affects the baseline performance of the model, as it reduces the number of possible word boundary locations (since a word boundary is necessarily a syllable boundary). As such performance over syllabified data cannot be directly compared to performance on non-syllabified data.

It may seem that syllabification is so closely tied to word segmentation that including the former in a model of the latter leaves little to the model. However, the determinants of syllable boundaries are not the same as those for word boundaries. The problem of assigning syllable boundaries is a question of deciding where a boundary goes between two syllable nuclei, with the assumption that there must be a boundary there. The problem of assigning word boundaries is a question of deciding whether there is a boundary between two syllable nuclei, and if so, where it is. Knowing the syllable boundaries reduces the set of possible word boundaries, but does not directly address the question of how likely a boundary is. The difference in these tasks is supported by the three-month gap between syllable and word identification in infants.

Table 4.1. Corpus stress patterns by types and tokens, showing an initial-stress bias in all lengths.

Types		**Tokens**	
Stress pattern	Count	Stress pattern	Count
S	21402	S	523
SW	2231	SW	208
SS	389	WS	40
WS	284	SWW	24
SWW	182	SS	7
WSW	33	WSW	7
Other	5	Other	2

4.4 Data

We use the Korman (1984) training corpus, as compiled by Christiansen et al. (1998), in this study. This is a 24493-word corpus of English spoken by adults to infants aged 6–16 weeks.[4] Phonemes, stresses, and syllable boundaries are the same as those used by Christiansen et al, which were based on citation forms in the MRC Psycholinguistic Database. All monosyllabic words were coded as stressed. Only utterances for which all words had citation forms were included.

This corpus is largely monosyllabic (87.3% of all word tokens), and heavily biased toward initial stress (89.2% of all multisyllable word tokens). No word is longer than three syllables, and most words have only one stressed syllable. A breakdown of the corpus by stress pattern is given in Table 4.1. This monosyllabic bias is an inherent property of English, not idiosyncratic to this corpus. The Bernstein-Ratner child-directed corpus is also over 80% monosyllabic. We expect that the results of segmentation on child-directed data will extend to adult speech, as the adult-directed corpus used by Gambell and Yang (2006) has an average word length of 1.17 syllables.

[4] Approximately 150 word tokens from the original corpus were omitted in our version of the corpus due to a disparity between recorded number of syllables and number of stresses.

4.5 Experiments

We test the model on three problems. First, we show that the addition of stress information improves segmentation performance compared to a stress-less model. Next, we apply the model to a question in human segmentation acquisition. Finally, we look at a task where the rational model deviates from human performance.

4.5.1 Parameter setting

The model has four free parameters: α_0 and α_1, which affect the likelihood of new words and bigrams, respectively, and $p_\#$ and $p_\$$, which affect the expected likelihood of word and utterance boundaries. Following Goldwater, Griffiths, and Johnson (2009), we set $\alpha_0 = 20$, $\alpha_1 = 100$, $p_\# = 0.8$ and $p_\$ = 0.5$ in all experiments.[5]

In all cases, the model performed five independent runs of 20000 iterations of Gibbs sampling the boundaries for the full corpus. Simulated annealing was performed during the burn-in period to improve convergence. All performance measures are reported as the mean of these five runs.

Performance is measured as word, boundary, and lexicon precision, recall, and F-scores. A word is matched iff both of its true boundaries are marked as boundaries and no internal boundaries are marked as word boundaries. Boundary counts omit utterance boundaries, which are assumed to be word boundaries. Lexical counts are based on word type counts.

4.5.2 Stress improves performance

We begin by showing that including a second cue type improves segmentation performance. We compare segmentation on a corpus with the attested stress pat-

[5] Performance was similar for a range of settings between 1 and 100 for α_0 and between 10 and 200 for α_1.

Table 4.2. Precision, recall, and F-score over corpora with and without stress information available. Stress information especially improves lexical performance.

	With stress			Without stress		
	Word	Bnd	Lex	Word	Bnd	Lex
Prec	.76	.99	.75	.76	.99	.72
Rec	.61	.70	.87	.60	.69	.84
F	.68	.82	.80	.67	.82	.77

terns to that of a corpus without stress. With stress information included in the model, word/boundary/lexicon F-scores are .68/.82/.80. Without stress, performance drops to .67/.82/.77.[6] Full results are given in Table 4.2.

Stress information primarily improves lexicon performance, along with a small improvement in token segmentation. Accounting for stress reduces both false positives and negatives in the lexicon; the fact that the lexical improvement is greater than that for words or boundaries suggests that much of the improvement rests is on rare words.

These effects are small but significant. For word token performance, we performed a paired *t*-test on utterance token F-scores between the with- and without-stress models. This difference was significant ($t = 11.28, df = 8125, p < .001$). We performed a similar utterance-by-utterance test on boundaries; again a small singificant improvement was found ($t = 8.92, df = 6084, p < .001$). To assess lexicon performance, we calculated for each word type in the gold-standard lexicon the proportion of the five trials in which that word appeared in the learned lexicon for the two models. We then examined the words where the proportions differed between the models. 89 true words appeared more often in the with-stress lexicons; 40 appeared more often in the without-stress lexicons. (683 appeared equally often in both.) By a sign test, this is significant at $p < .001$. We also tested lexicon performance with a binomial test on the two models' lexicon accuracy;

[6] Recall that due to the syllabified data, these results are not directly comparable to unsyllabified results in previous work.

this result was marginal ($p = .06$).

The explicit tracking of stress information also improves the model's acquisition of the stress bias of the language. Acquisition of the stress bias is potentially useful for generalization; stress patterns can be used for an initial segmentation if few or none of the words are familiar. In practice, we see children use their stress biases to segment new words from English speech (Jusczyk, Houston, & Newsome, 1999) as well as artificial languages (Thiessen & Saffran, 2003).

We assess the learned stress bias by dividing up the corpus as the model has segmented it, and count the number of tokens with SW versus WS stress patterns.[7] With stress representation, the learned stress bias is 6.77:1, and without stress representation, the stress bias is lower, at 6.33:1. Although these are both underestimates of the corpus's true stress bias (7.86:1), the stressed model is stronger and a better estimate of the true value.

The model's performance can be compared to various baselines, but perhaps the strongest is one with every syllable boundary being a word boundary. This baseline represents a shift from boundary *precision* being at ceiling (as in the model) to boundary *recall* being at ceiling. In fact, due to the preponderance of monosyllabic words in English child-directed speech, this baseline outperforms the model on word and boundary F-scores (.68 and .82 in the model, .82 and .91 in the baseline). However, the baseline's lexicon is much worse than the model's (F=.80 in the with-stress model, F=.64 in the baseline), and the baseline fails to learn anything about the language's stress biases. In addition, the baseline oversegments, whereas both the model and infant segmenters undersegment (Peters, 1983). This raises an important question about what the model should seek to optimize: though the baseline is more accurate by token, no structure is learned; type performance is more important if we want to learn the underlying structure.

[7] Note this defines a stress bias for the stressless model as well.

4.5.3 Are isolated words necessary?

We next use this model to test the necessity of isolated words in rational word segmentation. It is not immediately obvious how human learners begin to segment words from fluid speech. Stress biases and other phonological cues are dominant in all but the earliest of infant word segmentation (Johnson & Jusczyk, 2001). This raises a chicken-and-egg problem; if the cues infants favor to segment words, such as stress biases, are dependent on the words of the language, how do they learn enough words to determine the cues' biases?

One existing proposal is that human learners develop their stress biases based on words frequently heard in isolation (Jusczyk et al., 1999). In English, these include names and common diminutives (e.g., *mommy*, *kitty*) that generally have initial stress. These single-word utterances could offer the segmenter an initial guess of the stress bias, by supposing that short utterances are single words and recording their stress patterns. The most common stress patterns in short utterances could then be used as an initial guess at the stress bias to bootstrap other words and thereby improve the learned stress bias.

We test the rational learner's need for such explicit bootstrapping by learning to segment a corpus with all single-word utterances removed. The corpus is produced by excising all single-word utterances from the Korman corpus. This results in a 22081-word corpus, 10% fewer tokens than in the original. However, it does not substantially change the lexicon; the number of distinct word types only drops from 811 to 806.

We compare performance only on ambiguous boundaries and lexicon, as these are comparable between the corpora, and find that the model performs almost equally well. Without single-word utterances, boundary and lexical F-scores are .81 and .80, compared to .82 and .80 with single-word utterances. This shows that rational learners

are able to segment even without the possibility of bootstrapping stress patterns from single-word utterances.

