THE PROBABILISTIC TYPOLOGY OF VOWEL SYSTEMS

by

Ryan David Cotterell

A dissertation submitted to The Johns Hopkins University
in conformity with the requirements for the degree of
Doctor of Philosophy

Baltimore, Maryland
July, 2021

Abstract

Linguistic typology studies the range of structures present in human language. The main goal of the field is to discover which sets of possible phenomena are universal, and which are merely frequent. For example, all languages have vowels, while most—but not all—languages have an [u] sound. In this paper we present the first probabilistic treatment of a basic question in phonological typology: What makes a natural vowel inventory? We introduce a series of deep probability models.

In Chapter 1, we give an overview of the relevant background material in phonetics and the typology of vowel systems. In Chapter 2, we introduce a series of deep stochastic point processes, and contrast them with previous computational, simulation-based approaches. We provide a comprehensive suite of experiments on over 200 distinct languages. In Chapter 3, we work directly with the acoustic information—the first two formant values—rather than modeling discrete sets of phonemic symbols (IPA). We develop a novel generative probability model and report results based on the same corpus of over 200 languages. In Chapter 4, we focus on a functional account of vowel systems. The typology of vowel systems can, in part, be explained in part by functional pressures on communication. We find that a model of vowel token

formants is more predictive of held-out data if it is trained with the help of this prior (that is, by MAP rather than ML).

Thesis Committee

Jason Eisner (Primary Advisor)
Professor
Department of Computer Science
Johns Hopkins University

Colin Wilson
Professor
Department of Cognitive Science
Johns Hopkins University

Hal Daumé
Professor
Department of Computer Science
University of Maryland, College Park

Acknowledgments

First and foremost, I would like to acknowledge my advisor, Jason Eisner, who taught me so much over the years. He is a shining role model of an academic advisor. I would also like to thank the members of Jason's lab, jocularly titled the Argo, and other members of the broader Center for Language and Speech Processing for providing such an intellectually nurturing community. Special thanks, of course, must go to Timmy (Tim Vieira) who taught me so much. Next, I would like to acknowledge the funding source during my PhD. My PhD has been funded by the Fulbright Commission, Deutscher Akademischer Austauschdienst, and NDSEG fellowship and a Facebook Fellowship. Finally, I would like to thank my parents and my sister who have been so supportive over the years.

Previous Publications

Portions of this thesis have been published in slightly different forms in the following co-authored papers:

Chapter 2 Ryan Cotterell and Jason Eisner (2017), "Probabilistic Typology: Deep Generative Models of Vowel Inventories". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Chapter 3 Ryan Cotterell and Jason Eisner (2018). "A Deep Generative Model of Vowel Formant Typology". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Chapter 4 Ryan Cotterell and Jason Eisner (2021). "A Functionalist Account of Vowel System Typology". Under Submission.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Human languages exhibit a wide range of grammatical phenomena. Linguistic diversity is extensive, but also constrained or limited across languages: some grammatical structures occur or co-occur more frequently than others. Organizing and quantifying the range of natural variation in linguistic structures falls under the study of linguistic typology. This area attempts to describe and understand linguistic diversity and linguistic universals, which are typologically frequent linguistic patterns.

As an example, we can consider the basic typological universal that every language contains both vowels and consonants (Velupillai, 2012). This particular universal is, as far as we know, absolute, in that the description of the linguistic phenomenon has been observed in every language thus far studied. We can also consider linguistic universals that are statistical in nature, such as, which vowel types tend to occur and co-occur across languages? As is well-known from everyday experience and from the study of language-specific phonology, languages vary in the set of vowel categories, or the language-specific vowel *inventory*. While most languages have the vowel /i/ (as in English 'fleece'), only a handful of languages have the vowel /ɤ/, which is close in articulation to /o/ (as in English 'goat'), but without lip-rounding (Maddieson, 2013). Certain vowel categories are cross-linguistically more frequent, and certain co-occurrences of vowel categories are also universally more likely than others. This dissertation investigates the range of vowel systems that languages exhibit and proposes several probability models to account for the cross-linguistic patterns in vowel type frequency and co-occurrence. These models are then used to make concrete typological claims about language.

In the following sections, I describe and discuss several prerequisite concepts and assumptions that are necessary to construct probabilistic models of vowel inventories. In section 1.2, I provide a brief overview of phonology and phonetics, covering the critical questions of 1) what is a vowel category, and 2) what is the phonetic realization of a vowel category. The first question concerns itself with the phonology, or the study of speech sound patterns, and the second with phonetics, or the study of the physical properties of speech sounds. In section 1.3, I discuss linguistic typology with a strong focus on the linguistic typology of vowel inventories. Section 1.4 motivates the move to probabilistic modeling of typological patterning, and section 1.5 catalogs terminology used in the dissertation.

1.2 Phonology and Phonetics

Phonology and phonetics address aspects of the sounds of language: phonology with the abstract patterns and organization of speech sounds and phonetics with the physical realization of speech sounds. Traditionally, phonology is assumed to exist as an abstract level of representation that interfaces with a physical realization of those speech sounds, or the phonetic level of representation. Though these levels could be represented in the minds of speakers, the current proposal simply requires a distinction between the two levels, and that these two levels interface with one another.

1.2.1 Phonology and Vowel Phonemes

Phonology is a broad field covering how sounds pattern together in a spoken language.[1] Individual sounds of language naturally differ from one another on many physical dimensions—I return to this in the discussion of phonetics. Some physical differences, however, matter more than others. One primary goal of phonology is to determine the set of basic sound units within a language, or the sound inventory. How a basic unit of sound is defined can differ substantially across approaches to the study of phonology. In general, determining a basic unit of sound relates to establishing which sounds are meaningfully contrastive and which ones are equivalent. A traditional approach to determining meaningful contrast among sounds is to identify minimal pairs of words that differ by only one speech sound. As the sound difference gives rise to a meaningful difference at the lexical level, the speech

[1] The notion of phonology exists in signed language, but is not based on sound patterns.

sounds are assumed to be phonologically distinct.

For example, the way the sound /æ/ in the English word 'cat' will physically differ between two speakers (e.g., your voice and my voice), the physical changes in the vowel quality do not give rise to a lexical difference. At the phonological level of representation, these unique instantiations are considered equivalent. However, the English words 'cat', 'kit', and 'cot' are all distinct in meaning and differ from one another in one abstract unit of sound, the vowel. This diagnosis then leads to the conclusion that the vowels are meaningfully distinct units of sound in English, and are respectively the categories /æ/, /ɪ/, and /a/.

Throughout the dissertation, I will refer to these abstract categories of sound as phonemes, though it is to be acknowledged that the term phoneme has taken on myriad definitions and connotations within the literature (Chomsky and Halle, 1965; Dresher, 2011). I take the term **phoneme** to be a discrete phonological unit that abstracts away from individual phonetic values.[2] I contrast this with the term **phone**, which is a specific rendering of a sound. Under this definition, a phoneme is a discrete unit, whereas a phone is a continuous unit, closer to the phonetic realization. This distinction will become important as we progress from modeling phonology (Chapter 2) to phonetics (Chapters 3 and 4). For more introduction to phonology, I refer the reader to Hayes (2011).

[2] Note that we use the term phoneme in the way that many other linguists may use the term segment. However, I will consistently use the term phoneme for this concept.

1.2.2 Phonetics and Vowel Formants

The realization of language is physical in nature: every basic unit is articulated and in the case of speech, takes on an acoustic form. The impressionistic and quantitative analyses of these physical aspects of language fall under branch of linguistics known as **phonetics**. While phonetics encompasses the articulatory, auditory, and acoustic aspects of speech, the current dissertation focuses primarily on the acoustic description of speech sounds. This focus derives in part from the availability and accessibility of acoustic descriptions of speech, and also because of its approximate (though not direct) relation to the auditory perception of speech. For a more comprehensive introduction to phonetics, I refer the reader to Ladefoged and Maddieson (1996).

Spoken language is transmitted through sound waves, a continuous quantity, from speakers to hearers: one conversant emits a sound wave through the articulation of speech which is in turn comprehended by a second. Acoustic phonetics deals with the properties of those waves, and attempts to identify appropriate correlates between the sound wave and relevant units of speech. Here, I consider the nature of the portions of such sound waves that correspond to vowels. I briefly review the relevant bits of acoustic phonetics so as to give an overview of the data we are actually modeling and develop our notation.

The sound wave that carries spoken language is a function from time to sound pressure variation in the air. To distinguish vowels, it is helpful to transform this function into its constituent frequencies by using a short-time Fourier transform (Deng and O'Shaughnessy, 2003, Chapter 1). This decomposes each

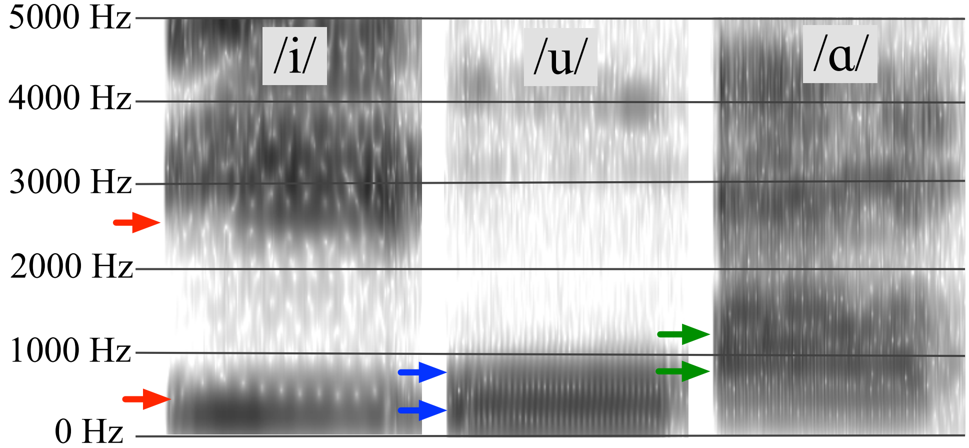

Figure 1.1: Example spectrogram of the three English vowels: [i], [u] and [ɑ]. The x-axis is time and y-axis is frequency. The first two formants F_1 and F_2 are marked in with colored arrows for each vowel. We used the Praat toolkit to generate the spectrogram and find the formants (Boersma and Weenink, 2019).

short interval of the wave function into a weighted sum of sinusoidal waves of different frequencies (measured in Hz). The Fourier transform thus returns the distribution of energy across the frequency spectrum. The constituent frequencies can be visually represented over time, or in three dimensions, by a **spectrogram** (Figure 1.1). At each interval, the darkness of the spectrogram corresponds to the weights, and in this figure the power spectral density (in Pa^2 / Hz), of the frequencies.

In acoustic-phonetic analysis, speech sounds are described with a set of acoustic correlates to the identification and discrimination of the sound unit. For vowel sounds, the most common correlates considered are the **formants**—local maxima of the (smoothed) frequency spectrum. These strong peaks of energy reflect resonant cavities in the vocal tract. Vowel sounds are articulated with continuous, unobstructed airflow, and the quality of the vowel is altered by changing the shape of the oral cavity with the tongue and lips. The differences in oral cavity shape then result in different resonant cavities, which are acoustically observed as different formant patterns. The first two formants F_1 and F_2 have been shown to be most informative for acoustic and auditory identification of vowels, which I discuss in more detail below (Peterson and Barney, 1952; Cooper et al., 1952; Fry et al., 1962). Additional correlates to vowel identity include higher formant values in the spectrum, such as F_3 and F_4, formant transitions from surrounding context, the fundamental frequency F_0 (which corresponds to the pitch), the vowel duration, among others (House and Fairbanks, 1953; Flanagan, 1957; Lehiste and Peterson, 1961; Lindblom and Studdert-Kennedy, 1967; Sussman, 1990; Whalen and Levitt, 1995).

The first two formants also map reasonably well onto a two-dimensional articulatory configuration of the tongue. This can be observed in the vowel chart in (Figure 1.2). The x-axis corresponds to tongue backness and F_2, and the y-axis to tongue height and F_1. From right to left, vowels are produced with a more fronted tongue shape and F_2 increases in value; from bottom to top, vowels are produced with increased tongue height and lower F_1 values. The vowel /i:/ therefore has a high and front tongue shape along with a low F_1 and a high F_2, whereas the vowel /a:/ has a low and back tongue shape with a high F_1 and a low F_2.

In terms of vowel recognition, humans can identify vowel categories with a high degree of accuracy using the first two formants of the sound wave alone (Cooper et al., 1952; Fry et al., 1962; Ladefoged, 2001, Chapter 5). This dissertation capitalizes on this observation, though it is important to note that additional correlates, or cues to vowel identity can be critical to vowel

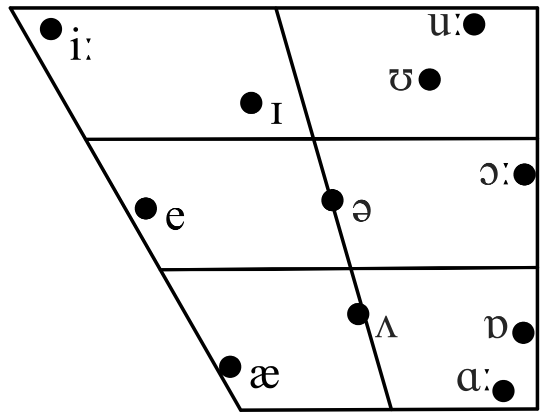

Figure 1.2: The standard vowel table in IPA for the Received Pronunciation accent of British English. The x-axis indicates the front-back spectrum and the y-axis indicates the high-low distinction.

identification. The two-formant assumption can break down in edge cases: e.g., the third formant F_3 helps to distinguish the roundness of the vowel (Ladefoged, 2001, Chapter 5). Other non-formant features may also play a role. For example, in tonal languages, the same vowel may be realized with different tones (which are signaled using F_0): Mandarin Chinese makes a distinction between mǎ (*horse*) and má (*hemp*) without substantially modifying the quality of the vowel /a/ (Gottfried and Suiter, 1997). Other features, such as creaky voice, can play a role in distinguishing vowel phonemes (Silverman et al., 1995; Garellek and Keating, 2011). We do not explicitly model any of these aspects of vowel space, limiting ourselves to (F_1, F_2) as in previous work (Liljencrants and Lindblom, 1972). However, it would be easy to extend all the models we will propose here to incorporate such information, given appropriate datasets.

1.3 The Typology of Vowel Inventories

Vowel inventories are perhaps one of the simplest entry points into the study of linguistic typology. Every spoken language has vowels, and vowel inventories contain between 3 and 46 vowels (mean of 8.7 and the majority of languages possessing between 5 and 7) (Gordon, 2016). Nevertheless, the empirical distribution over vowel inventories is remarkably peaked, i.e., there are only a handful of unique 4-vowel systems attested despite many possibilities. Crothers (1978) notes that all languages have the vowels /i a u/, or the corner vowels, though some languages realize each individual phoneme as a close variant (e.g., exchanging the rounded /u/ for the unrounded variant, /ɯ/). In addition, back vowels tend to be produced with lip rounding whereas front vowels are typically unrounded. Moreover, approximately 10 vowel inventory structures account for the inventories of almost 80% of observed languages (Crothers, 1978; Roark, 2001; Becker-Kristal, 2010). Reigning linguistic theory (Becker-Kristal, 2010) has maintained that vowel systems are guided by two core principles—dispersion and focalization—which we will address below.

1.3.1 Dispersion

The dispersion criterion, first proposed by Liljencrants and Lindblom (1972), states that the phonemes of a language must be "spread out" so that they are easily distinguishable by a listener. This presumes that the listener maps acoustic signals into internal perceptual representations, which fall in a perceptual space that has its own distance metric. A language seeks phonemes that are sufficiently distant from one another in perceptual space to avoid confusion. As noted in Liljencrants and Lindblom (1972), the principle of maximum contrast has a long history in linguistics (Jakobson, 1941). Lindblom (1986) later relaxed the notion of maximum contrast to "sufficient" contrast, where a good vowel system simply has to be contrastive enough to be functional—not necessarily maximally.

1.3.2 Focalization

The dispersion theory alone, however, does not seem to capture the whole story. In addition to vowel systems maximizing contrast, certain vowels are more popular cross-linguistically. That is, properties inherent in a vowel may make that vowel preferable independent of its context within a sound inventory. One explanation for the preference of some vowels over others is the *quantal theory of speech* (Stevens, 1989). Quantal theory states that certain regions of the acoustic space are relatively invariant to changes in articulation. Small variations in articulator position will have minimal impact on the acoustic realization. Such vowels may be preferable due to their relative acoustic stability. One common type of quantal space is when two or more formants within a vowel are relatively close together. All three universally attested vowels, /i a u/, exhibit some degree of formant proximity: in /i/ F_2, F_3, and F_4 are nearly merged, and in /a/ and /u/, both F_1 and F_2 are very close to each other. Close proximity between formants can result in the percept of a single, merged formant (Chistovich and Lublinskaya, 1979), which means that the speaker may have greater freedom in articulation, as long as the end result is a single, merged formant (Stevens, 1989). In general, languages will prefer these vowels over others. Thus, focalization adds the following to the theory of vowel inventories: A good vowel system should consist of vowels that are inherently easy to pronounce and distinguish.

1.3.3 Dispersion–Focalization Theory

The dispersion-focalization theory (DFT) (Schwartz et al., 2005) combines the notions of dispersion and focalization. Stated simply, a good vowel system now consists of vowels that are individually desirable *and* contrast with one another. For a thorough review, see Schwartz et al. (1997). The closest work to ours is the computer simulations performed in Roark (2001), which first proposed jointly integrating dispersion and focalization, i.e., preference for the quantal vowels, into a computational model. We take a cleaner approach, however, in providing a probabilistic treatment of DFT based on deep point process models. We additionally describe new evaluation techniques for future probabilistic and non-probabilistic treatments of vowel inventory typology.

1.4 The Move to Probabilistic Modeling

The primary methodological innovation of this thesis lies at a philosophical level. Independent of any specific model we describe, we propose a change in the mannner, by which we discuss typology. Much work has focused on a simulation-based paradigm of research. We believe, however, that a more natural way to discuss typology is through probability theory. We will argue that a typologist should operate in the following paradigm: they should construct a probability model of the specific lingusitic structure they wish to explain. In our case, we will discuss the construction of probability distributions over vowel systems.