4.5.4 Bounded rationality in human segmentation

Lastly, we use this model to examine rational performance in a multiple-cue segmentation task. We show that humans' segmentation does not adhere to these predictions, suggesting a bound on human rationality in word segmentation.

We consider an artificial language study by Thiessen and Saffran (2003). In this study, infants are exposed to an artificial language consisting of four bisyllabic word types uttered repeatedly without pauses. Each syllable appears in only one word type, so within-word transition probabilities are always 1, while across-word transition probabilities are less than 0.5. Segmentation strategies that hypothesize word boundaries at low transition probabilities or that seek to minimize the lexicon size will segment out the four word types as expected.

Segmentation in the experiment is complicated by the presence of stress in the artificial language. Depending on the condition, the words are either all strong-weak or all weak-strong. In the first condition, segmenting according to transition probabilities, lexicon size, or English stress bias favors the same segmentation. In the second condition, though, segmenting by the English stress bias to yield a lexicon of strong-weak words requires boundaries in the middle of the words. The segmenter must decide whether transition probabilities or preferred stress patterns are more important in segmentation. This situation is illustrated in Table 4.3, with a corpus consisting of two word types, *AB* and *CD*, each with weak-strong stress.

Thiessen and Saffran found that seven-month-old English-learning infants consistently segmented according to the transition probabilities, regardless of stress. However, nine-month-olds segmented according to the English stress bias, even if this meant going

Table 4.3. Examples of segmenting an artificial language according to transition probabilities (top) or stress bias (bottom), when the true words have weak-strong stress. Vertical lines represent word boundaries. The top segmentation produces a smaller lexicon, but the bottom segmentation produces primarily words with the preferred stress pattern.

Against bias, with TP				**With bias, against TP**				
AB	CD	CD	AB	A	BC	DC	DA	B
WS	WS	WS	WS	W	SW	SW	SW	S

against the transition probabilities.

Intuitively, this could be rational behavior according to our model. A child's increasing age means more exposure to data, potentially leading the child to develop more confidence in the stress bias. As confidence in the stress bias increases, the cost of segmenting against it increases as well. A sufficiently strong stress preference could lead the segmenter to accept a large lexicon, all of whose words have the preferred stress pattern, over a small lexicon, all of whose words have the dispreferred stress pattern.

To judge by the Korman corpus, English has a stress bias of approximately 7:1 in favor of SW bisyllabic stress over WS.[8] If human segmentation behavior follows the rational model, the model should predict segmentation to favor strong-weak words over the transition probabilities when the stress bias is approximately this strong.

We test this rationality hypothesis with a smaller version of the Thiessen and Saffran artificial language, consisting of 48 tokens.[9] In one version, all tokens have the preferred SW pattern, and in the other all tokens have the dispreferred WS pattern. We then adjust the P_S distribution such that $P_S(SW|M=2) = b * P_S(WS|M=2)$, where b is the bias ratio. We run the model otherwise the same as in the previous experiments, except with 10 runs instead of 5.

[8] The specific bias varies from corpus to corpus, but this appears to be a representative value.

[9] The 48 tokens come from four word types, with two types appearing 16 times and the other two appearing 8 times, mimicking the relative frequencies of Thiessen and Saffran's languages. Their test language had 270 tokens.

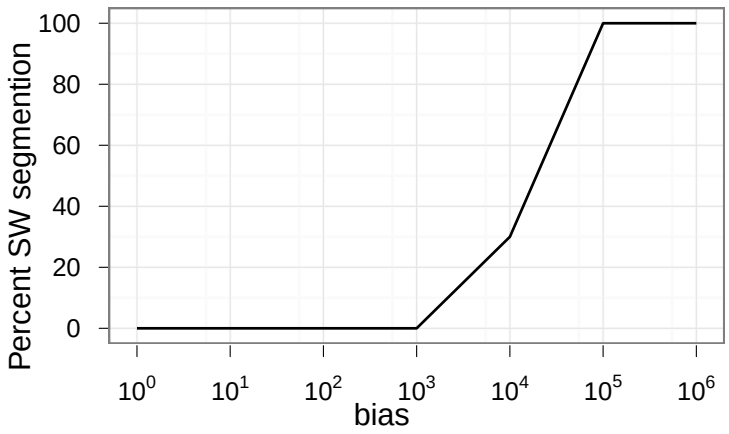


Figure 4.1. Percentage of runs segmented with the stress bias, against transition probabilities, as bias varies. At English-level biases, the rational model still overrules the stress bias when segmenting.

Contrary to this hypothesis, the model's segmentation with $b = 7$ was the same whether the true words were strong-weak or weak-strong. In all ten runs, transition probabilities dictated the segmentation. To switch to stress-based segmentation, the bias must be orders of magnitude greater than the English bias. Figure 4.1 shows the proportions of runs in the weak-strong condition that show segmentation according to the stress bias, as the bias increases by factors of 10. When $b = 10000$, three of the ten runs segmented according to the stress bias; below that, the stress bias did not affect the rational model's segmentation.

Why is this? In the Bayesian model, the stress bias of a language affects only the $P_S(s_i|M)$ term in the P_0 distribution, so non-novel words are not penalized for their stress pattern. The model pays only once to create a word; once the word is generated, no matter how a priori implausible the word was, it may be cheaply drawn again as a non-novel word. This effect can be illustrated with a brief calculation.

Consider a corpus built from four bisyllabic word types (AB, CD, EF, GH), each

appearing N times. If the corpus is segmented against the transition probabilities, the resulting lexicon will have 16 bisyllabic word types (BA, BC, BE, BG, DA, etc.), each occurring approximately $\frac{N}{4}$ times.

The probability of the against-bias corpus (C_{WS}) is proportional to the probability of generating the four word types, and then drawing them non-novelly from the lexicon.[10] (To simplify the calculations, we use the unigram version of the Goldwater et al model.)

$$p(C_{WS}) \propto P_W^4 P_S(WS)^4 (N!)^4 \frac{1}{4N!} \tag{4.6}$$

The first two terms are the probability of generating the four word types (Eqn. 4.5);[11] the second two terms are the Dirichlet process draws from the existing lexicon N times each (Eqn. 4.2). By comparison, the probability of the with-bias corpus C_{SW} depends on generating the 16 word types, and drawing each non-novelly $\frac{N}{4}$ times.

$$p(C_{SW}) \propto P_W^{16} P_S(SW)^{16} \left(\frac{N}{4}!\right)^{16} \frac{1}{4N!}$$

Given an SW bias b and a uniform distribution over syllables (so $P_W = \frac{1}{64}$), we find:

$$\frac{p(C_{WS})}{p(C_{SW})} = 64^{12} \frac{(b+1)^{12}}{b^{16}} \frac{(N!)^4}{(\frac{N}{4}!)^{16}} \tag{4.7}$$

This equation shows that the rational model is heavily biased toward the segmentation that fits the transition probabilities. Increasing the stress bias b or decreasing the number of observed word tokens makes the rational model more likely to segment with the stress bias (against transition probabilities), but as we see in the experimental results,

[10] It is also possible to generate this corpus by re-drawing the words novelly, but this is much less likely than non-novel draws.

[11] Because all syllables have equal unigram probabilities, the probability of all words' phonemic forms are equal, and will be written as P_W.

the stress bias must be very strong to overcome the efficient lexicon that the transition probability segmentation provides.

Since humans do not show this same inherent bias (or quickly lose it as they acquire the stress bias), we can ask how humans deviate from rationality. One possibility is that humans simply do not segment in this Bayesian manner. However, previous work (M. Frank et al., 2010) has shown that human word segmentation shows similar behavior to a resource-limited Bayesian model. Equation 4.7 suggests that human segmentation could deviate from rationality by having an effectively stronger bias than English would suggest (reducing the first fraction)[12] or, as with Phillips and Pearl's constrained learners, by having effectively less input than the model assumes (reducing the second fraction).

4.6 Future work

Introducing stress into the Bayesian segmentation model suggests a few additional expansions. One possibility is to add other cues into the generative model via P_0. Any cue that is based on the word itself can be added in this way, with little change to the general model structure. Phonotactics can be added using an n-gram distribution for P_0 (Blanchard & Heinz, 2008). Coarticulation between adjacent phonemes is also used in human segmentation (Johnson & Jusczyk, 2001), so the P_0 distribution could predict higher within-word coarticulation. Integrating additional cues used by human segmenters extends the investigation of the bounds on rationality in human segmentation and in balancing multiple conflicting cues.

A more complex view of the stress system of a language may also be useful. One possibility is to place a Dirichlet prior over the stress templates and allow P_S to be learned

[12] A potential source of an inflated bias is infants' preference for strong-weak patterns. Jusczyk, Cutler, and Redanz (1993) found English-hearing infants listened longer to strong-weak patterns than weak-strong. This could lead to overestimation of the stress bias by making possible strong-weak segmentations more prominent in the segmenter's mind.

as a latent variable in the model. Another possibility is to treat the stress templates more generally; in the present implementation, knowledge of the preferred stress patterns for word of one length tells the segmenter nothing about preferred stress patterns in another length. Cross-linguistically common stress rules (e.g., those that place stress a certain number of syllables from the left or right edge of a word) can be coded into P_S to improve generalization. Each rule dictates a specific stress pattern for each word length. When a word is generated in the Dirichlet process, the generative model would decide whether to assign stress according to one of these rules or to assign lexical stress from a default multinomial distribution. (This "default" distribution would handle idiosyncratic stress assignments, as one might see with names or morphologically complex words, like Spanish reflexive verbs.) A sparse prior over these rules, asymmetrically weighted against the default category, will encourage the model to explain as much of the observed stress patterns as possible with a few dominant rules, improving the phonological structure that the segmenter learns.

Improving the realism of the data is also important. The corpora used in much of segmentation research are idealized representations of the true data, and the dictionary-based phoneme and stress patterns used in this study are no exception. This ideal setting may paint a skewed picture of the segmentation problem, by providing a more consistent and learnable data source than humans actually receive. Elsner, Goldwater, and Eisenstein (2012)'s model unifying lexical and phonetic acquisition takes a significant step in showing that a rational segmenter can handle noisy input by recognizing phonetic variants of a base form. In terms of stress representations, dictionary-based stress has been standard in previous work (Christiansen et al., 1998; Gambell & Yang, 2006; Rytting, Brew, & Fosler-Lussier, 2010), but it is important to confirm such results against a (currently nonexistent) corpus with stresses based on the actual utterances. Effective use of stress in a less idealized setting may require a more complex representation of

stress in the model.

4.7 Conclusion

Effective word segmentation combines multiple factors to make predictions about word boundaries. We extended an existing Bayesian segmentation model to account for two factors, phonemes and stress, when segmenting. This improves segmentation performance and opens up new possibilities for comparing rational segmentation and human segmentation.

4.8 Acknowledgments

This research was partially supported by an Alfred P. Sloan Fellowship to RL and by NSF award 0830535. We also appreciate the feedback of the NAACL reviewers and the members of the UCSD Computational Psycholinguistics Lab.

This chapter, in full, is an exact copy of the material as it appears in Doyle and Levy (2013) [Combining multiple information types in Bayesian word segmentation. In L. Vanderwende, H. Daumé III, and K. Kirchhoff (Eds.), Proceedings of the 2013 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (pp. 117-126). Atlanta: Association for Computational Linguistics.] The dissertation author was the primary investigator and author of this paper.

Chapter 5

Accounting for burstiness in topic models

Abstract Many different topic models have been used successfully for a variety of applications. However, even state-of-the-art topic models suffer from the important flaw that they do not capture the tendency of words to appear in bursts; it is a fundamental property of language that if a word is used once in a document, it is more likely to be used again. We introduce a topic model that uses Dirichlet compound multinomial (DCM) distributions to model this burstiness phenomenon. On both text and non-text datasets, the new model achieves better held-out likelihood than standard latent Dirichlet allocation (LDA). It is straightforward to incorporate the DCM extension into topic models that are more complex than LDA.

5.1 Introduction

The effectiveness of a topic model is dependent on the appropriateness of its generative process for the task at hand. For most common tasks, any computationally feasible generative model will be a substantial simplification of the true generative process. Nevertheless, some tractable generative models are more reflective of the true generative process than others. In this paper, we propose a new generative process for topic models

that significantly improves the statistical fidelity of the process with minimal additional model complexity. Specifically, we replace the multinomial distributions in standard latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) by Dirichlet compound multinomial (DCM) distributions (Madsen, Kauchak, & Elkan, 2005; Elkan, 2006). The result is a better model for text data and for at least some other non-text data.

Our primary concern in the current study is accounting for the phenomenon of burstiness. Church and Gale (1995) note that real texts systematically exhibit this phenomenon: a word is more likely to occur again in a document if it has already appeared in the document. Importantly, the burstiness of a word and its semantic content are positively correlated; words that are more informative are also more bursty. The multinomial distribution does not take burstiness into account (Rennie, Shih, Teevan, & Karger, 2003, Sect. 4.1), so it gives an inaccurate model for the distribution of words in texts.

The phenomenon of burstiness is not limited to text. In Section 5.4 we present an example of bursty data in the financial realm. Burstiness also intuitively occurs in other types of data that have been modeled using topic models, including gene expression and computer vision data (Airoldi, Fienberg, & Xing, 2007; Fei-Fei & Perona, 2005). If a gene is transcribed once in a cell, then it is more likely to be transcribed again. And if a patch with certain properties occurs once in an image, then it is more likely that similar patches will occur again.

The new DCMLDA model is only slightly more complex than standard LDA. As a result, the LDA component in complex topic models, such as Pachinko allocation (Li & McCallum, 2006) and correlated topic models (Blei & Lafferty, 2005), can be replaced with a DCMLDA component. This should enable those models to account for burstiness and thereby improve their effectiveness.

Because it uses DCMs to represent topics, the DCMLDA model can capture the

tendency of the same topic to manifest itself with different words in different documents. Suppose that there is a natural "sports" topic in a corpus, with the words "rugby" and "hockey" being equally common overall. Within a document, though, one appearance of "rugby" makes a second appearance of "rugby" more likely than a first appearance of "hockey." The DCM distributions in DCMLDA can represent this fact, while a standard LDA model cannot. This property allows a single DCMLDA topic to explain related aspects of documents more effectively than a single LDA topic. Thus, we hypothesize, a DCMLDA model with a few topics can fit a corpus as well as an LDA model with many topics. This hypothesis is confirmed by the experimental results below.

5.2 Overview of Models

The DCMLDA model combines the DCM and LDA models, gaining the advantages of each. We review the two component models before discussing DCMLDA.

5.2.1 Latent Dirichlet allocation (LDA)

LDA has been discussed in detail elsewhere (Blei, Ng, & Jordan, 2001; Blei et al., 2003; Griffiths, Steyvers, Blei, & Tenenbaum, 2004; Heinrich, 2005), so we present only an overview here. The LDA generative model notionally posits that an author generates a document in two steps. First, the author determines the probability of each topic in the document. Each topic is a multinomial distribution over words, so to choose a word the author first draws a topic and then draws a word based on that topic. The graphical model for LDA is shown in Figure 5.1(a), with the unobserved variables distributed as follows:

$$\theta \sim \text{Dirichlet}(\alpha) \qquad z \sim \text{Multinomial}(\theta)$$
$$\phi \sim \text{Dirichlet}(\beta) \qquad w \sim \text{Multinomial}(\phi).$$

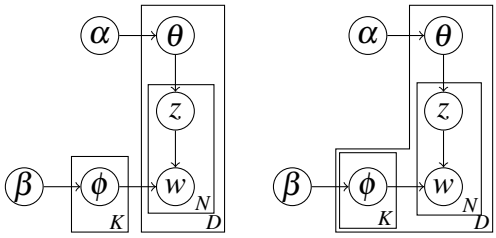
This generative process does not account for burstiness of words. The only way that burstiness can manifest itself is indirectly, as a consequence of how topics are distributed.

The fact that a document contains the word "rugby" from a sports topic, for instance, makes it more likely that the document contains other words from the same sports topic. Thus, the document is likely to contain a second instance of the word "rugby." However, because the sports topic is the same across the corpus, the presence of any sports word in a document will have a similar effect. That is, an appearance of the word "rugby" also indirectly makes an appearance of the word "hockey" more likely, which is not a desirable phenomenon.

The LDA model is bursty in topics, even though it is not in words: the presence of one word from a given topic in a document makes other words in the document more likely to be generated by the same topic. However, because the LDA generative process does not account for word-level burstiness, LDA may in fact be excessively bursty at the topic level. The reason is that each occurrence of a word is treated as independent extra evidence for its topic.

An LDA model has two Dirichlet hyperparameters, α and β, which condition θ and ϕ respectively. Different values for the hyperparameters cause different inferred values of ϕ and θ. In general α and β are vectors that can be learned (Blei et al., 2003; Fei-Fei & Perona, 2005). However, often they are kept fixed and uniform, meaning that each vector component is set to the same scalar value.