1.4.1 Terminological Zoo

In this section, I catalog the terminology and notation I will use throughout the dissertation to discuss the probabilistic models we introduce. As mentioned above, I take the term **phoneme** to be a discrete phonological unit that abstracts away from individual phonetic values. For instance, the first phoneme in the word [kʰæt] is /kʰ/. We contrast this with the term **phone**, which is a specific rendering of a sound. Under this definition, a phoneme is a discrete unit; I contrast this with a phone, a continuous unit, closer to the phonetic realization.

For a given language ℓ, I will write n_ℓ for the number of vowel phonemes the inventory has. I will define the following two sets:

- The set $\mathcal{V}$ consists of all IPA symbols.
- The set $\bar{\mathcal{V}}$ consists of all possible phone realizations.
- The set $\bar{\mathcal{V}}^{\mathcal{T}}$ consists of all possible sequences of phone realization of length T.

As should be clear, the progression is from modeling more abstract quantities, e.g. sets of phonological units, to more concrete, observed quantities, e.g. sequences of phonetic units. Whereas modeling sets of phonemes requires linguistically trained annotators to have annotated each language, simply modeling the sequences of phonetic measurements is considerably more a-theoretical. This is, arguably, a good thing, as the linguistic theory under which the data is annotated can alter the results of any empirical enterprise.

1.4.2 Three Models of Vowel Typology

Put in a phrase, the new idea in this thesis is the use of tools from probabilistic modeling to study vowel systems. We present a sequence of three probabilistic models, each increasing in complexity, to study the manner in which languages choose their vowels. Specifically, each model addresses short-comings of the previous model. We present each model in turn. In chapter 2, we present a distribution over subsets of phonological typess $\mathcal{V}$. In chapter 3, we present a distribution over subsets of phonetic types $\bar{\mathcal{V}}$. Finally, in chapter 4, we present a distribution over sequences of phonetic tokens $\bar{\mathcal{V}}^T$.

Chapter 2

Modeling Sets of Segments

2.1 Introduction

The first series of models in this thesis deals with distributions over sets of segments $\mathcal{V}$. We discuss the construction of three point process models that give us a distribution over the powerset of $2^{\mathcal{V}}$. These models are more phonological in nature, than phonetic, as they construct distributions over phonemes. We will bring phonetic information into the models by conditioning on formant values. This chapter should be read as a first step towards the larger aim of a distribution over possible phonetic linguistic systems. The primary result of this chapter is that stochastic point processes are a useful formal framework for considering the phonological typology of vowel systems. Indeed, a certain class of point processes, e.g. determinantal point processes, naturally model the phonetic principles of focalization and dispersion.

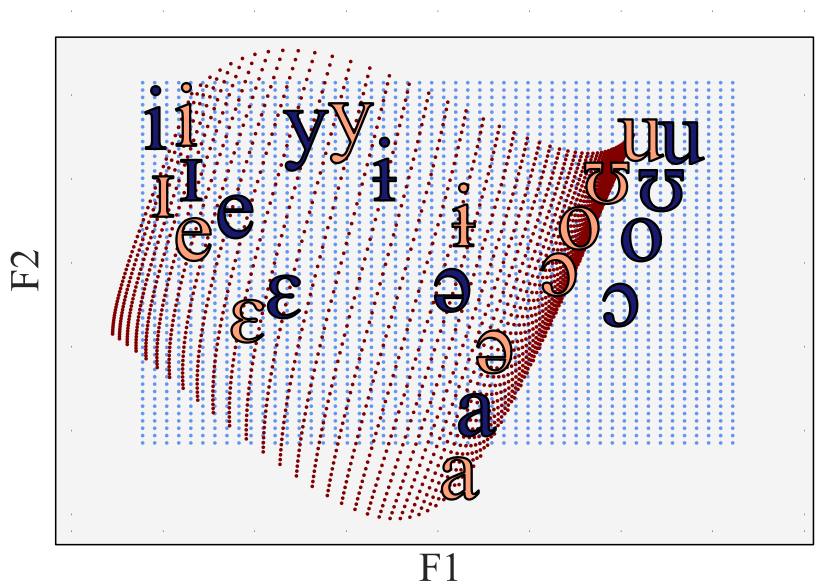

Figure 2.1: The transformed vowel space that is constructed within one of our deep generative models (see §2.6.1). A deep network nonlinearly maps the blue grid ("formant space") to the red grid ("metric space"), with individual vowels mapped from blue to red position as shown. Vowel pairs such as [ə]–[ɔ] that are brought close together are anti-correlated in the point process. Other pairs such as [y]–[ɨ] are driven apart. For purposes of the visualization, we have transformed the red coordinate system to place red vowels near their blue positions—while preserving distances up to a constant factor (a "Procrustes transformation").

2.2 Point Process Models

The technical exposition here starts with background on stochastic point processes. Given a base set $\mathcal{V}$, a point process is a distribution over its subsets.[1] In this paper, we take $\mathcal{V}$ to be the set of all IPA symbols corresponding to vowels. Thus a draw from a point process is a vowel inventory $V \subseteq \mathcal{V}$, and the point process itself is a distribution over such inventories. We will consider three basic point process models for vowel systems: the Bernoulli Point Process, the Markov Point Process and the Determinantal Point Process. In this section, we review the relevant theory of point processes, highlighting aspects related to §1.3.

2.2.1 Bernoulli Point Processes

Taking $\mathcal{V} = \{v_1, \ldots, v_N\}$, a Bernoulli point process (BPP) makes an *independent* decision about whether to include each vowel in the subset. The probability of a vowel system $V \subseteq \mathcal{V}$ is thus

$$p(V) \propto \prod_{v_i \in V} \phi(v_i), \tag{2.1}$$

where ϕ is a unary potential function, i.e., $\phi(v_i) \geq 0$. Qualitatively, this means that $\phi(v_i)$ should be large if the i^{th} vowel is good in the sense of §1.3.2, i.e. if the vowel is focalized. Marginal inference in a BPP is computationally trivial. The probability that the inventory V contains v_i is $\phi(v_i)/(1 + \phi(v_i))$, independent of the other vowels in V. Since a BPP predicts each vowel

[1] A point process is a specific kind of stochastic process, which is the technical term for a distribution over *functions*. Under this view, drawing some subset of $\mathcal{V}$ from the point process is regarded as drawing some *indicator function* on $\mathcal{V}$.

independently, it only models focalization. Thus, the model provides an appropriate baseline that will let us measure the importance of the dispersion principle—how far can we get with just focalization? A BPP may still tend to generate well-dispersed sets if it defines ϕ to be large only on certain vowels in $\mathcal{V}$ and these are well-dispersed (e.g., [i], [u], [a]). More precisely, it can define ϕ so that $\phi(v_i)\phi(v_j)$ is small whenever v_i, v_j are similar.[2] But it cannot actively encourage dispersion: including v_i does not lower the probability of also including v_j.

2.2.2 Markov Point Processes

A Markov Point Process (MPP; Van Lieshout, 2000)—also known as a Boltzmann machine (Ackley, Hinton, and Sejnowski, 1985; Hinton and Sejnowski, 1986)—generalizes the BPP by adding pairwise interactions between vowels. The probability of a vowel system $V \subseteq \mathcal{V}$ is now

$$p(V) \propto \prod_{v_i \in V} \phi(v_i) \prod_{v_i, v_j \in V} \psi(v_i, v_j), \tag{2.2}$$

where each $\phi(v_i) \geq 0$ is, again, a unary potential that scores the quality of the i^{th} vowel, and each $\psi(v_i, v_j) \geq 0$ is a binary potential that scores the combination of the i^{th} and j^{th} vowels. Roughly speaking, the potential $\psi(v_i, v_j)$ should

[2] We point out that such a scheme would break down if we extended our work to cover fine-grained phonetic modeling of the vowel inventory. In that setting, we ask not just whether the inventory includes /i/ but exactly which pronunciation of /i/ it contains. In the limit, ϕ becomes a function over a *continuous* vowel space $\mathcal{V} = \mathbb{R}^2$, turning the BPP into an inhomogeneous spatial Poisson process. A continuous ϕ function implies that the model places similar probability on similar vowels. Then if most vowel inventories contain some version of /i/, then many of them will contain several closely related variants of /i/ (independently chosen). By contrast, the other methods in this paper do extend nicely to fine-grained phonetic modeling.

be large if the i^{th} and j^{th} vowel often co-occur. Recall that under the principle of dispersion, the vowels that often co-occur are easily distinguishable. Thus, confusable vowel pairs should tend to have potential $\psi(v_i, v_j) < 1$.

Unlike the BPP, the MPP can capture both focalization *and* dispersion. In this work, we will consider a fully connected MPP, i.e., there is a potential function for each pair of vowels in $\mathcal{V}$. MPPs closely resemble Ising models (Ising, 1925), but with the difference that Ising models are typically lattice-structured, rather than fully connected.

Inference in MPPs Inference in fully connected MPPs, just as in general Markov Random Fields (MRFs), is intractable (Cooper, 1990) and we must rely on approximation. In this work, we estimate any needed properties of the MPP distribution by (approximately) drawing vowel inventories from it via Gibbs sampling (Geman and Geman, 1984; Robert and Casella, 2005). Gibbs sampling simulates a discrete-time Markov chain whose stationary distribution is the desired MPP distribution. At each time step, for some random $v_i \in \mathcal{V}$, it stochastically decides whether to replace the current inventory V with $\bar{V}$, where $\bar{V}$ is a copy of V with v_i added (if $v_i \notin V$) or removed (if $v_i \in V$). The probability of replacement is $\frac{p(\bar{V})}{p(V)+p(\bar{V})}$.

2.2.3 Determinantal Point Processes

A determinantal point process (DPP; Macchi, 1975) provides an elegant alternative to an MPP, and one that is directly suited to modeling both focalization and dispersion. Inference requires only a few matrix computations and runs tractably in $O(|\mathcal{V}|^3)$ time, even though the model may encode a rich set of multi-way interactions.

We focus on the L-ensemble parameterization of the DPP, due to ("Eynard–Mehta theorem, Schur process, and their pfaffian analogs").[3]

This type of DPP defines the probability of an inventory $V \subseteq \mathcal{V}$ as

$$p(V) \propto \det L_V, \tag{2.3}$$

where $L \in \mathbb{R}^{N \times N}$ (for $N = |\mathcal{V}|$) is a symmetric positive semidefinite matrix, and L_V refers to the submatrix of L with only those rows and columns corresponding to those elements in the subset V.

Although MAP inference remains NP-hard in DPPs (just as in MPPs), marginal inference becomes tractable. We may compute the normalizing constant in closed form as follows:

$$\sum_{V \in 2^{\mathcal{V}}} \det L_V = \det (L + I). \tag{2.4}$$

How does a DPP ensure focalization and dispersion? L is positive semidefinite iff it can be written as $E^\top E$ for some matrix $E \in \mathbb{R}^{N \times N}$. It is possible to express $p(V)$ in terms of the column vectors of E, which we call $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_N$:

- For inventories of size 2, $p(\{v_i, v_j\}) \propto (\phi(v_i)\phi(v_j)\sin\theta)^2$, where $\phi(v_i), \phi(v_j)$ represent the *quality* of vowels v_i, v_j (as in the BPP) while $\sin\theta \in [0, 1]$ represents their *dissimilarity*. More precisely, $\phi(v_i), \phi(v_j)$ are the lengths of vectors $\boldsymbol{e}_i, \boldsymbol{e}_j$ while θ is the angle between them. Thus, we should choose the columns of E so that focal vowels get *long* vectors and similar

[3] Most DPPs are L-ensembles (Kulesza and Taskar, 2012).

vowels get vectors *of similar direction*.

- Generalizing beyond inventories of size 2, $p(V)$ is proportional to the square of the volume of the parallelepiped whose sides are given by $\{\boldsymbol{e}_i : v_i \in V\}$. This volume can be regarded as $\prod_{v_i \in V} \phi(v_i)$ times a term that ranges from 1 for an orthogonal set of vowels to 0 for a linearly dependent set of vowels.
- The events $v_i \in V$ and $v_j \in V$ are anti-correlated (when not independent). That is, while both vowels may individually have high probabilities (focalization), having either one in the inventory lowers the probability of the other (dispersion).

2.3 Dataset

At this point it is helpful to introduce the empirical dataset we will model. For each of 223 languages,[4] Becker-Kristal (2010) provides the vowel inventory as a set of IPA symbols, listing the first 5 formants for each vowel (or fewer when not available in the original source). Some corpus statistics are shown in Figures 2.2 and 2.3.[5] For the present paper, we take $\mathcal{V}$ to be the set of all 53 IPA

[4]Becker-Kristal lists some languages multiple times with different measurements. When a language had multiple listings, we selected one randomly for our experiments.

[5]Caveat: The corpus is a curation of information from various phonetics papers into a common electronic format. No standard procedure was followed across all languages: it was up to individual phoneticians to determine the size of each vowel inventory, the choice of IPA symbols to describe it, and the procedure for measuring the formants. Moreover, it is an idealization to provide a single vector of formants for each vowel *type* in the language. In real speech, different *tokens* of the same vowel are pronounced differently, because of coarticulation with the vowel context, allophony, interspeaker variation, and stochastic intraspeaker variation. Even within a token, the formants change during the duration of the vowel. Thus, one might do better to represent a vowel's pronunciation not by a formant vector, but by a conditional probability distribution over its formant trajectories given its context, or by a parameter vector that characterizes such a conditional distribution. This

symbols that appear in the corpus. We treat these IPA labels as meaningful, in that we consider two vowels in different languages to be the *same* vowel in $\mathcal{V}$ if (for example) they are both annotated as [ɔ]. We characterize that vowel by its *average* formant vector across all languages in the corpus that contain the vowel: e.g., $(F_1, F_2, \ldots) = (500, 700, \ldots)$ for [ɔ]. In future work, we plan to relax this idealization (see footnote 2), allowing us to investigate natural questions such as whether [u] is pronounced higher (lower F_1) in languages that also contain [o] (to achieve better dispersion).

2.4 Model Parameterization

The BPP, MPP, and DPP models (§2.2) require us to specify parameters for each vowel in $\mathcal{V}$. In §2.4.1, we will accomplish this by deriving the parameters for each vowel v_i from a possibly high-dimensional embedding of that vowel, $\boldsymbol{e}(v_i) \in \mathbb{R}^r$.

In §2.4.2, $\boldsymbol{e}(v_i) \in \mathbb{R}^r$ will in turn be defined as some learned function of $\boldsymbol{f}(v_i) \in \mathbb{R}^k$, where $\boldsymbol{f} : \mathcal{V} \mapsto \mathbb{R}^k$ is the function that maps a vowel to a k-vector of its measurable acoustic properties. This approach allows us to determine reasonable parameters even for rare vowels, based on their measurable properties. It will even enable us in future to generalize to vowels that were unseen in the training set, letting us scale to very large or infinite $\mathcal{V}$ (footnote 2).

setting would require richer data than we present here.

2.4.1 Deep Point Processes

We consider deep versions of all three processes.

Deep Bernoulli Point Process We define

$$\phi(v_i) = ||\boldsymbol{e}(v_i)|| \geq 0 \tag{2.5}$$

Deep Markov Point Process The MPP employs the same unary potential as the BPP, as well as the binary potential

$$\psi(v_i, v_j) = \exp - \frac{1}{T \cdot ||\boldsymbol{e}(v_i) - \boldsymbol{e}(v_j)||^2} < 1 \tag{2.6}$$

where the learned temperature $T > 0$ controls the relative strength of the unary and binary potentials.

This formula is inspired by Coulomb's law for describing the repulsion of static electrically charged particles. Just as the repulsive force between two particles approaches ∞ as they approach each other, the probability of finding two vowels in the same inventory approaches $\exp -\infty = 0$ as they approach each other. The formula is also reminiscent of (Shepard, 1987)'s "universal law of generalization," which says here that the probability of responding to v_i as if it were v_j should fall off exponentially with their distance in some "psychological space" (here, embedding space).

A Deep Determinantal Point Process For the DPP, we simply define the vector $\boldsymbol{e}_i$ to be $\boldsymbol{e}(v_i)$, and proceed as before.

Summary In the deep BPP, the probability of a set of vowels is proportional to the product of the lengths of their embedding vectors. The deep MPP modifies this by multiplying in *pairwise* repulsion terms in $(0, 1)$ that increase as the vectors' endpoints move apart in Euclidean space (or as $T \to \infty$). The deep DPP instead modifies it by multiplying in a single *setwise* repulsion term in $(0, 1)$ that increases as the embedding vectors become more mutually orthogonal. In the limit, then, the MPP and DPP both approach the BPP.

2.4.2 Embeddings

Throughout this work, we simply have $\boldsymbol{f}$ extract the first $k = 2$ formants, since our dataset does not provide higher formants for all languages.[6] For example, we have $\boldsymbol{f}([\text{ɔ}]) = (500, 700)$. We now describe three possible methods for mapping $\boldsymbol{f}(v_i)$ to an embedding $\boldsymbol{e}(v_i)$. Each of these maps has learnable parameters.

Neural Embedding We first consider directly embedding each vowel v_i into a vector space $\mathbb{R}^r$. We achieve this through a feed-forward neural net

$$\boldsymbol{e}(v_i) = W_1 \tanh\left(W_0 \boldsymbol{f}(v_i) + \boldsymbol{b}_0\right) + \boldsymbol{b}_1, \tag{2.7}$$

Eq. (2.7) gives an architecture with 1 layer of nonlinearity; in general we consider stacking $d \geq 0$ layers. Here $W_0 \in \mathbb{R}^{r \times k}, W_1 \in \mathbb{R}^{r \times r}, \ldots W_d \in \mathbb{R}^{r \times r}$ are weight matrices, $\boldsymbol{b}_0, \ldots \boldsymbol{b}_d \in \mathbb{R}^r$ are bias vectors, and tanh could be replaced by any pointwise nonlinearity. We treat both the depth d and the embedding

[6] In lieu of higher formants, we could have extended the vector $\boldsymbol{f}(v_i)$ to encode the binary distinctive features of the IPA vowel v_i: round, tense, long, nasal, creaky, etc.

size r as hyperparameters, and select the optimal values on a development set.