Learning the hyperparameters can provide information about the corpus: α indicates how semantically diverse documents are, with lower α indicating increased diversity, while β indicates how similar the topics are, with higher β indicating more similarity between topics. Learning non-uniform values for the hyperparameters allows different words and topics to have different tendencies; some topics can be more general than others (e.g., function words versus medical jargon), and some words can be likely to appear in more topics than others (e.g., words with multiple senses).



(a) LDA **(b)** DCMLDA

Figure 5.1. Alternative graphical models.

Despite not accounting for burstiness, LDA is an effective model that has proven useful for its ability to model documents as varying mixtures of shared topics. From a trained LDA model, one can infer the multinomial distributions θ that give the probability of each topic in each document. These distributions can then be used for many tasks, including classifying new documents and measuring similarity between documents.

5.2.2 Dirichlet compound multinomial (DCM)

The DCM model (Madsen et al., 2005) captures burstiness, but it has no notion of topic. DCM uses a bag-of-bags-of-words generative process. In this process, each document is formed by drawing a document-specific multinomial distribution ϕ from a shared Dirichlet distribution, and then drawing words w according to ϕ. In the DCM model, each document is composed of words drawn from a single multinomial. This multinomial can be viewed as a document-specific subtopic, or aspect, of the high-level topic β. The β vector is the only parameter of DCM, so unlike the hyperparameters in LDA, it must be non-uniform.

Since topics are drawn from a Dirichlet distribution in LDA also, it is perhaps not immediately obvious why DCM accounts for the burstiness of words and LDA does not. The answer lies in the 1:1 mapping between subtopics and documents in the DCM model. In LDA, the multinomial distribution of words in each topic depends on the whole corpus, but DCM multinomial distributions are document-specific.

Turning to the mathematics of the models, a key difference is that multinomial parameters are constrained to sum to one, unlike Dirichlet parameters. This gives the DCM model one extra degree of freedom to represent a topic. By working out an exponential-family approximation of the DCM, Elkan (2006) shows explicitly that this degree of freedom allows the DCM to discount multiple observations of the same word. In bursty texts, additional appearances of a word are less surprising than its first appearance.

The smaller the sum of the Dirichlet parameters β, the more the emission of words is bursty. As the Dirichlet parameters tend to infinity, a DCM distribution approaches equivalence with a multinomial distribution.

A single DCM model represents one high-level topic that has alternative aspects. It cannot represent multiple distinct topics. Because of the 1:1 mapping between multinomials and documents, in a DCM model each document comes entirely from one subtopic. All these subtopics are closely related because the sum of the β vector is typically quite high (a few hundred). Elkan (2006) extended the DCM model to a mixture of DCM distributions. This model can be trained to represent a set of documents where each document comes from a different high-level topic, but it cannot represent the scenario where a single document contains words from more than one high-level topic.

Algorithm 1. DCMLDA Generative Model

for document $d \in \{1, \ldots, D\}$ **do**
  draw topic distribution $\theta_d \sim \text{Dir}(\alpha)$
  for topic $k \in \{1, \ldots, K\}$ **do**
    draw topic-word distribution $\phi_{kd} \sim \text{Dir}(\beta_k)$
  end for
  for word $n \in \{1, \ldots, N_d\}$ **do**
    draw topic $z_{dn} \sim \theta_d$
    draw word $w_{dn} \sim \phi_{z_{dn}d}$
  end for
end for

5.2.3 DCMLDA

To combine the advantages of DCM and LDA, we need a model that allows multiple topics in a single document, while still making the topics document-specific to account for burstiness. Figure 5.1 contrasts the LDA and DCMLDA graphical models, while Algorithm 1 is the DCMLDA generative process.

In LDA, for each topic k, one multinomial distribution ϕ_k is drawn from Dirichlet(β) and is used in all documents. In DCMLDA, for each topic k and each document d a fresh multinomial word distribution ϕ_{kd} is drawn. Each topic k has a different, non-uniform β_k vector. For each document d, ϕ_{kd} is drawn according to Dirichlet(β_k), so the instances of each topic are linked across documents. Having per-document instances of each topic allows for variations in the probability of each word in the same topic in different documents, which is the phenomenon of burstiness.

The change from a single set of multinomial topics to multiple sets of multinomial subtopics shifts the focus of attention in DCMLDA modeling. Let V be the size of the vocabulary, let K be the number of topics, and let D be the number of documents in the corpus. In LDA, ϕ is the focus, a $V \times K$ array of word probabilities given topics. In DCMLDA, ϕ is three-dimensional ($V \times K \times D$), measuring word likelihoods for each topic, for each document. Since in DCMLDA ϕ depends on the specific document, it is not a representation of the data that has sharply reduced dimensionality. Instead, with DCMLDA the focus of attention is β, which is a two-dimensional array of Dirichlet parameters for words given topics. As mentioned in the previous section, the β values are not constrained to sum to one. This gives DCMLDA an extra K degrees of freedom that allow it to capture word-level burstiness within each topic. The β values have a similar intuitive interpretation to the ϕ values in LDA. In particular, higher β values mean that a word is more likely in a given topic. Thus one can still use β values to identify the most

common words in each topic.

5.3 Methods of Inference

Both the standard LDA model and the DCMLDA model have five unobserved variables: α, β, ϕ, θ, and z. These variables can be classified into two groups: the per-document or per-word parameters ϕ, θ, and z, and the hyperparameters α and β. Given a training set of documents, we learn appropriate values for the variables by alternating between optimizing the topic parameters given the hyperparameters, and optimizing the hyperparameters given the topic parameters. Neither of these optimizations can be done analytically, but both yield to known estimation procedures. Specifically, for fixed values of the α vector and β array, we do collapsed Gibbs sampling to find the distribution of z given the documents. If desired, ϕ and θ can be computed straightforwardly from samples of z. Given a z sample, values of α and β that maximize the likelihood of the training documents are obtained by Monte Carlo expectation-maximization.

In this and subsequent sections, the notation $\beta_{\cdot k}$ indicates that β is a two-dimensional array, with one column for each topic k, so $\beta_{\cdot k}$ is what was informally called β_k previously. Similarly, the notation $\alpha_{\cdot}$ is used to emphasize that α is a vector.

Gibbs sampling Gibbs sampling for DCMLDA is similar to the method for LDA, which Heinrich (2005) explains in detail. We present a condensed derivation, highlighting what is novel for DCMLDA sampling. We start by factoring the complete likelihood of the model: $p(w,z|\alpha_{\cdot},\beta_{\cdot\cdot}) = p(w|z,\beta_{\cdot\cdot})p(z|\alpha_{\cdot})$. The first probability is an average over all

possible ϕ distributions:

$$\begin{aligned} p(w|z,\beta_{..}) &= \int_{\phi} p(z|\phi)p(\phi|\beta_{..})d\phi \\ &= \int_{\phi} p(\phi|\beta_{..}) \prod_{d} \prod_{n=1}^{N_d} \phi_{w_{dn} z_{dn} d} \, d\phi \\ &= \int_{\phi} p(\phi|\beta_{..}) \prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} d\phi. \end{aligned}$$

Expanding $p(\phi|\beta_{..})$ as a Dirichlet distribution yields

$$\begin{aligned} p(w|z,\beta_{..}) &= \int_{\phi} \left[\prod_{d,k} \frac{1}{B(\beta_{\cdot k})} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1} \right] \left[\prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} \right] d\phi \\ &= \prod_{d,k} \int_{\phi} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1+n_{tkd}} d\phi \\ &= \prod_{d,k} \frac{B(n_{\cdot kd} + \beta_{\cdot k})}{B(\beta_{\cdot k})}. \end{aligned} \tag{5.1}$$

Above, $B(\cdot)$ is the multidimensional Beta function, and n_{tkd} is how many times word t is assigned topic k in document d. DCMLDA and LDA are structurally identical over the α-to-z pathway, so $p(z|\alpha_{\cdot})$ in DCMLDA is the same as for LDA:

$$p(z|\alpha_{\cdot}) = \prod_{d} \frac{B(n_{\cdot\cdot d} + \alpha_{\cdot})}{B(\alpha_{\cdot})}. \tag{5.2}$$

Combining Equations 5.1 and 5.2 yields that the complete likelihood $p(w,z|\alpha_{\cdot},\beta_{..})$ is

$$\prod_{d} \left[\frac{B(n_{\cdot\cdot d} + \alpha_{\cdot})}{B(\alpha_{\cdot})} \prod_{k} \frac{B(n_{\cdot kd} + \beta_{\cdot k})}{B(\beta_{\cdot k})} \right]. \tag{5.3}$$