Interpretable Neural Embedding We are interested in the special case of neural embeddings when $r = k$ since then (for any d) the mapping $\boldsymbol{f}(v_i) \mapsto \boldsymbol{e}(v_i)$ is a diffeomorphism:[7] a smooth invertible function of $\mathbb{R}^k$. An example of such a diffeomorphism is shown in Figure 2.1.

There is a long history in cognitive psychology of mapping stimuli into some psychological space. The distances in this psychological space may be predictive of generalization (Shepard, 1987) or of perception. Due to the anatomy of the ear, the mapping of vowels from acoustic space to perceptual space is often presumed to be nonlinear (Rosner and Pickering, 1994; Nearey and Kiefte, 2003), and there are many perceptually-oriented phonetic scales, e.g., Bark and mel, that carry out such nonlinear transformations while preserving the dimensionality k, as we do here. As discussed in §1.3.1, vowel system typology is similarly believed to be influenced by distances between the vowels in a latent metric space. We are interested in whether a constrained k-dimensional model of these distances can do well in our experiments.

Prototype-Based Embedding Unfortunately, our interpretable neural embedding is unfortunately incompatible with the DPP. The DPP assigns probability 0 to any vowel inventory V whose $\boldsymbol{e}$ vectors are linearly dependent. If the vectors are in $\mathbb{R}^k$, then this means that $p(V) = 0$ whenever $|V| > k$. In our

[7] Provided that our nonlinearity in (2.7) is a differentiable invertible function like tanh rather than relu.

setting, this would limit vowel inventories to size 2.

Our solution to this problem is to still construct our interpretable metric space $\mathbb{R}^k$, but then map that nonlinearly to $\mathbb{R}^r$ for some large r. This latter map is constrained. Specifically, we choose "prototype" points $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r \in \mathbb{R}^k$. These prototype points are parameters of the model: their coordinates are learned and do not necessarily correspond to any actual vowel. We then construct $\boldsymbol{e}(v_i) \in \mathbb{R}^r$ as a "response vector" of similarities of our vowel v_i to these prototypes. Crucially, the responses depend on distances measured in the interpretable metric space $\mathbb{R}^k$. We use a Gaussian-density response function, where $\boldsymbol{x}(v_i)$ denotes the representation of our vowel v_i in the interpretable space:

$$\begin{aligned} \boldsymbol{e}(v_i)_\ell &= w_\ell \; p(\boldsymbol{x}(v_i); \boldsymbol{\mu}_\ell, \sigma^2 I) \\ &= w_\ell \; (2\pi\sigma^2)^{-\left(\frac{k}{2}\right)} \exp\left(\frac{-||\boldsymbol{x} - \boldsymbol{\mu}_\ell||^2}{2\sigma^2}\right). \end{aligned} \tag{2.8}$$

for $\ell = 1, 2, \ldots, r$. We additionally impose the constraints that each $w_\ell \geq 0$ and $\sum_{\ell=1}^{r} w_\ell = 1$.

Notice that the sum $\sum_{\ell=1}^{r} \boldsymbol{e}(v_i)$ may be viewed as the density at $\boldsymbol{x}(v_i)$ under a Gaussian mixture model. We use this fact to construct a prototype-based MPP as well: we redefine $\phi(v_i)$ to equal this positive density, while still defining ψ via Eq. (2.6). The idea is that dispersion is measured in the interpretable space $\mathbb{R}^k$, and focalization is defined by certain "good" regions in that space that are centered at the r prototypes.

2.5 Evaluation Metrics

Fundamentally, we are interested in whether our model has abstracted the core principles of what makes a *good* vowel system. Our choice of a probabilistic model provides a natural test: how surprised is our model by held-out languages? In other words, how likely does our model think unobserved, but attested vowel systems are? While this is a natural evaluation paradigm in NLP, it has not—to the best of our knowledge—been applied to a quantitative investigation of linguistic typology.

As a second evaluation, we introduce a *vowel system cloze task* that could also be used to evaluate non-probabilistic models. This task is defined by analogy to the traditional semantic cloze task (Taylor, 1953), where the reader is asked to fill in a missing word in the sentence from the context. In our vowel system cloze task, we present a learner with a subset of the vowels in a held-out vowel system and ask them to predict the remaining vowels. Consider, as a concrete example, the general American English vowel system (excluding long vowels) {[i], [ɪ], [u], [ʊ], [ɛ], [æ], [ɔ], [ɑ], [ə]}. One potential cloze task would be to predict {[i], [u]} given {[ɪ], [ʊ], [ɛ], [æ], [ɔ], [ɑ], [ə]} and the fact that two vowels are missing from the inventory. Within the cloze task, we report accuracy, i.e., did we guess the missing vowel right? We consider three versions of the cloze tasks. First, we predict *one* missing vowel in a setting where exactly one vowel was deleted. Second, we predict *up to one* missing vowel where a vowel *may* have been deleted. Third, we predict *up to two* missing vowels, where one or two vowels may be deleted.

2.6 Experiments

We evaluate our models using 10-fold cross-validation over the 223 languages. We report the mean performance over the 10 folds. The performance on each fold ("test") was obtained by training many models on 8 of the other 9 folds ("train"), selecting the model that obtained the best task-specific performance on the remaining fold ("development"), and assessing it on the test fold. Minimization of the parameters is performed with the L-BFGS algorithm (Liu and Nocedal, 1989). As a preprocessing step, the first two formants values F_1 and F_2 are centered around zero and scaled down by a factor of 1000 since the formant values themselves may be quite large.

Specifically, we use the development fold to select among the following combinations of hyperparameters. For neural embeddings, we tried $r \in \{2, 10, 50, 100, 150, 200\}$. For prototype embeddings, we took the number of components $r \in \{20, 30, 40, 50\}$. We tried network depths $d \in \{0, 1, 2, 3\}$. We sweep the coefficient for an L_2 regularizer on the neural network parameters.

2.6.1 Results and Discussion

Figure 2.1 visualizes the diffeomorphism from formant space to metric space for one of our DPP models (depth $d = 3$ with $r = 20$ prototypes). Similar figures can be generated for all of the interpretable models.

We report results for cross-entropy and the cloze evaluation in Table 2.1.[8] Under both metrics, we see that the DPP is slightly better than the MPP; both

[8] Computing cross-entropy exactly is intractable with the MPP, so we resort to an unbiased importance sampling scheme where we draw samples from the BPP and reweight according to the MPP (Liu et al., 2015).

	BPP	uBPP	uMPP	uDPP	iBPP	iMPP	iDPP	pBPP	pMPP	pDPP
x-ent	8.24	8.28	8.08	8.00	13.01	11.50	✗	12.83	10.95	10.29
cloze-1	69.55%	69.55%	72.05%	73.18%	64.13%	67.02%	✗	65.13%	68.18%	68.18%
cloze-01	60.00%	60.00%	61.01%	62.27%	61.78%	61.04%	✗	61.02%	63.04%	63.63%
cloze-012	53.18%	53.18%	57.92%	58.18%	39.04%	43.02%	✗	40.56%	45.01%	45.46%

Table 2.1: Cross-entropy in nats (lower is better) and cloze prediction accuracy (higher is better). "BPP" is a simple BPP with one parameter for each of the 53 vowels in $\mathcal{V}$. This model does artificially well by modeling an "accidental" feature of our data: it is able to learn not only which vowels are popular among languages, but also which IPA symbols are popular or conventional among the descriptive phoneticians who created our dataset (see footnote 5), something that would become irrelevant if we upgraded our task to predict actual formant vectors rather than IPA symbols (see footnote 2). Our point processes, by contrast, are appropriately allowed to consider a vowel only through its formant vector. The "*u*-" versions of the models use the uninterpretable neural embedding of the formant vector into $\mathbb{R}^r$: by taking r to be large, they are still able to learn special treatment for each vowel in $\mathcal{V}$ (which is why uBPP performs identically to BPP, before being beaten by uMPP and uDPP). The "*i*-" versions limit themselves to an interpretable neural embedding into $\mathbb{R}^k$, giving a more realistic description that does not perform as well. The "*p*-"versions lift that $\mathbb{R}^k$ embedding into $\mathbb{R}^r$ by measuring similarities to r prototypes; they thereby improve on the corresponding *i*- versions. For each result shown, the depth d of our neural network was tuned on a development set (typically $d = 2$). r was also tuned when applicable (typically $r > 100$ dimensions for the *u*- models and $r \approx 30$ prototypes for the *p*- models).

are better than the BPP. This ranking holds for each of the three embedding schemes. The embedding schemes themselves are compared in the caption.

Within each embedding scheme, the BPP performs several points worse on the cloze tasks, confirming that dispersion is needed to model vowel inventories well. Still, the BPP's respectable performance shows that much of the structure can be capture by focalization. As §2.2 noted, the BPP may generate well-dispersed sets, as the common vowels tend to be dispersed already (see Figure 2.2). In this capacity, however, the BPP is not explanatory as it cannot actually tell us why *these* vowels should be frequent.

We mention that depth in the neural network is helpful, with deeper

	BPP			MPP			DPP		
		changes from $n-1$			changes from $n-1$			changes from $n-1$	
n	MAP inventory	additions	deletions	MAP inventory	additions	deletions	MAP inventory	additions	deletions
1	i	i		ə	ə		ə	ə	
2	i, u	u		i, u	i, u	ə	i, u	i, u	ə
3	i, u, a	a		i, u, a	a		i, u, a	a	
4	i, u, a, o	o		i, u, a, e	e		i, u, a, o	o	
5	i, u, a, o, e	e		i, u, a, e, ə	ə		i, u, a, o, ə	o	

Table 2.2: Highest-probability inventory of each size according to our three models (prototype-based embeddings and $d = 3$). The MAP configuration is computed by brute-force enumeration for small n.

embedding networks performing slightly better than depth $d = 0$.

Finally, we identified each model's favorite complete vowel system of size n (Table 2.2). For the BPP, this is simply the n most probable vowels. Decoding the DPP and MPP is NP-hard, but we found the best system by brute force (for small n). The dispersion in these models predicts different systems than the BPP.

2.7 Discussion: Probabilistic Typology

Typology as Density Estimation? Our goal is to define a *universal* distribution over all possible vowel inventories. Is this appropriate? We regard this as a natural approach to typology because it directly describes which kinds of linguistic systems are more or less common. Traditional implicational universals ("all languages with v_i have v_j") are softened, in our approach, into conditional probabilities such as "$p(v_j \in V \mid v_i \in V) \approx 0.9$." Here the 0.9 is not merely an empirical ratio, but a smoothed probability derived from the complete estimated distribution. It is meant to make predictions about unseen languages.

Whether human language learners exploit any properties of this distribution[9] is a separate question that goes beyond typology. Jakobson (1941) did find that children acquired phoneme inventories in an order that reflected principles similar to dispersion ("maximum contrast") and focalization.

At any rate, we estimate the distribution given some set of attested systems that are assumed to have been drawn IID from it. One might object that this IID assumption ignores evolutionary relationships among the attested systems, causing our estimated distribution to favor systems that are coincidentally frequent among current human languages, rather than being natural in some timeless sense. We reply that our approach is then appropriate when the goal of typology is to estimate the distribution of *actual* human languages—a distribution that can be utilized in principle (and also in practice, as we show) to predict properties of *actual* languages from outside the training set.

A different possible goal of typology is a theory of *natural* human languages. This goal would require a more complex approach. One should not imagine that natural languages are drawn in a vacuum from some single, stationary distribution. Rather, each language is drawn *conditionally* on its parent language. Thus, one should estimate a stochastic model of the evolution of linguistic systems through time, and identify "naturalness" with the directions in which this system tends to evolve.

[9] This could happen because learners have evolved to expect the languages (the Baldwin effect), or because the languages have evolved to be easily learned (universal grammar).

Energy Minimization Approaches The traditional energy-based approach (Liljencrants and Lindblom, 1972) to vowel simulation minimizes the following objective (written in our notation):

$$\mathcal{E}(m) = \sum_{1 \le i < j \le m} \frac{1}{||\boldsymbol{e}(v_i) - \boldsymbol{e}(v_j)||^2}, \tag{2.9}$$

where the vectors $\boldsymbol{e}(v_i) \in \mathbb{R}^r$ are not spit out of a deep network, as in our case, but rather directly optimized. Liljencrants and Lindblom (1972) propose a coordinate descent algorithm to optimize $\mathcal{E}(m)$. While this is not in itself a probabilistic model, they generate diverse vowel systems through random restarts that find different local optima (a kind of *deterministic* evolutionary mechanism). We note that Eq. (2.9) assumes that the number of vowels m is given, and only encodes a notion of dispersion. Roark (2001) subsequently extended Eq. (2.9) to include the notion of focalization.

Vowel Inventory Size A fatal flaw of the traditional energy minimization paradigm is that it has no clear way to compare vowel inventories of different sizes. The problem is quite crippling since, in general, inventories with fewer vowels will have lower energy. This does not match reality—the empirical distribution over inventory sizes (shown in Figure 2.3) shows that the mode is actually 5 and small inventories are uncommon: no 1-vowel inventory is attested and only one 2-vowel inventory is known. A probabilistic model over *all* vowel systems must implicitly model the size of the system. Indeed, our models pit all potential inventories against each other, bestowing the extra burden to match the empirical distribution over size.

Frequency of Inventories Another problem is the inability to model frequency. While for inventories of a modest size (3–5 vowels) there are very few unique attested systems, there is a plethora of attested larger vowel systems. The energy minimization paradigm has no principled manner to tell the scientist how likely a novel system may be. Appealing again to the empirical distribution over attested vowel systems, we consider the relative diversity of systems of each size. We graph this in Figure 2.3. Consider all vowel systems of size 7. There are $\binom{|\mathcal{V}|}{7}$ potential inventories, yet the empirical distribution is remarkably peaked. Our probabilistic models have the advantage in this context as well, as they naturally quantify the likelihood of an individual inventory.

Typology is a Small-Data Problem In contrast to many common problems in applied NLP, e.g., part-of-speech tagging, parsing and machine translation, the modeling of linguistic typology is fundamentally a "small-data" problem. Out of the 7105 languages on earth, we only have linguistic annotation for 2600 of them (Comrie et al., 2013). Moreover, we only have *phonetic and phonological annotation* for a much smaller set of languages—between 300-500 (Maddieson, 2013). Given the paucity of data, overfitting on only those attested languages is a dangerous possibility—just because a certain inventory has never been attested, it is probably wrong to conclude that it is impossible—or even improbable—on that basis alone. By analogy to language modeling, almost all sentences observed in practice are novel with respect to the training data, but we still must employ a principled manner to discriminate high-probability sentences (which are syntactically and semantically coherent) from

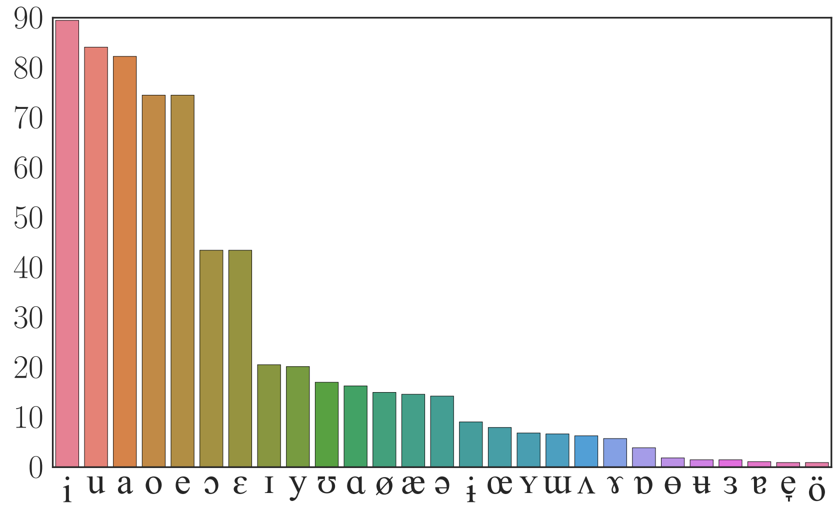

Figure 2.2: Percentage of the vowel inventories (y-axis) in the Becker-Kristal corpus (Becker-Kristal, 2010) that have a given vowel (shown in IPA along the x-axis).

low-probability ones. Probabilistic modeling provides a natural paradigm for this sort of investigation—machine learning has developed well-understood smoothing techniques, e.g., regularization with tuning on a held-out dev set, to avoid overfitting in a small-data scenario.

Related Work in NLP Various point processes have been previously applied to a potpourri of tasks in NLP. Determinantal point processes have found a home in the literature in tasks that require diversity. E.g., DPPs have achieved state-of-the-art results on multi-document document summarization (Kulesza and Taskar, 2011), news article selection (Affandi, Kulesza, and Fox, 2012) recommender systems (Gartrell, Paquet, and Koenigstein, 2016), joint clustering of verbal lexical semantic properties (Reichart and Korhonen, 2013), *inter alia*. Poisson point processes have also been applied to NLP problems, e.g., Yee, Keane, and Zhou (2015) model the emerging topic on social media. We are unaware of previous attempts to probabilistically model vowel inventory typology.

Future Work This work lends itself to several technical extensions. One could expand the function f to more completely characterize each vowel's acoustic properties, perceptual properties, or distinctive features (footnote 6). One could generalize our point process models to sample finite subsets from the *continuous* space of vowels (footnote 2). One could consider augmenting the MPP with a new factor that explicitly controls the size of the vowel inventory. Richer families of point processes might also be worth exploring. For example, perhaps the vowel inventory is generated by some temporal mechanism with latent intermediate steps, such as sequential selection of the vowels or evolutionary drift of the inventory. Another possibility is that vowel systems tend to reuse distinctive features or even follow factorial designs, so that an inventory with creaky front vowels also tends to have creaky back vowels.

2.8 Conclusions

We have presented a series of point process models for the modeling of vowel system inventory typology with the goal of a mathematical grounding for research in phonological typology. All models were additionally given a deep parameterization to learn representations similar to perceptual space in cognitive science. Also, we motivated our preference for *probabilistic* modeling in

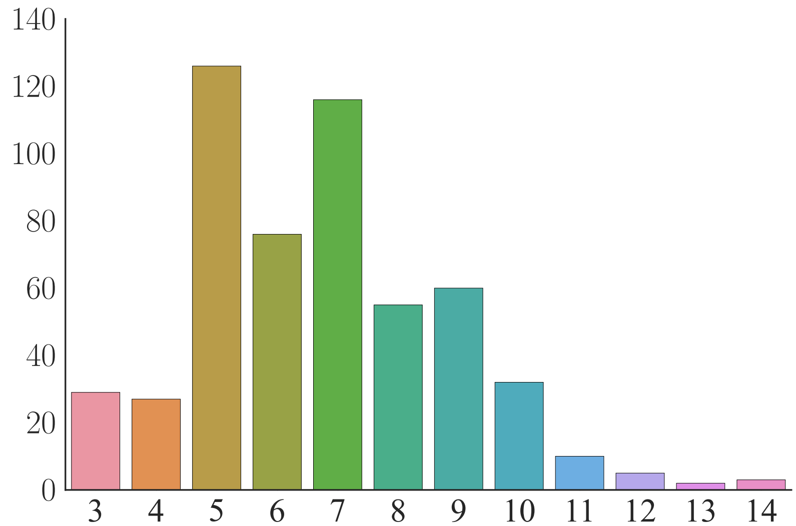

Figure 2.3: Histogram of the sizes of different vowel inventories in the corpus. The x-axis is the size of the vowel inventory and the y-axis is the number of inventories with that size.

linguistic typology over previously proposed computational approaches and argued it is a more natural research paradigm. Additionally, we have introduced several novel evaluation metrics for research in vowel-system typology, which we hope will spark further interest in the area. Their performance was empirically validated on the Becker-Kristal corpus, which includes data from over 200 languages.

Chapter 3

Modeling Sets of Formants

3.1 Introduction

In this chapter, we take a step beyond the work in Chapter 2. Rather than simply giving a distribution over subsets of $\mathcal{V}$, we will model a set of phonetic quantities, e.g. $\bar{\mathcal{V}}$. Specifically, we propose a Bayesian generative model of vowel inventories, where each language's inventory is a finite subset of acoustic vowels represented as points $(F_1, F_2) \in \mathbb{R}^2$. We deploy tools from the neural-network and point-process literatures and experiment on a dataset with 233 distinct languages. We show that our most complicated model outperforms simpler models. This model represents a development over the previous model in that we have moved closer to the phonetics of vowel inventories.

3.2 Notation

We will denote the universal set of international phonetic alphabet (IPA) symbols as $\mathcal{V}$. The observed vowel inventory for language ℓ has size n^ℓ and is denoted $V^\ell = \{(v_1^\ell, \mathbf{v}_1^\ell), \ldots, (v_{n^\ell}^\ell, \mathbf{v}_{n^\ell}^\ell)\} \subseteq \mathcal{V} \times \mathbb{R}^d$, where for each $k \in [1, n^\ell]$, $v_k^\ell \in \mathcal{V}$ is an IPA symbol assigned by a linguist and $\mathbf{v}_k^\ell \in \mathbb{R}^d$ is a vector of d measurable phonetic quantities. In short, the IPA symbol v_k^ℓ was assigned as a label for a phoneme with pronunciation $\mathbf{v}_k^\ell$. The ordering of the elements within V^ℓ is arbitrary.

Goals This framework recognizes that the same IPA symbol v (such as /u/) may represent a slightly different sound $\mathbf{v}$ in one language than in another, although they are transcribed identically. We are specifically interested in how the vowels in a language influence one another's fine-grained pronunciation in $\mathbb{R}^d$. In general, there is no reason to suspect that speakers of two languages, whose phonological systems contain the same IPA symbol, should produce that vowel with identical formants.

Data For the remainder of the paper, we will take $d = 2$ so that each $\mathbf{v} = (F_1, F_2) \in \mathbb{R}^2$, the vector consisting of the first two formant values, as compiled from the field literature by Becker-Kristal (2006). This dataset provides inventories V^ℓ in the form above. Thus, we do not consider further variation of the vowel pronunciation that may occur within the language (between speakers, between tokens of the vowel, or between earlier and later intervals within a token).

3.3 Phonemes versus Phones

Previous work ("Probabilistic Typology: Deep Generative Models of Vowel Inventories") has placed a distribution over discrete phonemes, ignoring the variation across languages in the *pronunciation* of each phoneme. In this paper, we crack open the phoneme abstraction, moving to a learned set of finer-grained phones.

Cotterell and Eisner ("Probabilistic Typology: Deep Generative Models of Vowel Inventories") proposed (among other options) using a *determinantal point process* (DPP) over a universal inventory $\mathcal{V}$ of 53 symbolic (IPA) vowels. A draw from such a DPP is a language-specific inventory of vowel *phonemes*, $V \subseteq \mathcal{V}$. In this paper, we say that a language instead draws its inventory from a larger set $\bar{\mathcal{V}}$, again using a DPP. In both cases, the reason to use a DPP is that it prefers relatively diverse inventories whose individual elements are relatively quantal.

While we could in principle identify $\bar{\mathcal{V}}$ with $\mathbb{R}^d$, for convenience we still take it to be a (large) discrete finite set $\bar{\mathcal{V}} = \{\bar{v}_1, \ldots, \bar{v}_N\}$, whose elements we call *phones*. $\bar{\mathcal{V}}$ is a learned cross-linguistic parameter of our model; thus, its elements—the "universal phones"—may or may not correspond to phonetic categories traditionally used by linguists.

We presume that language ℓ draws from the DPP a subset $\bar{V}^\ell \subseteq \bar{\mathcal{V}}$, whose size we call n^ℓ. For each universal phone $\bar{v}_i$ that appears in this inventory $\bar{V}^\ell$, the language then draws an observable language-specific pronunciation $\mathbf{v}_i^\ell \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$ from a distribution associated cross-linguistically with the universal phone $\bar{v}_i$. We now have an inventory of pronunciations.

As a final step in generating the vowel inventory, we could model IPA labels. For each $\bar{v}_i \in \bar{V}^\ell$, a field linguist presumably draws the IPA label v_i^ℓ conditioned on all the pronunciations $\{\mathbf{v}_i^\ell \in \mathbb{R}^d : \bar{v}_i \in \bar{V}^\ell\}$ in the inventory (and perhaps also on their underlying phones $\bar{v}_i \in \bar{V}^\ell$). This labeling process may be complex. While each pronunciation in $\mathbb{R}^d$ (or each underlying phone in $\bar{\mathcal{V}}$) may have a preference for certain IPA labels in $\mathcal{V}$, the n^ℓ labels must be drawn jointly because the linguist will take care not to use the same label for two phones, and also because the linguist may like to describe the inventory using a small number of distinct IPA features, which will tend to favor factorial grids of symbols. The linguist's use of IPA features may also be informed by phonological and phonetic processes in the language. We leave modeling of this step to future work; so our current likelihood term ignores the evidence contributed by the IPA labels in the dataset, considering only the pronunciations in $\mathbb{R}^d$.

The overall idea is that human languages ℓ draw their inventories from some universal prior, which we are attempting to reconstruct. A caveat is that we will train our method by maximum likelihood, which does not quantify our uncertainty about the reconstructed parameters. An additional caveat is that some languages in our dataset are related to one another, which belies the idea that they were drawn independently. Ideally, one ought to capture these relationships using hierarchical or evolutionary modeling techniques.

3.4 Determinantal Point Processes

Before delving into our generative model, we briefly review technical background used in Chapter 2. A DPP is a probability distribution over the subsets of a *fixed ground set* of size N—in our case, the set of phones $\bar{\mathcal{V}}$. The DPP is usually given as an L-ensemble ("Eynard–Mehta theorem, Schur process, and their pfaffian analogs"), meaning that it is parameterized by a positive semi-definite matrix $L \in \mathbb{R}^{N \times N}$. Given a discrete base set $\bar{\mathcal{V}}$ of phones, the probability of a subset $\bar{V} \subseteq \bar{\mathcal{V}}$ is given by

$$p(\bar{V}) \propto \det\left(L_{\bar{V}}\right) \tag{3.1}$$

where $L_{\bar{V}}$ is the submatrix of L corresponding to the rows and columns associated with the subset $\bar{V} \subseteq \bar{\mathcal{V}}$. The entry L_{ij}, where $i \neq j$, has the effect of describing the similarity between the elements $\bar{v}_i$ and $\bar{v}_j$ (both in $\bar{\mathcal{V}}$)—an ingredient needed to model dispersion. And, the entry L_{ii} describes the quality—focalization—of the vowel $\bar{v}_i$, i.e., how much the model wants to have $\bar{v}_i$ in a sampled set independent of the other members.

3.4.1 Probability Kernel

In this work, each phone $\bar{v}_i \in \bar{\mathcal{V}}$ is associated with a probability density over the space of possible pronunciations $\mathbb{R}^2$. Our measure of phone similarity will consider the "overlap" between the densities associated with two phones. This works as follows: Given two densities $f(x, y)$ and $f'(x, y)$ over $\mathbb{R}^2$, we

$$\prod_{\ell=1}^{M} \left[p(\mathbf{v}^{\ell,1}, \ldots, \mathbf{v}^{\ell,n^\ell} \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, N) \right] p(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_N \mid N)\, p(N) \tag{3.2}$$

$$= \prod_{\ell=1}^{M} \left[\sum_{\mathbf{a}^\ell \in A(n^\ell, N)} \left(\prod_{k=1}^{n^\ell} \underbrace{p(\mathbf{v}^{\ell,k} \mid \boldsymbol{\mu}_{a_k^\ell})}_{4} \right) \underbrace{p(\bar{V}(\mathbf{a}^\ell) \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N, N)}_{3} \right] \underbrace{p(\boldsymbol{\mu}_1, \ldots \boldsymbol{\mu}_N \mid N)}_{2} \underbrace{p(N)}_{1}$$

Figure 3.1: Joint likelihood of M vowel systems under our deep generative probability model for continuous-space vowel inventories. Here language ℓ has an observed inventory of pronunciations $\{\mathbf{v}^{\ell,k} : 1 \leq k \leq n^\ell\}$, and $a_k^\ell \in [1, N]$ denotes a phone that might be responsible for the pronunciation $\mathbf{v}^{\ell,k}$. Thus, $\mathbf{a}^\ell$ denotes some way to jointly label all n^ℓ pronunciations with distinct phones. We must sum over all $\binom{N}{n^\ell}$ such labelings $\mathbf{a}^\ell \in A(n^\ell, N)$ since the true labeling is not observed. In other words, we sum over all ways $\mathbf{a}^\ell$ of completing the data for language ℓ. Within each summand, the product of factors 3 and 4 is the probability of the completed data, i.e., the joint probability of generating the inventory $\bar{V}(\mathbf{a}^\ell)$ of phones used in the labeling and their associated pronunciations. Factor 3 considers the prior probability of $\bar{V}(\mathbf{a}^\ell)$ under the DPP, and factor 4 is a likelihood term that considers the probability of the associated pronunciations.

define the kernel (Jebara, Kondor, and Howard, 2004) as

$$\mathcal{K}(f, f'; \rho) = \int_x \int_y f(x, y)^\rho f'(x, y)^\rho \mathrm{d}x\, \mathrm{d}y \tag{3.3}$$

with inverse temperature parameter ρ.

In our setting, f, f' will both be Gaussian distributions with means $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ that share a fixed spherical covariance matrix $\sigma^2 I$. Then Eq. (3.3) and indeed its generalization to any $\mathbb{R}^d$ has a closed-form solution (Jebara, Kondor, and Howard, 2004, §3.1):

$$\mathcal{K}(f,f';\rho) = \tag{3.4}$$

$$(2\rho)^{\frac{d}{2}} \left(2\pi\sigma^2\right)^{\frac{(1-2\rho)d}{2}} \exp\left(-\frac{\rho||\boldsymbol{\mu} - \boldsymbol{\mu}'||^2}{4\sigma^2}\right)$$

Notice that making ρ small (i.e., high temperature) has an effect on (3.4) similar to scaling the variance σ^2 by the temperature, but it also results in changing the scale of $\mathcal{K}$, which affects the balance between dispersion and focalization in (3.6) below.

3.4.2 Focalization Score

The probability kernel given in Eq. (3.3) naturally handles the linguistic notion of dispersion. What about focalization? We say that a phone is focal to the extent that it has a high score

$$F(\boldsymbol{\mu}) = \exp\left(U_2 \tanh(U_1\boldsymbol{\mu} + \mathbf{b}_1) + \mathbf{b}_2\right) > 0 \tag{3.5}$$

where $\boldsymbol{\mu}$ is the mean of its density. To learn the parameters of this neural network from data is to learn which phones are focal. We use a neural network since the focal regions of $\mathbb{R}^2$ are distributed in a complex way.

3.4.3 The Gram Matrix

If $f_i = \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I)$ is the density associated with the phone $\bar{v}_i$, we may populate an $N \times N$ real matrix L where

$$L_{ij} = \begin{cases} \mathcal{K}(f_i, f_j; \rho) & \text{if } i \neq j \\ \mathcal{K}(f_i, f_j; \rho) + F(\boldsymbol{\mu}_i) & \text{if } i = j \end{cases} \tag{3.6}$$

Since L is the sum of two positive definite matrices (the first specializes a known kernel and the second is diagonal and positive), it is also positive definite. As a result, it can be used to parameterize a DPP over $\bar{\mathcal{V}}$. Indeed, since L is positive definite and not merely positive semidefinite, it will assign

Algorithm 1 Generative Process

$N \sim \text{Poisson}(\lambda)\ (\in \mathbb{N})$ ①
for $i = 1$ **to** N :
  $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, I)\ (\in \mathbb{R}^2)$ ②
define $L \in \mathbb{R}^{N \times N}$ via (3.6)
for $\ell = 1$ **to** M :
  $\bar{V}^{\ell} \sim \text{DPP}(L)\ (\subseteq [1, N])$; let $n^{\ell} = |\bar{V}^{\ell}|$ ③
  for $i \in \bar{V}^{\ell}$:
    $\tilde{\mathbf{v}}_i^{\ell} \sim \mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$ ④
    $\mathbf{v}_i^{\ell} = \nu_{\boldsymbol{\theta}}\left(\tilde{\mathbf{v}}_i^{\ell}\right)$ ④

positive probability to *any* subset of $\bar{\mathcal{V}}$.

As previously noted, this DPP does not define a distribution over an infinite set, e.g., the powerset of $\mathbb{R}^2$, as does recent work on continuous DPPs (Affandi, Fox, and Taskar, 2013). Rather, it defines a distribution over the powerset of a *set of densities with finite cardinality*. Once we have sampled a subset of densities, a real-valued quantity may be additionally sampled from each sampled density.

3.5 A Deep Generative Model

We are now in a position to expound our generative model of continuous-space vowel typology. We generate a set of formant pairs for M languages in a four step process. Note that throughout this exposition, language-specific quantities with be superscripted with an integral language marker ℓ, whereas universal quantities are left unsuperscripted. The generative process is written in algorithmic form in Algorithm 1. Note that each step is numbered and color-coded for ease of comparison with the full joint likelihood in Figure 3.1.

Step ①: $p(N)$ We sample the size N of the universal phone inventory $\bar{\mathcal{V}}$ from a Poisson distribution with a rate parameter λ, i.e.,

$$N \sim \text{Poisson}\,(\lambda) \tag{3.7}$$

That is, we do not presuppose a certain number of phones in the model.

Step ②: $p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)$ Next, we sample the means $\boldsymbol{\mu}_i$ of the Gaussian phones. In the model presented here, we assume that each phone is generated independently, so $p(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N) = \prod_{i=1}^{N} p(\boldsymbol{\mu}_i)$. Also, we assume a standard Gaussian prior over the means, $\boldsymbol{\mu}_i \sim \mathcal{N}(\mathbf{0}, I)$

The sampled means define our N Gaussian phones $\mathcal{N}\left(\boldsymbol{\mu}_i, \sigma^2 I\right)$: we are assuming for simplicity that all phones share a *single* spherical covariance matrix, defined by the hyperparameter σ^2. The dispersion and focalization of these phones define the matrix L according to Eqs. (3.4)–(3.6), where ρ in (3.4) and the weights of the focalization neural net (3.5) are also hyperparameters.

Step ③: $p(\bar{V}^{\ell} \mid \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N)$ Next, for each language $\ell \in [1, \ldots, M]$, we sample a diverse subset of the N phones, via a single draw from a DPP parameterized by matrix L:

$$\bar{V}^{\ell} \sim \text{DPP}(L) \tag{3.8}$$

where $\bar{V}^{\ell} \subseteq [1, N]$. Thus, $i \in \bar{V}^{\ell}$ means that language ℓ contains phone $\bar{v}_i$. Note that even the size of the inventory, $n^{\ell} = |\bar{V}^{\ell}|$, was chosen by the DPP. In general, we have $n^{\ell} \ll N$.

Step ④: $\prod_{i \in \bar{V}^\ell} p(\mathbf{v}_i^\ell \mid \boldsymbol{\mu}_i)$ The final step in our generative process is that the phones $\bar{v}_i$ in language ℓ must generate the pronunciations $\mathbf{v}_i^\ell \in \mathbb{R}^2$ (formant vectors) that are actually observed in language ℓ. Each vector takes two steps. For each $i \in \bar{V}^\ell$, we generate an underlying $\tilde{\mathbf{v}}_i \in \mathbb{R}^2$ from the corresponding Gaussian phone. Then, we run this vector through a feed-forward neural network $\nu_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$. In short:

$$\tilde{\mathbf{v}}_i^\ell \sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma^2 I) \tag{3.9}$$

$$\mathbf{v}_i^\ell = \nu_{\boldsymbol{\theta}}(\tilde{\mathbf{v}}_i^\ell) \tag{3.10}$$

where the second step is deterministic. We can fuse these two steps into a single step $p(\mathbf{v}_i \mid \boldsymbol{\mu}_i)$, whose closed-form density is given in Eq. (3.12) below. In effect, step 4 takes a Gaussian phone as input and produces the observed formant vector with an *underlying* formant vector in the middle.

This completes our generative process. We do not observe all the steps, but only the final collection of pronunciations $\mathbf{v}_i^\ell$ for each language, where the subscripts i that indicate phone identity have been lost. The probability of this incomplete dataset involves summing over possible phones for each pronunciation, and is presented in Figure 3.1.