To perform collapsed Gibbs sampling, we need to calculate $p(z_i|z_{-i},w)$, where z_{-i} is the set of topic assignments to all words but w_i. Letting n_{tkd} be the count of word t in topic k and document d in the complete corpus $\{w_{-i} \cup w_i\}$, and letting n'_{tkd} be the

count for the limited corpus w_{-i}, we get the DCMLDA Gibbs sampling equation:

$$\begin{aligned} p(z_i|z_{-i},w) &= \frac{p(z,w)}{p(z_{-i},w)} \\ &= \frac{B(n_{\cdot\cdot d_i}+\alpha_\cdot)B(n_{\cdot z_i d_i}+\beta_{\cdot z_i})}{B(n'_{\cdot\cdot d_i}+\alpha_\cdot)B(n'_{\cdot z_i d_i}+\beta_{\cdot z_i})} \\ &= \frac{(n_{\cdot z_i d_i}+\alpha_{z_i}-1)(n_{w_i z_i d_i}+\beta_{w_i z_i}-1)}{(\sum_k n_{\cdot k d_i}+\alpha_k-1)(\sum_t n_{t z_i d_i}+\beta_{t z_i}-1)}. \end{aligned}$$

Hyperparameter EM Many applications of LDA are successful using default uniform values for α and β, for example $\alpha = 50/K$ and $\beta = .01$, where K is the number of topics, as suggested by Griffiths and Steyvers (2004). Therefore it is not always necessary to learn the hyperparameters in LDA. However, it is imperative to learn the hyperparameters in DCMLDA. The information contained in the ϕ values with LDA is contained in the β values with the DCMLDA model.

Ideally, we would compute optimal α and β values by maximizing the likelihood $p(w|\alpha,\beta)$ directly. Unfortunately, even evaluating this likelihood is intractable. What can be computed is the complete likelihood $p(w,z|\alpha,\beta)$. Based on this, we use single-sample Monte Carlo EM to learn α and β. The single-sample method is recommended by Celeux, Chaveau, and Diebolt (1996) because it is computationally simple and generally outperforms multiple-sample Monte Carlo EM. Algorithm 2 summarizes the method as applied to DCMLDA.

To implement the M-step of the algorithm we need to find α and β that maximize

Algorithm 2. Single-Sample Monte Carlo EM

Start with initial $\alpha_\cdot$ and $\beta_{\cdot\cdot}$.
repeat
  Run Gibbs sampling to steady-state
  Choose a specific topic assignment for each word using Gibbs sampling
  Choose $\alpha_\cdot$ and $\beta_{\cdot\cdot}$ to maximize complete likelihood $p(w,z|\alpha_\cdot,\beta_{\cdot\cdot})$
until convergence of $\alpha_\cdot$ and $\beta_{\cdot\cdot}$.

Equation 5.3, given the current topic assignments. Expanding the Beta functions yields

$$p(w,z|\alpha,\beta) = \prod_d \left[\frac{(\prod_k \Gamma(n_{\cdot kd}+\alpha_k))\Gamma(\sum_k \alpha_k)}{(\prod_k \Gamma(\alpha_k))\Gamma(\sum_k n_{\cdot kd}+\alpha_k)}\right] \times \prod_{d,k}\left[\frac{(\prod_t \Gamma(n_{tkd}+\beta_{wk})\Gamma(\sum_t \beta_{tk})}{(\prod_t \Gamma(\beta_{tk}))\Gamma(\sum_t n_{tkd}+\beta_{tk})}\right].$$

Now we convert to log-likelihood:

$$\begin{aligned} L(\alpha,\beta;w,z) = &\sum_{d,k}[\log\Gamma(n_{\cdot kd}+\alpha_k) - \log\Gamma(\alpha_k)] \\ &+ \sum_d[\log\Gamma(\sum_k \alpha_k) - \log\Gamma(\sum_k n_{\cdot kd}+\alpha_k)] \\ &+ \sum_{d,k,t}[\log\Gamma(n_{tkd}+\beta_{tk}) - \log\Gamma(\beta_{tk})] \\ &+ \sum_{d,k}[\log\Gamma(\sum_t \beta_{tk}) - \log\Gamma(\sum_t n_{tkd}+\beta_{tk})]. \end{aligned}$$

This is a separable function, since the first term depends only on α and the second only on β. Furthermore, the second term is a sum over topics, so each $\beta_{\cdot k}$ can be independently maximized. This gives a collection of $K+1$ equations to maximize:

$$\begin{aligned} \alpha'_{\cdot} &= \text{argmax}\sum_{d,k}(\log\Gamma(n_{\cdot kd}+\alpha_k) - \log\Gamma(\alpha_k)) \\ &\quad + \sum_d[\log\Gamma(\sum_k \alpha_k) - \log\Gamma(\sum_k n_{\cdot kd}+\alpha_k)] \\ \beta'_{\cdot k} &= \text{argmax}\sum_{d,t}(\log\Gamma(n_{tkd}+\beta_{tk}) - \log\Gamma(\beta_{tk})) \\ &\quad + \sum_d[\log\Gamma(\sum_t \beta_{tk}) - \log\Gamma(\sum_t n_{tkd}+\beta_{tk})]. \end{aligned}$$

Each equation above defines a vector, either $\alpha_{\cdot}$ or $\beta_{\cdot k}$. We use limited memory BFGS (Zhu, Byrd, Lu, & Nocedal, 1997) to perform the maximizations. For one iteration of

EM with 20 topics on S&P500 data explained below, a careful Matlab implementation requires about 100 seconds on a 2.4GHz CPU with 6GB memory.

The implementation of DCMLDA allows both the α vector and β array to be non-uniform. For the DCMLDA model to be useful, β must be non-uniform, since it carries the information that ϕ carries in LDA. The vector α could be uniform in DCMLDA, but learning non-uniform values allows the model to give certain topics higher prior probability than others.

5.4 Experimental Design

Our experimental goal is to test whether the handling of burstiness in a DCMLDA model creates a better topic model than standard LDA. We compare DCMLDA models with LDA models, rather than with more complex topic models, for two reasons. First, DCMLDA and LDA are of comparable conceptual complexity. Second, and more important, they are competing models. DCMLDA is not in competition with more complex topic models, because these models can be modified to include DCM components.

Given a test set of documents not used for training, we estimate the held-out likelihood $p(w|\alpha,\beta)$ for LDA and DCMLDA models. The latter probability uses a vector $\alpha_{\cdot}$ and an array $\beta_{\cdot\cdot}$ learned as described above. The former probability uses $\alpha = \bar{\alpha}_{\cdot}$ and $\beta = \bar{\bar{\beta}}_{\cdot\cdot}$, the scalar means of the values learned by DCMLDA training. We also compare these two models to LDA using the values proposed by Griffiths and Steyvers (2004).

We compare LDA and DCMLDA as models for both text and non-text data. The textual dataset is a collection of papers from the 2002 and 2003 NIPS dataset compiled by Globerson, Chechik, Pereira, and Tishby (2004) and organized by Elkan (2006). This dataset comprises 520955 words (6871 unique word types) in 390 documents. The second is a newly-compiled dataset of stock price fluctuations for the stocks that compose the S&P 500. This dataset contains 501 days of stock transactions between January 2007

and September 2008, with each document being a single day of trading. Each word is a concatenation of a stock symbol and a direction (+ or –), and each day contains one copy of a word for each (rounded) percentage point change between the opening and closing price of the stock. This dataset contains 469642 words in 501 documents. Both datasets are bursty, and approximately equally so, with $B = 2.63$ for NIPS and $B = 2.51$ for S&P500, where B is the burstiness measure from Church and Gale (1995), with $B = 1$ indicating no burstiness and higher values indicating more burstiness.

In analyzing the S&P500 data, the goal is to find groups of companies whose stock prices tend to move together. For example, a learned topic might hypothetically include the words IBM+, MSFT+, and AAPL–. This would indicate that IBM and Microsoft frequently rise together, while Apple tends to fall on the same days. Because different groups of stocks can move independently, each day can be a combination of a different set of topics.

5.5 Empirical Likelihood

Comparing the goodness-of-fit of topic models is a notoriously tricky endeavor. Ideally, we would calculate the incomplete likelihood $p(w|\alpha,\beta)$ for each model and compare those values. However, the incomplete likelihood is intractable for topic models. The complete likelihood $p(w,z|\alpha,\beta)$ is tractable, so previous work (Griffiths & Steyvers, 2004, e.g.) has calculated the harmonic mean of the complete likelihood from the topic assignments generated during Gibbs sampling. This approach is based on a true mathematical identity, but Newton and Raftery (1994) have argued that it is unreliable.