3.5.1 A Neural Transformation of a Gaussian

A crucial bit of our model is running a sample from a Gaussian through a neural network. Under certain restrictions, we can find a closed form for the resulting density; we discuss these below. Let $\nu_{\boldsymbol{\theta}}$ be a depth-2 multi-layer perceptron

$$\nu_{\boldsymbol{\theta}}(\tilde{\mathbf{v}}_i) = W_2 \tanh\left(W_1 \tilde{\mathbf{v}}_i + \mathbf{b}_1\right) + \mathbf{b}_2 \tag{3.11}$$

In order to find a closed-form solution, we require that (3.5) be a diffeomorphism, i.e., an invertible mapping from $\mathbb{R}^2 \to \mathbb{R}^2$ where both $\nu_{\boldsymbol{\theta}}$ and its inverse $\nu_{\boldsymbol{\theta}}^{-1}$ are differentiable. This will be true as long as $W_1, W_2 \in \mathbb{R}^{2\times 2}$ are square matrices of full-rank and we choose a smooth, invertible activation function, such as tanh. Under those conditions, we may apply the standard theorem for transforming a random variable (see Stark and Woods, 2011):

$$\begin{aligned} p(\mathbf{v}_i \mid \boldsymbol{\mu}_i) &= p(\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i) \mid \boldsymbol{\mu}_i) \ \det J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i)} \\ &= p(\tilde{\mathbf{v}}_i \mid \boldsymbol{\mu}_i) \ \det J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{v}_i)} \end{aligned} \tag{3.12}$$

where $J_{\nu_{\boldsymbol{\theta}}^{-1}(\mathbf{x})}$ is the Jacobian of the inverse of the neural network at the point $\mathbf{x}$. Recall that $p(\tilde{\mathbf{v}}_i \mid \boldsymbol{\mu}_i)$ is Gaussian-distributed.

3.6 Modeling Assumptions

Imbued in our generative story are a number of assumptions about the linguistic processes behind vowel inventories. We briefly draw connections between our theory and the linguistics literature.

Why underlying phones? A technical assumption of our model is the existence of a universal set of underlying phones. Each phone is equipped with a probability distribution over reported acoustic measurements (pronunciations), to allow for a single phone to account for multiple slightly different pronunciations in different languages (though never in the same language). This distribution can capture both actual interlingual variation and also random noise in the measurement process.

While our universal phones may seem to resemble the universal IPA symbols used in phonological transcription, they lack the rich featural specifications of such phonemes. A phone in our model has no features other than its mean position, which wholly determines its behavior. Our universal phones are not a substantive linguistic hypothesis, but are essentially just a way of partitioning $\mathbb{R}^2$ into finitely many small regions whose similarity and focalization can be precomputed. This technical trick allows us to use a discrete rather than a continuous DPP over the $\mathbb{R}^2$ space.[1]

Why a neural network? Our phones are Gaussians of spherical variance σ^2, presumed to be scattered with variance 1 about a two-dimensional *latent* vowel space. Distances in this latent space are used to compute the dissimilarity of phones for modeling dispersion, and also to describe the phone's ability to vary across languages. That is, two phones that are *distant* in the latent space can appear in the same inventory—presumably they are easy to discriminate in both perception and articulation—and it is easy to choose which one better explains an acoustic measurement, thereby affecting the other measurements that may appear in the inventory.

[1] Indeed, we could have simply taken our universal phone set to be a huge set of tiny, regularly spaced overlapping Gaussians that "covered" (say) the unit circle. As a computational matter, we instead opted to use a smaller set of Gaussians, giving the learner the freedom to infer their positions and tune their variance σ^2. Because of this freedom, this set should not be too large, or a MAP learner may overfit the training data with zero-variance Gaussians and be unable to explain the test languages—similar to overfitting a Gaussian mixture model.

We relate this *latent* space to measurable acoustic space by a learned diffeomorphism $\nu_{\boldsymbol{\theta}}$ ("Probabilistic Typology: Deep Generative Models of Vowel Inventories"). $\nu_{\boldsymbol{\theta}}^{-1}$ can be regarded as warping the acoustic distances into perceptual/articulatory distances. In some "high-resolution" regions of acoustic space, phones with fairly similar (F_1, F_2) values might yet be far apart in the latent space. Conversely, in other regions, relatively large acoustic changes in some direction might not prevent two phones from acting as similar or two pronunciations from being attributed to the same phone. In general, a unit circle of radius σ in latent space may be mapped by $\nu_{\boldsymbol{\theta}}$ to an oddly shaped connected region in acoustic space, and a Gaussian in latent space may be mapped to a multimodal distribution.

3.7 Inference and Learning

We fit our model via MAP-EM (Dempster, Laird, and Rubin, 1977). The E-step involves deciding which phones each language has. To achieve this, we fashion a Gibbs sampler (Geman and Geman, 1984), yielding a Markov-Chain Monte Carlo E-step (Levine and Casella, 2001).

3.7.1 Inference: MCMC E-Step

Inference in our model is intractable even when the phones $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_N$ are fixed. Given a language with n vowels, we have to determine which subset of the N phones best explains those vowels. As discussed above, the alignment $\mathbf{a}$ between the n vowels and n of the N phones represents a latent variable. Marginalizing it out is #P-hard, as we can see that it is equivalent to summing over all bipartite matchings in a weighted graph, which, in turn, is as costly as computing the permanent of a matrix (Valiant, 1979). Our sampler[2] is an approximation algorithm for the task. We are interested in sampling $\mathbf{a}$, the labeling of observed vowels with universal phones. Note that this implicitly samples the language's phone inventory $\bar{V}(\mathbf{a})$, which is fully determined by $\mathbf{a}$.

Specifically, we employ an MCMC method closely related to Gibbs sampling. At each step of the sampler, we update our vowel-phone alignment $\mathbf{a}^\ell$ as follows. Choose a language ℓ and a vowel index $k \in [1, n^\ell]$, and let $i = a_k^\ell$ (that is, pronunciation $\mathbf{v}^{\ell,k}$ is currently labeled with universal phone $\bar{v}_i$). We will consider changing a_k^ℓ to j, where j is drawn from the $(N - n^\ell)$ phones that do *not* appear in $\bar{V}(\mathbf{a}^\ell)$, heuristically choosing j in proportion to the likelihood $p(\mathbf{v}^{\ell,k} \mid \boldsymbol{\mu}_j)$. We then stochastically decide whether to keep $a_k^\ell = i$ or set $a_k^\ell = j$ in proportion to the resulting values of the product 4 $\cdot$ 3 in Eq. (3.2).

For a single E-step, the Gibbs sampler "warm-starts" with the labeling from the end of the previous iteration's E-step. It sweeps $S = 5$ times through all vowels for all languages, and returns S sampled labelings, one from the end of each sweep.

We are also interested in automatically choosing the number of phones N, for which we take the Poisson's rate parameter $\lambda = 100$. To this end, we employ reversible-jump MCMC (Green, 1995), resampling N at the start of every E-step.

[2]Taken from Volkovs and Zemel (2012, p. 3.1).

3.7.2 Learning: M-Step

Given the set of sampled alignments provided by the E-step, our M-step consists of optimizing the log-likelihood of the now-complete training data using the inferred latent variables. We achieved this through SGD training of the diffeomorphism parameters $\boldsymbol{\theta}$, the means $\boldsymbol{\mu}_i$ of the Gaussian phones, and the parameters of the focalization kernel $\mathcal{F}$.

3.8 Experiments

3.8.1 Data

Our data is taken from the Becker-Kristal corpus (Becker-Kristal, 2006), which is a compilation of various phonetic studies and forms the largest multilingual phonetic database. Each entry in the corpus corresponds to a linguist's phonetic description of a language's vowel system: an inventory consisting of IPA symbols where each symbol is associated with two or more formant values. The corpus contains data from 233 distinct languages. When multiple inventories were available for the same language (due to various studies in the literature), we selected one at random and discarded the others.

3.8.2 Baselines

Baseline #1: Removing dispersion The key technical innovation in our work lies in the incorporation of a DPP into a generative model of vowel formants—a continuous-valued quantity. The role of the DPP was to model the linguistic principle of dispersion—we may cripple this portion of our model, e.g., by forcing $\mathcal{K}$ to be a diagonal kernel, i.e., $K_{ij} = 0$ for $i \neq j$. In this case the DPP becomes a Bernoulli Point Process (BPP)—a special case of the DPP. Since dispersion is widely accepted to be an important principle governing naturally occurring vowel systems, we expect a system trained without such knowledge to perform worse.

Baseline #2: Removing the neural network $\nu_{\boldsymbol{\theta}}$ Another question we may ask of our formulation is whether we actually need a fancy neural mapping $\nu_{\boldsymbol{\theta}}$ to model our typological data well. The human perceptual system is known to perform a non-linear transformation on acoustic signals, starting with the non-linear cochlear transform that is physically performed in the ear. While $\nu_{\boldsymbol{\theta}}^{-1}$ is intended as loosely analogous, we determine its benefit by removing Eq. (3.10) from our generative story, i.e., we take the observed formants $\mathbf{v}_k$ to arise directly from the Gaussian phones.

Baseline #3: Supervised phones and alignments A final baseline we consider is *supervised* phones. Linguists standardly employ a finite set of phones—symbols from the IPA. In phonetic annotation, it is common to map each sound in a language back to this universal discrete alphabet. Under such an annotation scheme, it is easy to discern, cross-linguistically, which vowels originate from the same phoneme: an /ɪ/ in German may be roughly equated with an /ɪ/ in English. However, it is not clear how consistent this annotation truly is. There are several reasons to expect high-variance in the cross-linguistic acoustic signal. First, IPA symbols are primarily useful for interlinked phonological distinctions, i.e., one applies the symbol /ɪ/ to distinguish it from /i/

N	metric	DPP+$\nu_{\boldsymbol{\theta}}$	BPP+$\nu_{\boldsymbol{\theta}}$	DPP−$\nu_{\boldsymbol{\theta}}$	Sup.
15	x-ent	540.02	540.05	600.34	✗
	cloze1	5.76	5.76	6.53	✗
	cloze12	4.89	4.89	5.24	✗
25	x-ent	280.47	275.36	335.36	✗
	cloze1	5.04	5.25	6.23	✗
	cloze12	4.76	4.97	5.43	✗
50	x-ent	222.85	231.70	320.05	1610.37
	cloze1	3.38	3.16	4.02	4.96
	cloze12	2.73	2.93	3.04	6.95
57	x-ent	212.14	220.42	380.31	✗
	cloze1	2.21	3.08	3.25	✗
	cloze12	2.01	3.05	3.41	✗
100	x-ent	271.95	301.45	380.02	✗
	cloze1	2.26	2.42	3.03	✗
	cloze12	1.96	2.01	2.51	✗

Table 3.1: Cross-entropy in nats per language (lower is better) and expected Euclidean-distance error of the cloze prediction (lower is better). The overall best value for each task is boldfaced. The case $N = 50$ is compared against our supervised baseline. The $N = 57$ row is the case where we allowed N to fluctuate during inference using reversible-jump MCMC; this was the N value selected at the final EM iteration.

in the given language, rather than to associate it with the sound bearing the same symbol in a second language. Second, field linguists often resort to the closest common IPA symbol, rather than an exact match: if a language makes no distinction between /i/ and /ɪ/, it is more common to denote the sound with a /i/. Thus, IPA may not be as universal as hoped. Our dataset contains 50 IPA symbols so this baseline is only reported for $N = 50$.

3.8.3 Evaluation

Evaluation in our setting is tricky. The scientific goal of our work is to place a bit of linguistic theory on a firm probabilistic footing, rather than a downstream engineering-task, whose performance we could measure. We consider three metrics.

Cross-Entropy Our first evaluation metric is cross-entropy: the average negative log-probability of the vowel systems in held-out test data, given the universal inventory of N phones that we trained through EM. We find this to be the cleanest method for scientific evaluation—it is the metric of optimization and has a clear interpretation: how surprised was the model to see the vowel systems of held-out, but attested, languages?

The cross-entropy is the negative log of the $\prod [\cdots]$ expression in Eq. (3.2), with ℓ now ranging over held-out languages.[3] Wallach et al. (2009) give several methods for estimating the intractable sum in language ℓ. We use the simple harmonic mean estimator, based on 50 samples of $\mathbf{a}^\ell$ drawn with our Gibbs sampler (warm-started from the final E-step of training).

Cloze Evaluation In addition, following Cotterell and Eisner ("Probabilistic Typology: Deep Generative Models of Vowel Inventories"), we evaluate our trained model's ability to perform a cloze task (Taylor, 1953). Given $n^\ell - 1$ or $n^\ell - 2$ of the vowels in held-out language ℓ, can we predict the pronunciations

[3] Since that expression is the product of both probability distributions and probability densities, our "cross-entropy" metric is actually the sum of both entropy terms and (potentially negative) differential entropy terms. Thus, a value of 0 has no special significance.

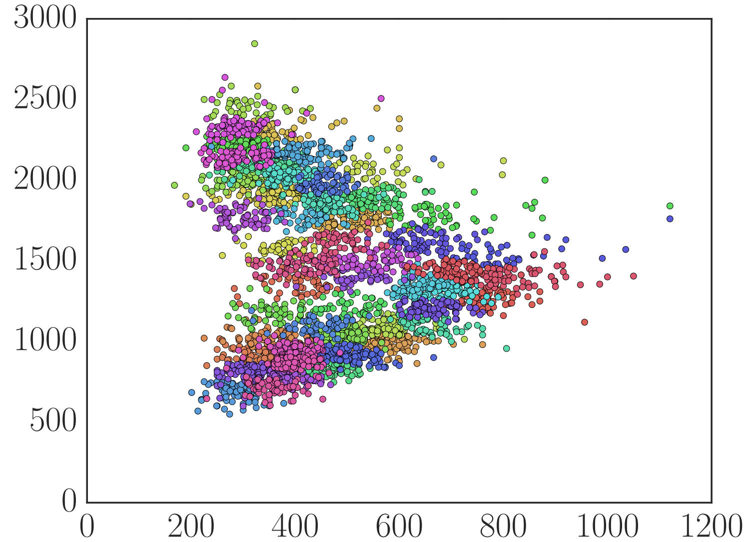

Figure 3.2: A graph of $\mathbf{v} = (F_1, F_2)$ in the union of all the training languages' inventories, color-coded by inferred phone ($N = 50$).

$\mathbf{v}_k$ of the remaining 1 or 2? We predict $\mathbf{v}_k$ to be $\nu_{\boldsymbol{\theta}}(\boldsymbol{\mu}_i)$ where $i = a_k^\ell$ is the phone inferred by the sampler. Note that the sampler's inference here is based only on the observed vowels (the likelihood) and the focalization-dispersion preferences of the DPP (the prior). We report the expected error of such a prediction—where error is quantified by Euclidean distance in (F_1, F_2) formant space—over the same 50 samples of $\mathbf{a}^\ell$.

For instance, consider a previously unseen vowel system with formant values $\{(499, 2199), (861, 1420), (571, 1079)\}$. A "cloze1" evaluation would aim to predict $\{(499, 2199)\}$ as the missing vowel, given $\{(861, 1420), (571, 1079)\}$, and the fact that $n^\ell = 3$. A "cloze12" evaluation would aim to predict two missing vowels.

3.8.4 Experimental Details

Here, we report experimental details and the hyperparameters that we use to achieve the results reported. We consider a neural network $\nu_{\boldsymbol{\theta}}$ with $k \in [1, 4]$ layers and find $k = 1$ the best performer on development data. Recall that our *diffeomorphism* constraint requires that each layer have exactly two hidden units, the same as the number of observed formants. We consider $N \in \{15, 25, 50, 100\}$ phones as well as letting N fluctuate with reversible-jump MCMC (see footnote 1). We train for 100 iterations of EM, taking $S = 5$ samples at each E-step. At each M-step, we run 50 iterations of SGD for the focalization NN and also for the diffeomorphism NN. For each N, we selected (σ^2, ρ) by minimizing cross-entropy on a held-out development set. We considered $(\sigma^2, \rho) \in \{10^k\}_{k=1}^{5} \times \{\rho^k\}_{k=1}^{5}$.

3.8.5 Results and Error Analysis

We report results in Table 3.1. We find that our DPP model improves over the baselines. The results support two claims: (i) dispersion plays an important role in the structure of vowel systems and (ii) learning a non-linear transformation of a Gaussian improves our ability to model sets of formant-pairs. Also, we observe that as we increase the number of phones, the role of the DPP becomes more important. We visualize a sample of the trained alignment in Figure 3.2.

Frequency Encodes Dispersion Why does dispersion not always help? The models with fewer phones do not reap the benefits that the models with more phones do. The reason lies in the fact that the most common vowel formants are *already* dispersed. This indicates that we still have not quite modeled the mechanisms that select for good vowel formants, despite our work at the phonetic level; further research is needed. We would prefer a model that explains the *evolutionary motivation* of sound systems as communication systems.

Number of Induced Phones What is most salient in the number of induced phones is that it is close to the number of IPA phonemes in the data. However, the performance of the phoneme-supervised system is much worse, indicating that, perhaps, while the linguists have the right idea about the *number* of universal symbols, they did not specify the correct IPA symbol in all cases. Our data analysis indicates that this is often due to pragmatic concerns in linguistic field analysis. For example, even if /ɪ/ is the proper IPA symbol for the sound, if there is no other sound in the vicinity the annotator may prefer to use more common /i/.

3.9 Conclusion

Our model combines existing techniques of probabilistic modeling and inference to attempt to fit the actual distribution of the world's vowel systems. We presented a generative probability model of sets of measured (F_1, F_2) pairs. We view this as a necessary step in the development of generative probability models that can explain the distribution of the world's languages. Previous work on generating vowel inventories has focused on how those inventories were transcribed into IPA by field linguists, whereas we focus on the field linguists' acoustic measurements of how the vowels are actually pronounced.

Chapter 4

Modeling Sequences of Formants

4.1 Introduction

Functionalism is the view that languages are shaped by communicative pressures (Bybee, 2003). Functionalist research seeks to explain how the distribution of human languages is affected by their need to function as communication channels. If one accepts the premise that a key evolutionary pressure on language is communication (Pinker and Jackendoff, 2005—though see Fitch, Hauser, and Chomsky, 2005)— then languages that better serve the needs of communication will be more common, other things equal.[1] In phonology, this line of research is called phonetically driven phonology. It considers how naturally occurring phonological systems have developed to communicate via human production and perception of sound (Boersma, 1998; Hayes, 1999;

[1] A common objection to functionalist analyses is that a language is not *optimal* with respect to a given function. For example, while frequent words tend to be short, there are many examples of frequent words that are longer than rare words, so swapping their forms would have increased the rate of communication even further. One explanation for suboptimality is that the language was shaped by other functional pressures as well. Another explanation is that while languages may tend to evolve toward greater functionality, evolution is a slow and stochastic process, during which the language may also be affected by other shocks, so one cannot expect a language to have achieved global optimality at any given point in time.