Another possibility is to measure classification accuracy, but that entwines the usefulness of the topics with the separability of the dataset. This is an important consideration because datasets do not always lend themselves to obvious classification schemes. Also, learned topics can be meaningful even if they are not well correlated with

pre-assigned class labels.

We follow a third approach suggested by Li and McCallum (2006). This approach is to approximate the true held-out likelihood with so-called empirical likelihood (EL). To measure EL, we first train each model to obtain its parameter values α and β. These parameter values are then fed into the generative model, and a large set of pseudo documents is produced. Each of these documents has θ and ϕ distributions. (For DCMLDA the ϕ distribution of each document is different, while for LDA they are identical.) The pseudo documents are then used to train a tractable model. In the present case, we use a mixture of multinomials. Following Li and McCallum (2008), each multinomial model is inferred directly from the generated ϕ and θ distributions; individual words are not generated in the pseudo documents. The true likelihood of the test set is then estimated as its likelihood under the tractable model of the pseudo corpus. We report the arithmetic mean of log likelihoods of documents in the test set.

We investigate the stability of EL as a measure of goodness-of-fit by running it multiple times for the same DCMLDA model. Specifically, we train three separate 20-topic DCMLDA models on the S&P500 dataset, and run the EL method five times for each of these models. The mean absolute difference between EL values for the same model is 0.08%, with maximum 0.20%. Furthermore, the mean absolute difference between EL values for separately trained DCMLDA models is 0.11%, with maximum 0.29%, showing that likelihood values are stable over DCMLDA models with the same number of topics. The relationship between empirical likelihood and other measures of goodness-of-fit measures is unclear, but this stability suggests that EL is a sensible measure.

Table 5.1. Sample topics found by a 20-topic DCMLDA model trained on the S&P 500 dataset. The six most likely words for each topic are listed.

"Computer Related"		"Real Estate"	
Stock	Company	Stock	Company
NVDA+	Nvidia	SPG+	Simon Prop.
SNDK+	SanDisk	AIV+	Apt. Invest.
BRCM+	Broadcom	KIM+	Kimco Realty
JBL+	Jabil Circuit	AVB+	AvalonBay
KLAC+	KLA-Tencor	DDR+	Developers
NSM+	Nat'l Semicon.	EQR+	Equity Resid.

Table 5.2. Sample topics found by a 20-topic LDA model trained on the same S&P 500 dataset. The six most likely words for each topic are listed.

"Computer Related"		"Real Estate"	
Stock	Company	Stock	Company
NVDA+	Nvidia	LEN+	Lennar
SNDK+	SanDisk	CTX+	Centex
AMD+	AMD	PHM+	Pulte Homes
MU+	Micron	DHI+	D. R. Horton
BRCM+	Broadcom	KBH+	KB Home
CIEN+	Ciena	PLD+	ProLogis

5.6 Results

An important, but informal, measure of the success of a topic model is the plausibility of the topics that it proposes. Since DCMLDA creates document-specific subtopics based on corpus-level topics, it is fair to ask if these corpus-level topics are as interpretable as LDA topics. Table 5.1 shows two topics from a 20-topic DCMLDA model trained on the S&P500 dataset. The words shown are the most likely based on the rank-order of the β_{tk} values over words t for a given topic k, in the same way that ϕ_{tk} indicates the most likely words for an LDA topic. The topics discovered by DCMLDA generally follow accepted stock classification systems. The 25 most likely stocks in the "computer related" topic are all in the Information Technology sector of the Global Industry Classification Standard (GICS), and 24 of the 25 most likely stocks in the "real estate" topic are in the Financials sector.

The DCMLDA topics are similar to topics from a 20-topic LDA model trained on the same data, as shown in Table 5.2. Three of the top six companies in the computer topic are shared between the models. The LDA topic most similar to the DCMLDA "real estate" topic is also shown; all six top companies in the DCMLDA topic are among the top 15 of the LDA topic. Subjectively, the interpretability of the DCMLDA topics is comparable to the interpretability of the LDA topics. Looking closely suggests that the DCMLDA topics may be better. For example, all six top stocks for the DCMLDA "computer related" topic are suppliers to computer manufacturers, while Ciena in the matching LDA topic is not. In the LDA "real estate" topic the top five stocks are homebuilders but ProLogis is quite different. In contrast, all six stocks in the DCMLDA topic are corporate landlords.

As discussed in Section 5.5, we use empirical likelihood to compare the goodness-of-fit of the DCMLDA and LDA models on the NIPS and S&P500 datasets. We perform five 5-fold cross-validation trials for each number of topics and each dataset. We first

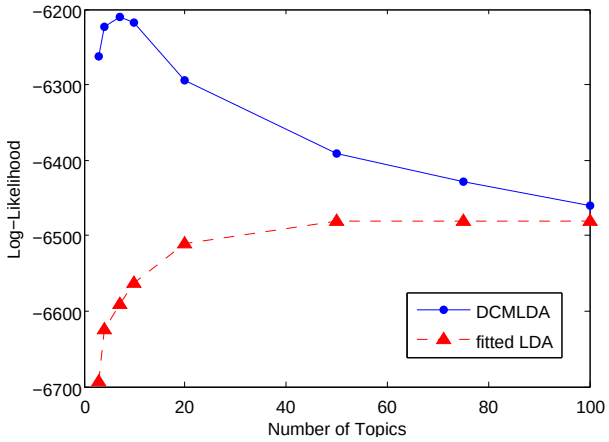


Figure 5.2. Mean per-document log-likelihood on the S&P500 dataset for DCMLDA and fitted LDA models. The heuristic model is omitted here because its likelihood is too low. The maximum standard error is 11.2.

train a DCMLDA model, then create two LDA models. One ("fitted LDA") uses the mean values of the DCMLDA hyperparameters. The other ("heuristic LDA") uses the uniform hyperparameter values suggested by Griffiths and Steyvers (2004). For both datasets, DCMLDA is better than fitted LDA, which in turn is better than heuristic LDA.

Figure 5.2 shows performance on the S&P500 dataset. The highest likelihood comes from DCMLDA with seven topics, where DCMLDA has a major advantage over the fitted LDA model. This supports the idea that a DCMLDA model with few topics is comparable to an LDA model with many topics. This may also indicate that the a natural set of topics for this dataset has cardinality about seven.

Above 100 topics, the likelihood of the fitted LDA model remains approximately constant, while that of DCMLDA continues dropping, ending up lower than that of LDA. This is likely a result of data sparsity preventing the estimation of good β values. As there are only 1000 unique symbols in the dataset, poor behavior with more than 100 topics is not a major source of concern. The likelihoods for heuristic LDA model are not shown in Figure 5.2 because they are much lower than those of the other models, especially

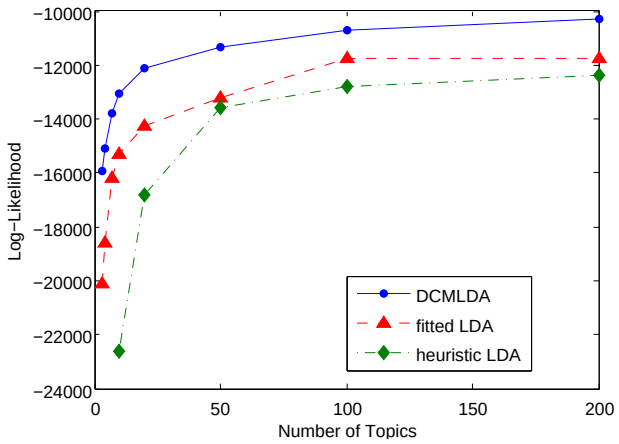


Figure 5.3. Mean per-document log-likelihood on the NIPS dataset for DCMLDA and LDA models. The maximum standard error is 183.6.

when the number of topics is low. For 100 topics, heuristic LDA has mean log-likelihood -7383, which approaches that of the other two models, but for three topics, its mean log-likelihood is -34130.

Figure 5.3 shows performance on the NIPS dataset. For this dataset, the DCMLDA model does not exhibit the few-topics bump seen in the S&P500 dataset. DCMLDA outperforms the fitted LDA model at every tested number of topics. For the NIPS dataset, LDA never surpasses DCMLDA as the number of topics grows, presumably because the larger number of unique words (6871) in this corpus keeps data sparsity from becoming a major issue. LDA with the heuristic hyperparameter values is not as bad on the NIPS dataset as on the S&P500 dataset, almost catching up with the fitted LDA model at 50 topics. This confirms that the suggestions of Griffiths and Steyvers (2004) are reasonable for textual data. However, the fitted LDA model retains a substantial advantage, especially when the number of topics is small.

5.7 Discussion

While the choice of α and β in a topic model is sometimes viewed as a formality, and heuristic values are used without much consideration, we find that heuristic values can lead to much worse likelihood than fitted values, especially when the number of topics is small. Thus learning α and β can be beneficial, and optimized values can be significantly different from previously suggested heuristic values. In addition, we see that accounting for burstiness improves held-out likelihood for both text and non-text data. To be completely confident that the EL improvement is due to modeling burstiness, DCMLDA should be compared also to a version of LDA with a single optimized non-uniform β parameter.

Recent years have seen a profusion of topic model variants, such as the correlated topic model (Blei & Lafferty, 2005) and the Pachinko allocation model (Li & McCallum, 2006). These newer models outperform LDA on many tasks, so comparing the performance of DCMLDA only to that of LDA may seem inappropriate. However, DCMLDA is not in competition with the more complex topic models, but rather with LDA. The more complex topic models share an LDA core, in that they use multinomials to represent topics. These multinomials can be replaced by DCMs to improve, potentially, the performance of these models. Thus the DCMLDA idea and complex topic models are complementary.

5.8 Acknowledgments

The first author was supported in part by NIH Training Grant T32-DC000041. We wish to thank the UCSD Computational Psycholinguistics Lab for insights and advice.

Chapter 5, in full, is an exact copy of the material as it appears in Doyle and Elkan (2009) [Accounting for burstiness in topic models. In L. Bottou and M. Littman (Eds.),

Proceedings of the 26th International Conference on Machine Learning (pp. 281-288). Montreal: Omnipress.] The dissertation author was the primary investigator and author of this paper.

Chapter 6

Conclusion

This dissertation has examined the acquisition of latent linguistic structure from positive-only data using unsupervised computational models. These models span a range of linguistic problems, from human-like acquisition of phonological constraints or word segmentation behavior to applied problems in semantic topic modeling. Despite their range, two major threads connect these models. First, all of the models rely primarily on the observed language data to determine the latent linguistic structure, and argue for a more emergentist approach to language acquisition. Second, the models use appropriate representations of the basic linguistic structure to improve their explanatory power.

Chapter 2 proposed the IBPOT model, a method for learning constraint definitions within Optimality Theory that is driven almost entirely by distributional data, minimizing the amount of innate phonological structure used by a learner. The ability of the IBPOT model to capture basic constraint structure in Wolof vowel harmony shows that distributional data can be highly informative about phonological constraints, and suggests that the standardly-assumed fully-innate OT constraint set may not be necessary. These results are extended and strengthened by Chapter 3, which proposed the Rational Rules OT model. This model introduces a more appropriate representation of constraint structures into the constraint learning problem, and as a result is able to capture more robust constraint definitions, as well as making as accurate of predictions about novel

forms as the fully-innate constraint set does. These results show that a small amount of innate structure can be as informative as a larger amount, if the data is sufficiently informative.

Chapter 4 turned to word segmentation behavior and introduced a model for incorporating multiple segmentation cues coherently. By capturing the information from multiple cues, the model was able to outperform existing single-cue models in segmentation accuracy. In addition, the model showed how multiple cues could trigger changes in segmentation behavior that could not be explained by single-cue models. These changes are driven by the data itself, and capturing the aspects of data that drive this change require appropriate representation and integration of the cues within the model.

Finally, Chapter 5 moved into applications of computational models beyond human learning behavior, and showed that better models of the core linguistic structure can improve application performance as well. The DCMLDA model proposed in this chapter accounts for burstiness in the observed data, and this more appropriate model structure results in improved performance on both linguistic and non-linguistic problems. The high interpretability of the learned topics in this model further supports the richness of language data for determining latent structure without significant innate knowledge.

References

Airoldi, E. M., Fienberg, S. E., & Xing, E. P. (2007). *Mixed membership analysis of genome-wide expression data.* (Arxiv preprint arXiv:0711.2520)

Anderson, J. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anttila, A. (1997). *Variation in finnish phonology and morphology* (Unpublished doctoral dissertation). Stanford U.

Archangeli, D., & Pulleyblank, D. (1994). *Grounded phonology*. MIT Press.

Bernhardt, B., & Stemberger, P. (1998). *Handbook of phonological development*. San Diego: Academic Press.

Bicknell, K. (2011). *Eye movements in reading as rational behavior* (Unpublished doctoral dissertation). University of California, San Diego.

Blake, A., Bülthoff, H. H., & Sheinberg, D. (1993). Shape from texture: Ideal observers and human psychophysics. *Vision research*, *33*(12), 1723–1737.

Blanchard, D., & Heinz, J. (2008). Improving word segmentation by simultaneously learning phonotactics. In *Proceedings of CoNLL* (pp. 65–72).

Blei, D., & Lafferty, J. (2005). Correlated topic models. In *Advances in Neural Information Processing Systems 18* (Vol. 18, p. 147-154).

Blei, D., Ng, A., & Jordan, M. (2001). Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14* (Vol. 14, pp. 601–608).

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *J. of Machine Learning Research*, *3*, 993-1022.

Boersma, P. (1998). *Functional phonology*. Holland Academic Graphics.

Boersma, P. (1999). Optimality-theoretic learning in the Praat program. In *Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam.*

Boersma, P., & Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, *32*, 45–86.

Borfeld, H., Morgan, J., Golinkoff, R., & Rathbun, K. (2005). Mommy and me: familiar names help launch babies into speech-stream segmentation. *Psychological Science*, *16*(4), 298–304.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning*, *34*, 71–105.

Celeux, G., Chaveau, D., & Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *J. of Statistical Computation and Simulation*, *55*, 287–314.

Chater, N., & Oaksford, M. (2008). The probabilistic mind: Prospects for Bayesian cognitive science. In N. Chater & M. Oaksford (Eds.), (pp. 59–75). Oxford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and Cognitive Processes*, *13*, 221–268.

Church, K., & Gale, W. A. (1995). Poisson mixtures. *Natural Language Engineering*, *1*, 163–190.

Cole, R., & Jakimik, J. (1980). A model of speech perception. In *Perception and production of fluent speech* (pp. 136–163). Hillsdale, NJ: Erlbaum.

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the English vocabulary. *Comp. Speech Lang.*, *2*, 133–142.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavioral Research Methods, Instruments, and Computers*, *28*, 125–127.

Danks, D. (2008). The probabilistic mind: Prospects for Bayesian cognitive science. In N. Chater & M. Oaksford (Eds.), (pp. 59–75). Oxford University Press.

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present.*

Dekker, P., & van Rooy, R. (2000). Bi-directional Optimality Theory: An application of game theory. *Journal of Semantics*, *17*, 217–242.

de Lacy, P. (2004). Markedness conflation in Optimality Theory. *Phonology*, *21*, 145–199.

Della Pietra, S., Della Pietra, V., & Lafferty, J. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *19*, 380–393.

Doyle, G., Bicknell, K., & Levy, R. (2014). Nonparametric learning of phonological constraints in Optimality Theory. In *Proceedings of the Association for Computational Linguistics.*

Doyle, G., & Elkan, C. (2009a). Accounting for burstiness in topic models. In L. Bottou & M. Littman (Eds.), *Proceedings of the 26th International Conference on Machine Learning* (pp. 281–288).

Doyle, G., & Elkan, C. (2009b). Financial topic models. In *Proceedings of the NIPS Applications for topic models: Text and beyond workshop.*

Doyle, G., & Levy, R. (2013). Combining multiple information types in Bayesian word segmentation. In *Proceedings of the 2013 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 117–126).

Eimas, P. (1999). Segmental and syllabic representations in the perception of speech by young infants. *Journal of the Acoustic Society of America*, *105*, 1901–1911.

Elkan, C. (2006). Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 289–296).

Elsner, M., Goldwater, S., & Eisenstein, J. (2012). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics.*

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Vol. 2, pp. 524–531).

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Conference on Cognitive Science.*

Flack, K. (2007). *The sources of phonological markedness* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Frank, M., Goldwater, S., Griffiths, T., & Tenenbaum, J. (2010). Modeling human

performance in statistical word segmentation. *Cognition*, *117*, 107–125.

Frank, M., Goodman, N., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 579–585.