Bronson, 2004).

The phonological system of a spoken language includes a vowel system—an inventory of vowel types such as /a/, /e/, and /æ/. A given vowel type such as /a/ is not pronounced identically each time it is uttered. If we simplify by ignoring the effect of context on pronunciation, then each vowel type can be regarded as a probability distribution over tokens of that vowel. Then a vowel system is a small collection of probability distributions corresponding to the different vowel types.

Our technical contribution in this paper is to propose a universal generative model of such vowel systems. Our model simply states that more functional vowel systems should be more common. How *much* more common is controlled by a single parameter T. Our model is so simple that it can be abstractly formulated using little more than this one parameter. However, then the vowels live only in an abstract phonetic space. Additional parameters of our model nonlinearly relate this abstract space to concrete phonetic measurements. They serve to describe which specific acoustic distinctions are easy for humans to produce and perceive.

Our model recognizes two functional pressures: rate and ease of communication. These pressures are inherently in competition. Consider a language with K vowel types. If its speakers pronounce these vowels precisely and its hearers attend closely to those pronunciations, then the vowels can be accurately distinguished, so information is communicated at a high rate, approaching $\log_2 K$ bits per token in the limit where distinction is perfect and all vowels are equally probable. Conversely, in a version of the language where speakers or hearers invest less effort, there will be more stochastic noise in the communication channel, and the information content will drop as a result. We operationalize both communicative effort and information rate, and construct a prior that directly favors "efficient" phonetic systems—those with a high rate/effort ratio.

We experiment on data from 29 languages taken from the Epitran (Mortensen, Dalmia, and Littell, 2018) portion of the VoxClamantis dataset (Salesky et al., 2020). By jointly training our model on many languages, we obtain a description of what is easy for humans—and test the model's goodness of fit.

In addition to simply measuring performance on held-out languages, we also confirm that our model does generate vowel systems that are human-like on various metrics. For instance, it produces vowel systems that are more dispersed than a baseline model and tends to produce vowel systems with an appropriate number of vowels when tuned. In contrast to Chapter 2, where we estimate a distribution over the space of all languages and asks how well the model generalizes to *new* languages, here we only ask how well our universal prior helps learning on a *per-language* basis. We find that on our two languages (English and Mandarin Chinese), our prior leads to more successful extrapolation.

4.2 Linguistic Background

The study of the sound structure of language is subdivided into two disciplines: **phonology** and **phonetics**. Phonology deals with discrete units—the **phonemes** of a language—that are strung together to form larger units such as words (Hayes, 2011). Phonemes are abstract categories that are not observed directly: "the major motivation for the phoneme is its role as a functional unit that allows one to make concrete the intuition that sounds that are objectively different are functionally 'the same' at some level of analysis" (Dresher, 2011). Phonetics deals with the objective differences: the production and perception by humans of specific sound waves. Different tokens of the same phoneme can have different phonetic **measurements**. In this paper, we will regard this variation as random variation due to imprecise articulation.[2]

Many phonetic measurements of a token are possible. A linguist could measure the speaker's articulators (Ladefoged and Johnson, 2014), or extract various objective or subjective properties of the recorded sound wave. In particular, a vowel token's **formants** $F_1, F_2, F_3, \ldots$ are the frequencies at which its sound energy is concentrated.

Our work builds on phonetically driven phonology (Hayes, 1999), a prominent functionalist research area that seeks to explain phonological patterns (e.g. distinctive features) through phonetic properties. Newmeyer (2005) poses the key question: "How direct is the linkage between functional pressures and the typological distribution of formal elements that represents a response to those pressures?" In this paper, we posit a very direct link, suggesting that vowel systems have been sampled directly from a functionalist prior. This is an example of **holistic functionalism** (Newmeyer, 2005), the claim that functional pressures affect the structure of a language as a whole. We

[2]In future, by adopting a more complex model, we could predict some of the variation from differences in the tokens' contexts (allophonic variation) and the tokens' speakers (as different speakers have different vocal tracts and accents). The remaining unexplained variation would still be treated as random, like the residuals of linear regression.

will not here investigate the **atomic functionalism** hypothesis that individual utterances are also shaped by functional pressures.[3]

4.3 Measurement Space vs. Phonetic Space

Spoken language is communicated through sound waves. The speaker transmits discrete phonemes by converting them into an acoustic form, from which the listener tries to recover the discrete phonemes.

Our observed data will consist of tokens $\mathbf{m}$ that fall in some universal **measurement space** $\mathcal{M}$, whose shape is determined by the process that collected these observations. $\mathcal{M}$ may have continuous dimensions (acoustic formants, articulator positions) and/or discrete dimensions (IPA features annotated by linguists). In the case of vowel system typology, much work has focused on the first two formants (F_1, F_2) (Liljencrants and Lindblom, 1972; Schwartz et al., 1997; Cotterell and Eisner, 2018), as this is the primary information that human listeners rely on to discriminate among vowel phonemes within a language. Using a new dataset, our experiments will take $\mathcal{M}$ to be the 3-dimensional space of possible (F_1, F_2, F_3) values. We will consider the distribution of (F_1, F_2, F_3) token measurements for each vowel type in each language. However, our methods could also handle an $\mathcal{M}$ that had been augmented with further dimensions such as tone, duration, and creaky voice.

[3]Atomic functionalism might predict that English speakers pronounce the vowel in "bed" more carefully than in "fed," in order to distinguish "bed" from "bid," particularly in sentences where "bid" would have been plausible. One might try to extend our model in order to quantify this kind of effect. However, our present model assumes that the phonetic distribution of a vowel is independent of the word or sentence in which it appears. Relaxing this assumption would force us to confront the question of why the vowels in "bed" and "fed" remain as similar as they do—that is, why it is functional for a language's morphemes to be composed from a fixed inventory of phonemes whose phonetics are largely context-invariant.

Our model says that each token also has an underlying representation $\mathbf{v}$ in an universal abstract **phonetic space**, $\mathbb{R}^d$, which yielded the measurement via $\mathbf{m} = \Psi(\mathbf{v})$. Here $\Psi : \mathbb{R}^d \to \mathcal{M}$ is a universal deterministic mapping that represents the measurement process. More generally, to accommodate noisy or subjective measurements, we could allow Ψ to be a conditional probability distribution, saying that each measurement was sampled independently according to $\Psi(\mathbf{m} \mid \mathbf{v})$. In either case Ψ should be surjective, so that any possible measurement $\mathbf{m} \in \mathcal{M}$ can be explained by at least one underlying $\mathbf{v}$.

Our model defines a **phoneme** to be a probability distribution over abstract phonetic space—indeed, some Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. A **system** is then a probability distribution over a finite **inventory** of phonemes, yielding a mixture model over phonetic space. Our goal is to characterize which systems are probable *a priori*, in terms of their communicative effort and communicative capacity. Thus, we aim to give a universal prior over Gaussian mixture models.

Why do we make use of an underlying phonetic space? It provides a latent embedding of tokens such that cross-linguistic phonetic properties correspond to Euclidean distances. In phonetic space, points $\mathbf{v}$ that are *far from the origin* $\mathbf{0}$ are those that universally require more effort to pronounce or perceive. Pairs of points $\mathbf{v}, \mathbf{v}'$ that are *close together* are those that are universally hard to distinguish: a speaker who *wishes* to produce $\mathbf{v}$ might *actually* produce $\mathbf{v}'$ or be *misperceived* by a hearer as having done so—and thus if $\mathbf{v}$ and $\mathbf{v}'$ are assigned to different phonemes, the system will require high effort from speakers and hearers to maintain a low confusion rate between these phonemes.[4]

There is no guarantee that distances in *measurement* space behave in this way. The density of a given phoneme i in measurement space may be irregularly shaped; but we will assume that this is due to a nonlinear distortion by Ψ of some simple Gaussian density in phonetic space, whose covariance matrix Σ_i describes the underlying directions of typical variation in the phoneme's pronunciation. Two sounds $\mathbf{m}, \mathbf{m}'$ that are close together in measurement space may nonetheless be easily distinguished and thus assigned to different phonemes; we take this as evidence that their underlying forms $\mathbf{v}, \mathbf{v}'$ must be far apart in phonetic space. Similarly, if a measurement $\mathbf{m}$ has high probability in the languages of the world, we take this as evidence that many points $\mathbf{v}$ close to $\mathbf{0}$ are capable of generating $\mathbf{m}$.

Our universal prior has two main parameters: the mapping Ψ and a temperature T. In general Ψ is a complex function that could be modeled with some neural network (see e.g. Cotterell and Eisner, 2018, §6.1), whose own parameters could be estimated so as to explain a collection of observed speech datasets in different languages. The true Ψ and T should be likely to generate systems that predict the observed phonetics. Perhaps they should also assess those systems to be approximately locally optimal, on the grounds that functional pressures would otherwise have corrected them.[5]

[4] We can think of $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ as the result of corrupting $\boldsymbol{\mu}$ by both articulatory error and perceptual error, provided that these errors are distributed as independent Gaussians in a phonetic space, $\mathcal{N}(\mathbf{0}, \Sigma_{\mathrm{a}})$ and $\mathcal{N}(\mathbf{0}, \Sigma_{\mathrm{p}})$, with $\Sigma_{\mathrm{a}} + \Sigma_{\mathrm{p}} = \Sigma$. This view motivates our prior, which is based on communicative effort and rate when using $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ (§4.5). However, our measurements of phonemes usually measure actual sounds, *not* taking perceptual error into account, so they should really be regarded as samples from $\mathcal{N}(\boldsymbol{\mu}, \Sigma_{\mathrm{a}})$. We gloss over this distinction in this paper, effectively assuming $\Sigma_{\mathrm{p}} = \mathbf{0}$. This treatment could be improved.

[5] In addition, the true Ψ determines the probability that a token of phoneme j will be

Algorithm 2 Generative Process

for $\ell = 1$ **to** L :
  ▷ *For language ℓ, sample the parameter $\boldsymbol{\xi}^{(\ell)}$ of a distribution over abstract phonetic space*
  $\boldsymbol{\xi}^{(\ell)} \sim \pi$ ①
for $n = 1$ **to** N :
  ▷ *Sample a phoneme $i^{(n)}$ for the nth token*
  $i^{(n)} \sim p(\cdot \mid \boldsymbol{\xi}^{(\ell_n)})$ ②
  ▷ *Sample a representation $\mathbf{v}^{(n)}$ for the nth token*
  $\mathbf{v}^{(n)} \sim p(\cdot \mid i^{(n)}, \boldsymbol{\xi}^{(\ell_n)})$ ②
  ▷ *Obtain a measurement $\mathbf{m}^{(n)}$ of the nth token*
  $\mathbf{m}^{(n)} \leftarrow \Psi(\mathbf{v}^{(n)})$ or $\mathbf{m}^{(n)} \sim \Psi(\cdot \mid \mathbf{v}^{(n)})$ ③

4.4 A Generative Model: Overview

If we observe measurements of N tokens in languages $\ell_1, \ldots \ell_N$ respectively, we assume they were generated by the story in Algorithm 2. ① For each language ℓ represented in the dataset, functionalist language evolution independently drew a distribution over abstract phonetic space from our universal prior. More precisely, it drew a Gaussian mixture model with parameter $\boldsymbol{\xi}^{(\ell)}$. ② For each token $1 \leq n \leq N$, a speaker of language ℓ_n drew an abstract phone $\mathbf{v}^{(n)}$ from the $\boldsymbol{\xi}^{(\ell_n)}$ distribution, using some mixture component $i^{(n)}$. We regard $i^{(n)}$ as a phoneme of language ℓ_n; it will be observed in our experiments (§4.10.2). ③ The speaker pronounced $\mathbf{v}^{(n)}$ and a linguist obtained a surface measurement $\mathbf{m}^{(n)}$ of it, which is observed in our experiments.

We will first detail step ② and phonetic space. Once we formalize the distributions we will be working with, we will go back and explain our construction of the prior ① over this space of distributions. Finally, we will

perceived as phoneme i ($\forall i, j$). Thus, lab measurements of these confusion probabilities could further inform the estimation of Ψ.

conclude with a discussion of how to transform a distribution over phonetic space into one over measurement space in ③.

4.5 Distribution Over Phonetic Space

In a language with K phonemes, each phoneme $1 \leq i \leq K$ is some Gaussian distribution over the universal abstract phonetic space $\mathbb{R}^d$,

$$G_i \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i) \tag{4.1}$$

with mean $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and covariance matrix $\Sigma_i \in \mathbb{S}^d_+$ (where $\mathbb{S}^d_+$ is the set of positive definite symmetric $d \times d$ matrices). Σ_i characterizes noise in the articulation and/or perception of phoneme i. We will sometimes express Σ_i in terms of its Cholesky factorization $\mathbf{L}_i\mathbf{L}^\top$. For each token of the language, we have random variables

- I is a phoneme in $\{1, \ldots, K\}$
- V is its abstract phonetic realization in $\mathbb{R}^d$
- M is its measurement in $\mathcal{M}$

The first two have joint distribution

$$p(I = i, V = \mathbf{v}) = \underbrace{p(I = i)}_{\text{call this } w_i}\ \underbrace{p(V = \mathbf{v} \mid I = i)}_{G_i(\mathbf{v})} \tag{4.2}$$

Marginalizing over i, this means that each token $\mathbf{v}$ is drawn (independently) from a Gaussian mixture model with mixture weights w_i and components G_i. This distribution is fully specified by a parameter $\boldsymbol{\xi}$ that consists of K mixture weights $\mathbf{w}$, K means $\{\mu_i\}$, and K positive definite covariance matrices $\{\Sigma_i\}$.

Thus, for a fixed K, $\boldsymbol{\xi}$ falls in

$$\Omega(K) = \Delta^{K-1} \otimes \underbrace{\mathbb{R}^d \otimes \cdots \otimes \mathbb{R}^d}_{K \text{ times}} \otimes \underbrace{\mathbb{S}^d_+ \otimes \cdots \otimes \mathbb{S}^d_+}_{K \text{ times}} \tag{4.3}$$

where $\Delta^{K-1} = \{(w_1, \ldots, w_K) \in \mathbb{R}_{\geq 0} : \sum_i w_i = 1\}$ is the probability simplex of possible $\mathbf{w}$vectors. More generally, $\boldsymbol{\xi}$ can range over the space $\bar{\Omega} \stackrel{\text{def}}{=} \bigsqcup_{K=1}^{\infty} \Omega(K)$. We now define a prior distribution over this parameter space $\bar{\Omega}$.

4.6 Construction of the Prior

In this section, we will define a probability measure π over vowel systems as in (4.2). Our goal is to construct a distribution that reflects linguistic intuition about what makes a good vowel system $\boldsymbol{\xi}$. Classic work in vowel system theory has posited that there are two principles in the construction of a "good" vowel system. The first principle is known as **focalization** (Schwartz et al., 1997), which states that certain vowels are acoustically "better" than others—vowel systems should favor these vowels. The second principle, **dispersion** (Liljencrants and Lindblom, 1972), states that good vowel systems should consist of well-separated vowels. These two principles, however, are descriptive in nature and do not quantify the functional pressures. In this work, we justify focalization and dispersion information-theoretically, taking a view of vowel systems as a communication channel.

Given a vowel system $\boldsymbol{\xi}$, we will focus on two functional pressures: the average effort it takes to communicate one phoneme and the average information communicated by doing so. Our prior simply favors systems that are efficient in the sense that they achieve a high ratio (bits per unit effort):

$$p_{\text{prior}}(\boldsymbol{\xi}) \;\propto\; \pi(\boldsymbol{\xi}) \;\stackrel{\text{def}}{=}\; \left(\frac{\mathit{rate}(\boldsymbol{\xi})}{\mathit{effort}(\boldsymbol{\xi})}\right)^{1/T} \tag{4.4}$$

The **temperature** $T > 0$ controls the degree to which more efficient systems are more probable.[6]

Articulatory and Perceptual Effort

In our work, we take ease of articulation itself to reflect three forces. The first is **focalization**, which encourages vowels to live in a certain part of the phonetic space. The second is **imprecision**—vowels with higher variance are easier to produce. The third is **complexity**, which encourages the overall number of phonemes to be small.

Focalization. Focalized vowels (Stevens, 2000) are those that are easier to articulate and perceive. Much work has been done on determining the correlates of focalized vowels in measurement space. We simply stipulate that a sound is focalized to the degree that it has a small squared-length in our abstract phonetic space. The ease of pronouncing vowel i is the exp of its expected focality:

$$\exp\left(-\mathop{\mathbb{E}}_{\mathbf{v}\sim G_i}\left[\,||\mathbf{v}||_2^2\,\right]\right) \tag{4.5}$$

Imprecision. If we defined e_i as (4.5), then the easiest vowel would be the Dirac delta function, a degenerate Gaussian that always pronounces the most

[6] One could imagine using different T values in the numerator and denominator, to adjust the relative importance of the two quantities in the ratio.

focal possible sound, $\mathbf{0}$. However, that would require superhuman effort to articulate a perfect $\mathbf{0}$ every time, and noiseless transmission and perception to ensure that it is perfectly received. If a speaker is allowed some wiggle room, the vowel is easier to pronounce. Thus, languages prefer distributions G_i that have higher variation, which we quantify by $|\Sigma_i|^{\frac{1}{2}}$. Geometrically, $|\Sigma_i|^{\frac{1}{2}}$ is the volume of space that is "covered" by the Gaussian $G_i = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$, relative to what would be "covered" by the standard Gaussian $\mathcal{N}(\mathbf{0}, I)$.[7]

Putting It All Together. We define the ease of articulation of G_i as

$$e_i = |\Sigma_i|^{\frac{1}{2}} \cdot \exp\big(-\mathop{\mathbb{E}}_{\mathbf{v}\sim G_i}[\,||\mathbf{v}||_2^2\,]\big) \tag{4.6}$$

The following proposition will be needed later to show that (4.4) is a proper prior over vowel systems:

Proposition 1. *Recall that the ease e_i is a function of the Gaussian G_i over abstract phonetic space $\mathbb{R}^d$. For any $b > 0$, the integral of e_i^b over the space of possible Gaussians G_i is finite. Specifically,*

$$\int_{G_i} e_i^b = \sqrt{2\pi}\,\Gamma_d\left(\frac{b+d+1}{2}\right)\sqrt{\frac{1}{b^{bd+d^2+d+1}}} \tag{4.7}$$

We use $\int_{G_i}$ as shorthand for an integral $\int\int\cdots \mathrm{d}\boldsymbol{\mu}_i\, d\Sigma_i$ over the parameters of the Gaussian G_i (as in lemma 1).