Frank, R., & Satta, G. (1998). Optimality theory and the generative complexity of constraint violability. *Computational Linguistics*, *24*, 307–315.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*(6471), 542–545.

Gambell, T., & Yang, C. (2006). *Word segmentation: Quick but not dirty.* (Unpublished manuscript)

Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2004). Euclidean embedding of co-occurrence data. In *Advances in Neural Information Processing Systems 17* (Vol. 17, p. 497-504).

Goad, H., & Rose, Y. (2004). Input elaboration, head faithfulness, and evidence for representation in the acquisition of left-edge clusters in West Germanic. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in phonological acquisition* (pp. 109–157). Cambridge Univ. Press.

Goldwater, S. (2007). *Nonparametric Bayesian models of lexical acquisition* (Unpublished doctoral dissertation). Brown Univ.

Goldwater, S., Griffiths, T., & Johnson, M. (2006). Contextual dependencies in unsupervised word segmentation. In *Proceedings of Coling/ACL.*

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, *112*, 21–54.

Goldwater, S., & Johnson, M. (2003). Learning OT constraint rankings using a Maximum Entropy model. In *Proceedings of the workshop on variation within optimality theory.*

Goodman, N., Tenebaum, J., Feldman, J., & Griffiths, T. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*, 108–154.

Görür, D., Jäkel, F., & Rasmussen, C. (2006). A choice model with infinitely many latent features. In *Proceedings of the 23rd International Conference on Machine Learning.*

Griffiths, T., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian*

buffet process (Tech. Rep. No. 2005-001). Gatsby Computational Neuroscience Unit.

Griffiths, T., & Ghahramani, Z. (2011). The Indian Buffet Process: An introduction and review. *Journal of Machine Learning Research*, *12*, 1185-1224.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *104*((suppl. 1)), 5228–5235.

Griffiths, T., Steyvers, M., Blei, D., & Tenenbaum, J. (2004). Integrating topics and syntax. In *Advances in neural information processing systems 17* (Vol. 17, pp. 537–544).

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies* (pp. 1–8).

Hayes, B. (1999). Phonetically driven phonology: the role of optimality theory and inductive grounding. In M. Darnell, E. Moravcsik, M. Noonan, F. Newmeyer, & K. Wheatley (Ed.), *Formalism and functionalism in linguistics, vol. 1.* Benjamins.

Hayes, B., & Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, *39*, 379-440.

Heinrich, G. (2005). *Parameter estimation for text analysis.* (Unpublished manuscript)

Heinz, J., Kobele, G., & Riggle, J. (2009). Evaluating the complexity of Optimality Theory. *Linguistic Inquiry*, *40*, 277–288.

Hyde, B. (2012). Alignment constraints. *Natural Language and Linguistic Theory*, *30*, 789–836.

Idsardi, W. (2006). A simple proof that Optimality Theory is computationally intractable. *Linguistic Inquiry*, *37*, 271-275.

Johnson, E., & Jusczyk, P. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *J. of Memory and Language*, *44*, 548–567.

Jusczyk, P., Cutler, A., & Redanz, N. (1993). Preference for predominant stress patterns of English words. *Child Development*, *64*, 675–687.

Jusczyk, P., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kager, R., Pater, J., & Zonneveld, W. (2004). Introduction: Constraints in phonolog-

ical acquisition. In R. Kager, J. Pater, & W. Zonneveld (Eds.), *Constraints in phonological acquisition* (pp. 1–53). Cambridge Univ. Press.

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values, and frames*. Cambridge Univ. Press.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality* (Unpublished doctoral dissertation). U. of Edinburgh.

Körding, K. P., & Wolpert, D. M. (2004). Bayesian integration in sensorimotor learning. *Nature*, *427*(6971), 244–247.

Korman, M. (1984). Adaptive aspects of maternal vocalizations in differing contexts at ten weeks. *First language*, *5*, 44–45.

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the International Conference on Machine Learning.*

Legendre, G., Miyata, Y., & Smolensky, P. (1990). Can connectionism contribute to syntax?: Harmonic grammar, with an application. In *Proceedings of the 26th meeting of the Chicago Linguistics Society.*

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 577–584).

Li, W., & McCallum, A. (2008). *Pachinko allocation: Scalable mixture models of topic correlations.* (Unpublished manuscript)

Madsen, R., Kauchak, D., & Elkan, C. (2005). Modeling word burstiness using the Dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 545–552).

Marr, D. (1982). *Vision*. W. H. Freeman.

Maurits, L., Perfors, A., & Navarro, D. (2009). Joint acquisition of word order and word reference. In *Proceedings of 31st Annual Conference of the Cognitive Science Society.*

McCarthy, J. (2008). *Doing Optimality Theory*. Blackwell.

McCarthy, J., & Prince. (1993). Generalized alignment. In *Yearbook of Morphology 1993.*

Merchant, N., & Tesar, B. (2005). Learning underlying forms by searching restricted lexical subspaces. In *Proceedings from the annual meeting of the Chicago Linguistic Society.*

Miyazaki, M., Nozaki, D., & Nakajima, Y. (2005). Testing Bayesian models of human coincidence timing. *Journal of Neurophysiology*, *94*(1), 395–399.

Navarro, D., & Griffiths, T. (2007). A nonparametric Bayesian method for inferring features from similarity judgments. In *Advances in Neural Information Processing Systems 19.*

Neal, R. (1994). Response to approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*, 3–48.

Newton, M., & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, *56*, 3–48.

Parker, S. (2002). *Quantifying the sonority hierarchy* (Unpublished doctoral dissertation). University of Massachusetts, Amherst.

Pater, J. (1997). Minimal violation and phonological development. *Language Acquisition*, *6*, 201–253.

Pater, J., & Werle, A. (2003). Direction of assimilation in child consonant harmony. *Canadian Journal of Linguistics*, *48*, 385–408.

Peters, A. (1983). *The units of language acquisition: Monographs in applied psycholinguistics*. Cambridge Univ. Press.

Phillips, L., & Pearl, L. (2012). Less is more in Bayesian word segmentation: When cognitively plausible learners outperform the ideal. In *Proceedings of the 34th annual conference of the Cognitive Science Society.*

Poon, H., Cherry, C., & Toutanova, K. (2009). Unsupervised morphological segmentation with log-linear models. In *Proceedings of the North American chapter of the Association for Computational Linguistics.*

Prince, A., & Smolensky, P. (1993). *Optimality Theory: Constraint interaction in generative grammar* (Tech. Rep.). Rutgers Center for Cognitive Science.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia* (pp. 251–260).

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive Bayes text classifiers. In *Proceedings of 20th International Conference on Machine Learning* (pp. 616–623).

Riggle, J. (2006). Using entropy to learn OT grammars from surface forms alone. In *Proceedings of the 25th West Coast Conference on Formal Linguistics.*

Riggle, J. (2009). Generating contenders. *Rutgers Optimality Archive*, *1044*.

Riggle, J. (2012, December). *Phonological feature chart (v. 12.12).*

Rytting, C. A., Brew, C., & Fosler-Lussier, E. (2010). Segmenting words from natural speech: subsegmental variation in segmental cues. *Journal of Child Language*, *37*, 513–543.

Saffran, J., Aslin, R., & Newport, E. (1996). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Shafir, E. B., Smith, E. E., & Osherson, D. N. (1990). Typicality and reasoning fallacies. *Memory & Cognition*, *18*(3), 229–239.

Sher, S., & McKenzie, C. (2008). The probabilistic mind: Prospects for Bayesian cognitive science. In N. Chater & M. Oaksford (Eds.), (pp. 79–96). Oxford University Press.

Smith, J. (2004). Making constraints compositional: toward a compositional model of Con. *Lingua*, *114*, 1433–1464.

Smith, N. (1973). *The acquisition of phonology: a case study*. Cambridge Univ. Press.

Smith, N., & Eisner, J. (2005). Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd meeting of the Association for Computational Linguistics.*

Stevens, K. (1998). *Acoustic phonetics*. MIT Press.

Stokoe, W. (1960). *Sign language structure*. Linstok Press.

Swingley, D. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology*, *50*, 86–132.

Tesar, B., & Smolensky, P. (2000). *Learnability in optimality theory*. MIT Press.

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Tu, Z., Chen, X., Yuille, A. L., & Zhu, S.-C. (2005). Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, *63*(2), 113–140.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience*, *5*, 598–604.

Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*, 49–63.

Zhu, C., Byrd, R. H., Lu, P., & Nocedal, J. (1997). Algorithm 778: L-BFGS-B: Fortran routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software*, *23*(4), 550–560.