Proof: See §4.12.

[7] The linear transform $\mathbf{x} \mapsto \mathbf{L}_i\mathbf{x} + \boldsymbol{\mu}$ transforms samples from $\mathcal{N}(\mathbf{0}, I)$ into samples from G_i, and transforms any sphere centered at $\mathbf{0}$ into an ellipsoid. The ellipsoid contains the same fraction of samples from G_i that the sphere contained from $\mathcal{N}(\mathbf{0}, I)$, but its volume is $|\mathbf{L}_i| = |\Sigma_i|^{\frac{1}{2}}$ times larger.

We can rewrite e_i^b as proportional to the product of a probability density over $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and a probability density over $\Sigma_i \in \mathbb{S}_+^d$:[8]

$$e_i^b = \exp\big(- b \cdot \mathop{\mathbb{E}}_{\mathbf{v}\sim G_i} [\, \mathrm{tr}(\mathbf{v}\mathbf{v}^\top) \,]\big)|\Sigma_i|^{\frac{b}{2}} \tag{4.8}$$

$$= \exp\big(- b \cdot \mathrm{tr}(\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top + \Sigma_i)\big)\; |\Sigma|^{\frac{b}{2}} \tag{4.9}$$

$$= \underbrace{\exp(-b \cdot ||\boldsymbol{\mu}_i||_2^2)}_{\propto\ \textit{Gaussian}}\ \underbrace{\exp(-b \cdot \mathrm{tr}\,\Sigma_i)\ |\Sigma_i|^{\frac{b}{2}}}_{\propto\ \textit{matrix gamma}}$$

The double integral of e_i^b over $\boldsymbol{\mu}_i$ and Σ_i is thus the product of the finite normalizing constants for the two unnormalized densities shown.

Effort. Now that we have defined ease of articulation for a single phoneme i, we define effort as the inverse, $1/e_i$. The expected effort per token for vowel system $\boldsymbol{\xi}$ is then

$$\mathop{\mathbb{E}}_i[\frac{1}{e_i}] = \sum_{i=1}^{K} \frac{w_i}{e_i} \tag{4.10}$$

where w_i is the probability of phoneme i in the system. We define the effort of a system as

$$\textit{effort}(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \sum_{i=1}^{K} \frac{w_i + \varepsilon}{e_i} \tag{4.11}$$

This view of effort takes into account the pronunciability of each individual phoneme. We highlight that adding the small constant $\varepsilon > 0$ posits a small cognitive cost for including each vowel i in the system description (particularly an effortful vowel), regardless of that vowel's rate of actual use w_i. This means

[8]Step (4.9) uses a standard and easily proved identity about expectations (Petersen, Pedersen, et al., 2008, §7.3.2). Below (4.9), the first factor is the unnormalized density for the Gaussian distribution $\mathcal{N}(\mathbf{0}, \frac{1}{2b}I)$, while the second factor is the unnormalized density for the matrix gamma distribution $\mathrm{MG}_p(\frac{b+d+1}{2}, \frac{1}{b}, I)$.

that in contrast to Eq. (4.10), the effort of a 5-vowel system would now be slightly increased by adding an "invisible" sixth vowel with $w_6 = 0$, and thus, MAP inference on a dataset with 5 observed vowel types will not regard this 6-vowel solution as equally good. The modification is also needed to obtain Proposition 2 below.

4.6.1 Rate of Transmission

Taking an information-theoretic perspective on communication, a vowel system defines a noisy channel through which the speaker can send a phoneme $i \in [1, K]$ to a hearer. The quality of this vowel system depends on how well the transmitted signal $\mathbf{v}$ helps the hearer to recover the intended phoneme i. This is measured as the **mutual information** between two random variables:

$$rate(\boldsymbol{\xi}) \stackrel{\text{def}}{=} \mathrm{MI}(I;V) = \mathrm{H}(I) - \mathrm{H}(I \mid V) \tag{4.12}$$

This quantity is measured in units of bits per token. If semantic messages must be encoded as sequences of vowels[9] and transmitted through this channel, then (4.12) is the **channel capacity**. According to the channel coding theorem, codes can be constructed that transmit messages arbitrarily close to this rate, but no faster. Thus, languages whose vowel systems provide a higher rate (4.12) have room to be more functional.

Computing the mutual information $\mathrm{MI}(I;V)$ is not trivial. The first term

[9] In practice, of course, speech can employ consonants as well. However, we have left consonants for future work. Consonant systems seem to undergo an additional pressure to be "systematic," in the sense of having a factorial design.

in (4.12) is the entropy of the distribution over phonemes,

$$\mathrm{H}(I) = -\sum_{i=1}^{K} w_i \log_2 w_i \tag{4.13}$$

This quantity can be computed straightforwardly in $\mathcal{O}(K)$ time. The second term, $\mathrm{H}(I \mid V)$, however, is intractable.[10] Since it is an expectation, we can approximate it using sampling:

$$\begin{aligned}\mathrm{H}(I \mid V) &= \mathop{\mathbb{E}}_{(i,\mathbf{v})\sim p}[-\log_2 p(i \mid \mathbf{v})] \\ &= \sum_{i=1}^{K} w_i \mathop{\mathbb{E}}_{\mathbf{v}\sim G_i}[-\log_2 p(i \mid \mathbf{v})] \end{aligned}\tag{4.14}$$

Each $p(i \mid \mathbf{v})$ is computed with Bayes' theorem, representing an optimal hearer recovering information from $\mathbf{v}$ about random variable I:

$$p(i \mid \mathbf{v}) = \frac{p(i)\, p(\mathbf{v} \mid i)}{\sum_{j=1}^{K} p(j)\, p(\mathbf{v} \mid j)} = \frac{w_i\, G_i(\mathbf{v})}{\sum_{j=1}^{K} w_j\, G_j(\mathbf{v})} \tag{4.15}$$

This can be computed exactly in $\mathcal{O}(K)$ time.

We further use the reparameterization trick (Kingma and Welling, 2014) to rewrite (4.14) as

$$\sum_{i=1}^{K} w_i \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{N}(\mathbf{0},I)}[-\log_2 p(i \mid V = \mathbf{L}_i\mathbf{x} + \boldsymbol{\mu}_i)] \tag{4.16}$$

As before, a single sample gives us an unbiased stochastic estimate of the expectation—but now the gradient of that estimate with respect to $\boldsymbol{\mu}_i$ and Σ_i (holding $\mathbf{x}$ fixed), which can be found by automatic differentiation, gives us

[10] Otherwise $\mathrm{MI}(I; V)$ would be tractable. Since we can also write $\mathrm{MI}(I; V) = \mathrm{H}(V) - \mathrm{H}(V \mid I)$ and $\mathrm{H}(V \mid I)$ is tractable, it would follow that $H(V)$ was tractable. However, $H(V)$ is the entropy of a Gaussian mixture model, which is known to be hard (Huber et al., 2008).

an unbiased estimate of the gradient of the expectation. Averaging S samples to reduce variance, we can estimate $\mathrm{H}(I \mid V)$ and thus $\mathrm{MI}(I;V)$—along with their gradients—in $\mathcal{O}(SK^2)$ time. We improve to $\mathcal{O}(SK)$ time by moving the expectation in (4.16) outside the sum. Then each sample is shared across all i; we compute (4.15) for all i together in $\mathcal{O}(K)$ total time.

4.6.2 Choice of the Base Measure

Eq. (4.4) converts a functionalist score for vowel systems—the rate/effort ratio—into an unnormalized density π over the space $\bar{\Omega}$ of vowel systems. The space is continuous, so each *individual* system $\boldsymbol{\xi} \in \bar{\Omega}$ has probability 0. A measurable *set* of vowel systems, $A \subseteq \bar{\Omega}$, may have positive probability:

$$p_{\mathrm{prior}}(A) = \frac{\int_A \pi\, \mu(\mathrm{d}\boldsymbol{\xi})}{\int_{\bar{\Omega}} \pi\, \mu(\mathrm{d}\boldsymbol{\xi})} \tag{4.17}$$

The integrations are with respect to the base measure μ over $\bar{\Omega}$, which is usually taken to be Lebesgue measure.

Alas, the prior probability (4.17) is sensitive to the specific way that we described the systems. A cosmetic change of variable that transformed each $\xi \in A$ to an alternative description η of the same system would not change the rate, effort, or functional score of any system—but sadly, it would in general change the prior probability of the set.

This is because the change of variable stretches or compresses regions of $\bar{\Omega}$, increasing or decreasing their Lebesgue measure. It is equivalent to keeping the old parameterization but changing the measure μ in (4.17). In short, (4.4) specifies only a functionalist probability *density* over vowel systems.

To convert this to an actual probability *distribution* (4.17)—i.e., a measure on $\bar{\Omega}$—requires the further choice of a base measure μ, indeed one for which the denominator of (4.17) will be finite.

The choice of μ is not itself functionalist. Rather, it reflects the dynamics of language evolution. Suppose for example that language evolution under functionalist pressures follows the Metropolis algorithm (Metropolis et al., 1953). That is, at each time step, the current system $\boldsymbol{\xi}$ considers mutating into a specific new $\boldsymbol{\xi}'$ drawn from some $q(\cdot \mid \boldsymbol{\xi})$. If $\boldsymbol{\xi}'$ is at least as functional—that is, $\pi(\boldsymbol{\xi}') > \pi(\boldsymbol{\xi})$—then the mutation survives. If not, it still survives with probability $\pi(\boldsymbol{\xi}')/\pi(\boldsymbol{\xi})$. If this Markov chain on $\bar{\Omega}$ is irreducible and aperiodic, then its unique stationary distribution is the very Boltzmann distribution given by (4.17)—provided that μ is such that the denominator is finite and such that the proposal distribution q, when expressed as a density with respect to μ, satisfies $q(x' \mid x) = q(x \mid x')$ for all x, x'. Thus, we should choose a μ that is compatible with evolution's actual proposal distribution.

Attempting to choose a plausible μ from synchronic or diachronic linguistic data is beyond the scope of this paper. In our experiments, we make an admittedly arbitrary choice that at least keeps the denominator finite. Given our parameterization of vowel systems in §4.5 using a parameter space $\bar{\Omega} \stackrel{\text{def}}{=} \bigsqcup_{K=1}^{\infty} \Omega(K)$, we take μ to be very similar to Lebesgue measure on that space (e.g., the measure of the simplex Δ^{K-1} is $\frac{\sqrt{K}}{(K-1)!}$), except that we multiply the measures of $\Omega(K)$ and its subsets by $\left(\frac{1}{K!}\right)^{\lambda}$. Without this scaling factor (i.e., $\lambda = 0$), we suspect that the denominator becomes infinite, but when

$\lambda \geq \frac{d^2+d+1}{2}$, we can prove that the denominator is finite. Further increasing λ has the effect of causing the marginal probability $p_{\text{prior}}(K)$ to fall off more rapidly as K increases: if we assume Metropolis evolutionary dynamics, this corresponds to increasing the probability that proposed mutations lower rather than raise K.

Proposition 2. *For any fixed K and any $\varepsilon > 0$ and $T > 0$, the function $\pi(\boldsymbol{\xi})$ may be normalized into a probability density over $\Omega(K)$, which can be used as a prior over K-vowel systems.*

See §4.12 for proof.

Proposition 3. *For any $\lambda \geq \frac{d^2+d+1}{2}$, the function $\pi(\boldsymbol{\xi})$ in Proposition 2 may be normalized into a probability distribution over all of $\bar{\Omega} = \sqcup_{K=1}^{\infty} \Omega(K)$. This can be used as a prior over all vowel systems (of any size). The parameter λ may be regarded as a complexity penalty that discourages large K* a priori, *which enables the sum to converge.*

See §4.12 for proof.

4.6.3 Discussion

Our prior over systems looks only at the **efficiency** of each system, that is, its rate/effort ratio. Notice that increasing K allows a higher communication rate provided that the K phonemes are kept fairly distinct, but it requires higher effort to keep them distinct (by making the Gaussians compact or placing some of them far from the origin).

Importantly, when the systems are described in phonetic space, our prior scores the vowel systems with nearly no tuning required. It has does have three parameters to tune: the temperature T from (4.4), the per-vowel cost ε from (4.11) (which we expect to have little effect in practice), and the choice of measure μ from §4.6.2 (which we have reduced to a choice of λ). It

This contrasts with Cotterell and Eisner (2018), who fit an elaborate distribution over vowel systems without *explaining* why the many parameters of the distribution (a universal inventory of underlying phones with shared variance but different means) should have the values that they have. The universal prior in that paper was merely descriptive, whereas the one in this paper aims to be explanatory.

4.7 Measurement Model

In our experiments, the measurement space $\mathcal{M}$ is the space of formant triples (F_1, F_2, F_3) in $\mathbb{R}^3_{>0}$. While there are many plausible choices for the measurement mapping, in this work we opt to fit an m-layer feed-forward neural network for Ψ, where m is a hyperparameter. Following Cotterell and Eisner ("Probabilistic Typology: Deep Generative Models of Vowel Inventories"), we ensure that the neural network computes a diffeomorphism by using $\mathbb{R}^3$ for our abstract phonetic space as well as for every hidden layer of the neural network. The diffeomorphism property allows us to compute a closed-form solution to the neurally transformed Gaussian using the standard change-of-variable technique. Specifically, we can merge lines 5 and 6 in Algorithm 2 to sample $\mathbf{m}^{(n)} \sim \widetilde{p}(\cdot \mid i^{(n)}, \boldsymbol{\xi}^{(\ell)})$ where we define

$$\widetilde{p}\left(\mathbf{m}^{(n)} \,\middle|\, i^{(n)}, \boldsymbol{\xi}^{(\ell)}\right) \stackrel{\text{def}}{=} \tag{4.18}$$
$$p\left(\Psi^{-1}(\mathbf{m}^{(n)}) \,\middle|\, i^{(n)}, p^{(\ell)}\right) \left|\frac{\partial}{\partial \mathbf{v}} \Psi^{-1}\right|$$

where $\left|\frac{\partial}{\partial \mathbf{v}} \Psi^{-1}\right|$ is the absolute value of the Jacobian determinant of the measurement mapping. Note that Eq. (4.18) is not true *unless* $\Psi(\cdot)$ is a diffeomorphism. It is possible to relax the diffeomorphism assumption (e.g., in order to expand the dimensionality of phonetic space), but this requires a more complex inference procedure than in §4.8.

4.8 Inference

Given p_{prior} (specified by T, ε, λ), we may want to sample a vowel system from it. More generally, if we are also given token measurements (as in §4.4) along with the measurement model Ψ, we can sample a vowel system from the posterior as well. In this section, we provide a method for this.

Note that the sampler must automatically choose the number of phonemes K, much like a Dirichlet process mixture model (Ferguson, 1973) or a mixture of finite mixtures (Richardson and Green, 1997). In any of these nonparametric Bayesian clustering models, including ours, the prior discourages having too many clusters. However, as our prior is non-conjugate, we cannot simply use Gibbs sampling, as is common in other such models (Gelman et al., 2013). What distinguishes our functionalist prior is that it does not choose the *number* of clusters independently from their *shape* and *position*, but rather considers the communicative properties of the system as a whole.

Cotterell and Eisner (2018) also estimated Gaussian mixture models from unlabeled data. However, they modeled vowel types as *points* in abstract phonetic space. Its mixture components corresponded to cross-lingual clusters of vowel types, rather than monolingual clusters of vowel tokens. There was no notion of a vowel type in a language as a *region* or *distribution* in phonetic space, nor any notion of the communicative efficiency achieved by a system of K such regions.

We use a Metropolis–Hastings sampler (MH; Hastings, 1970; Metropolis et al., 1953). At a high level, MH iteratively updates latent variables $\boldsymbol{\theta}$ by taking randomized moves. In our case, $\boldsymbol{\theta}$ includes the assignment of tokens to clusters (phonemes) as well as the vowel system parameters $\boldsymbol{\xi}$ (the probabilities, means, and covariances of the phonemes). A move proposal $\boldsymbol{\theta}'$ is drawn from a proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is the current iterate. We accept the proposed move with probability

$$\min\left(1, \frac{p(\boldsymbol{\theta}', \mathcal{D})\, q(\boldsymbol{\theta} \mid \boldsymbol{\theta}')}{p(\boldsymbol{\theta}, \mathcal{D})\, q(\boldsymbol{\theta}' \mid \boldsymbol{\theta})}\right) \tag{4.19}$$

If the proposal is accepted, we overwrite $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}'$. If the proposal is rejected, we do not overwrite $\boldsymbol{\theta}$. If this sequence is continued for sufficiently long, the iterates (samples) are representative of the posterior distribution, $p(\boldsymbol{\theta} \mid \mathcal{D})$. Our MH sampler makes use of six move categories; we will explain each of these moves in detail.

Move I: Token Assignments. We propose a new phoneme i' for the n^{th} token from:

$$q_{\text{token}}(i') \propto w_{i'}\, \widetilde{p}(\mathbf{m}^{(n)} \mid i', \boldsymbol{\xi}^{(\ell_n)}) \tag{4.20}$$

As Eq. (4.20) is a discrete distribution over a set of K elements, we can sample from it in $\mathcal{O}(K)$ time. It can be shown that move is *always* accepted as it corresponds to Gibbs sampling.

Move II: Mixture Weights. We propose new mixture weights $\mathbf{w}'^{(\ell)}$ from:

$$q_{\text{weights}}(\mathbf{w}'^{(\ell)}) = \text{Dirichlet}(\mathbf{c}^{(\ell)} + \delta \cdot \mathbf{1}) \tag{4.21}$$

where $\mathbf{w}'^{(\ell)}$ are the language-specific mixture weights and $c_i^{(\ell)}$ counts the number of tokens assigned to phoneme i in language ℓ. The proposal distribution's hyperparameter $\delta > 0$ acts as a smoothing parameter that ensures a valid parameters to the Dirichlet, which are positive.

Move III: Gaussian Means. We propose a new language-specific mean $\boldsymbol{\mu}_i^{'(\ell)}$ from the following distribution:

$$q_{\text{mean}}(\boldsymbol{\mu}_i^{'(\ell)} \mid \boldsymbol{\mu}_i^{(\ell)}) = \mathcal{N}(\boldsymbol{\mu}_i^{(\ell)}, \tau^2\, \mathbf{I}) \tag{4.22}$$

where τ^2 is a hyperparameter of the proposal distribution. The hyperparameter τ^2 determines how conversative the sampler is in its proposal: The smaller τ^2, the more likely it is that $\boldsymbol{\mu}_i^{'(\ell)} \approx \boldsymbol{\mu}_i^{(\ell)}$.

Move V: Merge Phonemes. To update the number of clusters K, we make use of reversible-jump MCMC (RJMCMC; Green, 1995). Specifically, we employ the adaptation of Marrs (1998) for multivariate Gaussian mixture models. RJMCMC generalizes standard MH in that it allows number of parameters to change during sampling; this has the effect of allowing model selection during MH. In our setting, we require RJ-MH due to the fact that we have an unknown number of phonemes K. We introduce two new MH actions **merge** and **split**. Both operations function as the names suggest. The first action, merge, selects two existing phonemes and merges them together. The second action, split, selects an existing phoneme and splits it into two. We only provide Moves V and IV for the special case of a diagonal covariance matrix, which we denote as $\Sigma_i = \boldsymbol{\sigma}_i^2$. See Zhang et al. (2004) for an alternative approach that does not assume a diagonal covariance. However, we find that the diagonal covariance matrices are sufficient in this setting because the non-linear measurement map Ψ can learn the correlation structure.

Let Π be a partitioning of the tokens in the corpus (a set of sets that are mutually exclusive and exhaustive). At each step we update the partition. Sample $i, j \in \Pi$ as follows. Pick i uniformly at random; then pick j with probability proportional to $1/||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||_2^2$. Let $k = i \cup j$ denote the new phoneme. Thus, the proposed partition is $\Pi' = (\Pi - \{i, j\}) \cup \{k\}$. Given the sampled phonemes i and j, the merge action deterministically proposes the following parameters for the new cluster k:

$$w_k \leftarrow w_i + w_j \tag{4.23}$$

$$\boldsymbol{\mu}_k \leftarrow \frac{w_i\boldsymbol{\mu}_i + w_j\boldsymbol{\mu}_j}{w_i + w_j} \qquad\qquad \boldsymbol{\sigma}_k^2 \leftarrow \frac{w_i\boldsymbol{\sigma}_i^2 + w_j\boldsymbol{\sigma}_j^2}{w_i + w_j}$$

This proposal was given in Marrs (1998) and in Richardson and Green (1997).

It is a reasonable choice as it preserves the zeroth, first, and second moments of the distribution before and after the proposal. This keeps the acceptance probability relatively high. Furthermore, the heuristic is deterministic and differentiable.

Move VI: Split Phonemes. To split a phonene, we sample a phonene k uniformly from Π. Next, we randomly partition k into two sets i and j such that $k = i \cup j$ and $i \cap j = \emptyset$. Interestingly, we will sample the parameters of the split without committing to a specific paritioning. After we have the parameters of the split, we build a randomized partition by assigning vowels to the phoneme i according the a coin flip with bias equal to the probability that the vowel was generated from i versus j; this is a slightly modified version of Eq. (4.20) where i' is constrained to be either i or j.

To remedy this inverse problem, we sample an additional set of variables: $\alpha \sim \mathrm{Beta}(a, b)$, $\beta_t \sim \mathrm{Beta}(a, b)$, and $\gamma_t \sim \mathrm{Beta}(a, b)$ for $t \in \{1, \ldots, d\}$; we write $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ for the concatenation of the β_t and γ_t, respectively, into vectors. The general idea is that we use these additional variables to specify a differentiable map between the previous two phonemes and the new one. There are many different principles one could choose to create two new phonemes. A common approach is to ensure that the phonemes have the same first and second moments (Richardson and Green, 1997; Marrs, 1998), which is achieved by the updates

$$w_i \leftarrow \alpha \cdot w_k, \quad w_j \leftarrow (1-\alpha) \cdot w_k \tag{4.24}$$

$$\boldsymbol{\mu}_i \leftarrow \boldsymbol{\mu}_k - \boldsymbol{\beta} \cdot \boldsymbol{\sigma}_k \sqrt{\frac{w_j}{w_i}} \tag{4.25}$$

$$\boldsymbol{\mu}_j \leftarrow \boldsymbol{\mu}_k + \boldsymbol{\beta} \cdot \boldsymbol{\sigma}_k \sqrt{w_i \,/\, w_j} \tag{4.26}$$

$$\boldsymbol{\sigma}_i^2 \leftarrow \boldsymbol{\gamma} \cdot \boldsymbol{\sigma}_k^2 w_k \,/\, w_i \tag{4.27}$$

$$\boldsymbol{\sigma}_j^2 \leftarrow (\mathbf{1} - \boldsymbol{\gamma}) \cdot \boldsymbol{\sigma}_k^2 w_k \,/\, w_j \tag{4.28}$$

The proposed partition is $\Pi' = (\Pi - \{i \cup j\}) \cup \{i, j\}$. A sanity check: Plugging Eqs. 4.24–4.28 into Eqs. Eq. (4.23) gives you back the original weights, means and covariances.

Choosing Actions. We describe the acceptance criterion used by our MH sample. In contrast to standard MH, RJ-MH, which we need due to the variable number of phonemes, requires a slightly modified acceptance criterion. If we choose a split or a merge action, then our acceptance probability in Eq. (4.19) requires us to multiply in a factor of $\frac{1}{r(\mathbf{u})}\left|\frac{\partial \boldsymbol{\theta}'}{\partial(\boldsymbol{\theta},\mathbf{u})}\right|$ in the second argument of the min. This can be computed by automatic differentiation. The merge analogue is given by the reciprocal $\frac{1}{r(\mathbf{u})}\left|\frac{\partial \boldsymbol{\theta}'}{\partial(\boldsymbol{\theta},\mathbf{u})}\right|^{-1}$.

4.9 Learning

§4.8 assumed that we already knew T, ε, λ, and Ψ. If not, we use approximate maximum-likelihood estimation to fit these parameters on observed tokens, with L2 regularization of Ψ's parameters.

We use grid search to explore values of the first three parameters, trying $T \in \{.1 \cdot k \mid k \in \{1, \ldots, 20\}\}$, $\lambda \in \{1.0 + .1k \mid k \in \{0, \ldots, 10\}\}$ and $\varepsilon \in \{10^{-k} \in \{1, \ldots, 10\}\}$. For each setting of those variables, an innermost loop optimizes Ψ using a version of expectation–maximization (EM; Dempster, Laird, and Rubin, 1977). EM is a well-known algorithm for performing parameter estimation on models with latent variables. We approximate the E-step using Monte Carlo Markov Carlo (Wei and Tanner, 1990), as described in §4.8. As in generalized EM, our M-step does not maximize the expected log-likelihood but only improves it for several steps (Neal and Hinton, 1998). Specifically we run Adam (Kingma and Welling, 2014), a gradient-based optimizer, for a fixed number of iterations. We perform early stopping of EM using a held-out development set, which is also used to choose the regularization coefficient and neural network depth for Ψ.

We return the $T, \varepsilon, \lambda, \Psi$ for which the innermost loop achieves the highest expected log-likelihood, which lower-bounds log-likelihood.

4.10 Experiments

The experiments in this section represent a first step towards showing the efficacy of the universal prior over vowels.

4.10.1 Data

Our data are taken from the VoxClamantis v1.0 corpus (Salesky et al., 2020), which provides token-level phonetic measures for vowels and sibilants in more than 600 languages.

We here use the vowel token measurements derived from the 29 languages with highest-quality alignments, in which Epitran, a grapheme-to-phoneme (G2P) tool based on language-specific rules (Mortensen, Dalmia, and Littell, 2018), was used to create phonemic transcripts, which were then aligned to utterance-level audio segments using a multilingual ASR model (Wiesner et al., 2019) trained in Kaldi (Povey et al., 2011). The first four formants for each vowel token were extracted using Praat (Boersma and Weenink, 2019) at each quartile and decile. In this work, we use the measurements for the first three formants as measured at the midpoint.

4.10.2 Phonemic Annotations

Our dataset has the advantage that each token has already been labeled with the vowel type (phoneme) that it was likely intended to communicate.[11] Thus, a dataset that uses K distinct phoneme labels requires a system with K Gaussian mixture components (or more, but extra unused components are discouraged by $\epsilon > 0$).

[11] In general, such labeling involves choosing an inventory of phonemes and an inventory of words built out of them, and using these to produce an aligned phonemic transcript of raw speech (Salesky et al., 2020). Fully automating these steps as part of our inference would require additional models for word and sentence formation, as in Elsner et al. (2013).

$\frac{1}{1}$	$\frac{rate}{1}$	$\frac{1}{effort}$	$\frac{rate}{effort}$
150.23	120.43	129.14	**97.42**

Table 4.1: Average log-likelihood on the held-out fold.

The shape of each phoneme's Gaussian is largely inferrable from the observed tokens of that phoneme. Additional information comes from our prior, which says that the Gaussians—particularly the frequently used ones—should be high variance (for ease) and well-separated (for rate).

The phonemic annotations are valuable information because they tell us which acoustic distinctions actually convey meaning—something that is not obvious from unlabeled acoustic measurements.

4.10.3 Log-Likelihood Experiments

We randomly divided our 29 languages into four folds: three with 7 languages and one with 8 languages. The first set of experiments focuses on evaluating the log-likelihood on the model (and various baselines) using four-fold cross validation. In this setting, the phoneme labels for the individual vowel tokens are observed, so we only require MH without Moves I, V and VI. We run 1000 iterations of Monte Carlo EM. The first 100 iterations are performed without Ψ—however, after this "warm-up" period, we interleave samples from the MH sampler with gradient-based training; there, we run 100 iterations of our optimizer. We compare to three baselines that use pared-down versions of our prior: non-prior ($1/1$), $1/effort$ (just effort) and $rate/1$ (just rate). This allows us to test how much each bit of our prior contributes to the results. The results are displayed in Table 4.1. Our prior outperforms all three baselines.

	actual (test)	sampled
# vowels	7	9.52
dispersion	24.00	27.34

Table 4.2: Sampled vowel system metrics.

4.10.4 Sampled Vowel Systems

For each of four folds discussed above in §4.10.3, we sample 10,000 vowel systems from the prior distribution that we estimated on training data (Table 4.2) using the other three folds. We then look at diagnostic criteria of these sampled systems. The first bit we look at is the number of vowels in the sampled systems. The simplest statistic here is the average numbers of vowels in the sampled systems. This value is displayed in Table 4.1 We also display the average percentage of vowels above a certain number and the second is a determinantal dispersion metric: Specifically, we arrange the sampled vowels in a matrix $M \in \mathbb{R}^{K\times 3}$ where M_i is the vector of formants for the i^{th} sampled vowel and compute $\det(M^\top M)$. These two qualitative metrics give us some insight into the naturalness of the generated systems, compared to real systems that appear in test data.

4.10.5 Caveats

It is important to note that the experimental section is far from complete. First and foremost, we must acknowledge that 29 languages are not enough to make a broad scientific claim about the nature of human language. No matter how we carve up the languages or how many experiments we run, 29 languages will never be representative of the full spectrum of human language. However, as discussed in detail in Salesky et al. (2020), it is not currently possible to get a larger sample of languages with sufficiently clean formant values. We hope that the availability of data will change in the coming years and more comprehensive experiments will be able to be run. For now, it is only fair to characterize the experimental evidence in this chapter as weak.

So what do the experiments here show? They primarily show that, when holding out roughly a 1/4 of the 29 languages we have as test data, our model does fairly well in terms of predicting those held-out languages. This is not an overly good metric as it does not tell us what sampling from the prior looks like. To handle this case, we take samples from the prior using the full RJ-MCMC scheme. The idea is to ask—using qualitative diagnostics—whether the sampled systems look like real vowel systems. Using these experiments, the evidence tentatively says they do, but more work is needed to be done. Indeed, a major drawback of our experimental set-up is the lack of strong baselines. There has never been another modeling effort that generate vowel systems in a similar manner.

4.11 Conclusion

We developed a novel prior over vowel systems as a tension between two functionalist principles of communication: rate and effort. Our experimentation on 29 languages shows that using this prior helps us better fit the empirical distribution over vowel measurements in an ablation study.

4.12 Proofs

Proposition 2. *Recall that the ease e_i is a function of the Gaussian G_i over abstract phonetic space $\mathbb{R}^d$. For any $b > 0$, the integral of e_i^b over the space of possible Gaussians G_i is finite. Specifically,*

$$\int_{G_i} e_i^b = \sqrt{2\pi}\,\Gamma_d\left(\frac{b+d+1}{2}\right)\sqrt{\frac{1}{b^{bd+d^2+d+1}}} \tag{4.29}$$

We use $\int_{G_i}$ as shorthand for an integral $\int\int\cdots \mathrm{d}\boldsymbol{\mu}_i\, d\Sigma_i$ over the parameters of the Gaussian G_i (as in lemma 1).

Proof.

$$\begin{aligned}
&\int\int \exp(-b\cdot||\boldsymbol{\mu}_i||_2^2)\,\exp(-b\cdot\operatorname{tr}\Sigma_i)\,|\Sigma_i|^{\frac{b}{2}}\,d\boldsymbol{\mu}_i\,d\Sigma_i\\
&=\int \exp(-b\cdot||\boldsymbol{\mu}_i||_2^2)\,d\boldsymbol{\mu}_i\int \exp(-b\cdot\operatorname{tr}\Sigma_i)\,|\Sigma_i|^{\frac{b}{2}}\,d\Sigma_i\\
&=\sqrt{\frac{2\pi}{b}}\int \exp(-b\cdot\operatorname{tr}\Sigma_i)\,|\Sigma_i|^{\frac{b}{2}}\,d\Sigma_i\\
&=\sqrt{\frac{2\pi}{b}}\,\frac{\Gamma_d(\frac{b+d+1}{2})}{b^{\frac{b+d+1}{2}d}}\\
&=\sqrt{2\pi}\,\Gamma_d\left(\frac{b+d+1}{2}\right)\sqrt{\frac{1}{b^{bd+d^2+d+1}}}
\end{aligned}$$

Both integrals are standard results, corresponding respectively to the partition functions of the Gaussian distribution and the matrix gamma distribution.[12] Note that as $b \to 0$, the value (4.7) diverges, since $d^2+d+1>0$. □

[12]See https://en.wikipedia.org/wiki/Matrix_gamma_distribution.

Lemma 1. *We have* $\Gamma_d\left(\frac{b+d+1}{2}\right) \in [\pi^{(d-1)/2}, \infty)$.

Proof. First we show that $\Gamma_d\left(\frac{b+d+1}{2}\right) \geq \pi^{(d-1)/2}$:

$$\Gamma_d\left(\frac{b+d+1}{2}\right) \tag{4.30}$$

$$= \pi^{(d-1)/2}\prod_{i=1}^{d}\Gamma\left(\frac{b+d+1}{2}+\frac{(1-i)}{2}\right) \tag{4.31}$$

$$= \pi^{(d-1)/2}\prod_{i=1}^{d}\Gamma\left(\frac{b+d+2-i}{2}\right) \tag{4.32}$$

$$\geq \pi^{(d-1)/2} \tag{4.33}$$

Next we show that $\Gamma_d\left(\frac{b+d+1}{2}\right) < \infty$:

$$\Gamma_d\left(\frac{b+d+1}{2}\right) \tag{4.34}$$

$$= \pi^{(d-1)/2}\prod_{i=1}^{d}\Gamma\left(\frac{b+d+1}{2}+\frac{(1-i)}{2}\right) \tag{4.35}$$

$$= \pi^{(d-1)/2}\prod_{i=1}^{d}\Gamma\left(\frac{b+d+2-i}{2}\right) \tag{4.36}$$

$$\leq \pi^{(d-1)/2}\prod_{i=1}^{d}\left\lceil\frac{b+d+2-i}{2}\right\rceil! \tag{4.37}$$

$$< \infty \tag{4.38}$$

□

Lemma 2. *For any $T > 0$, $K > \frac{1}{T}$, $\varepsilon > 0$, and $w_i \in [0,1]$, with $w_i' = \frac{w_i+\epsilon}{1+K\epsilon} \in (0,1)$,*

$$\prod_{i=1}^{K} \int_{G_i} e_i^{w_i'/T} \leq C\left(\left(K + {}^{1}\!/\!{}_{\varepsilon}\right)T\right)^{\frac{d}{2T}+(d^2+d+1)\frac{K}{2}} \tag{4.39}$$

where $C = \sqrt{2\pi}\,\Gamma_d\left(\frac{1/T+d+1}{2}\right) < \infty$.

Proof.

$$\prod_{i=1}^{K} \int_{G_i} e_i^{w_i'/T} = \prod_{i=1}^{K} \sqrt{2\pi}\, \Gamma_d \left(\frac{w_i'/T + d + 1}{2} \right) \sqrt{\frac{1}{\left(\frac{w_i'}{T}\right)^{\left(\frac{w_i'}{T}\right) d + d^2 + d + 1}}} \tag{4.40}$$

$$\leq C \prod_{i=1}^{K} \sqrt{\frac{1}{\left(\frac{w_i'}{T}\right)^{\left(\frac{w_i'}{T}\right) d + d^2 + d + 1}}} \tag{4.41}$$

$$\leq C \prod_{i=1}^{K} \sqrt{\frac{1}{\left(\frac{\varepsilon}{(1+K\varepsilon)T}\right)^{\left(\frac{\varepsilon}{(1+K\varepsilon)T}\right) d + d^2 + d + 1}}} \tag{4.42}$$

$$= C \prod_{i=1}^{K} \sqrt{\frac{1}{\left(\frac{1}{(K+1/\varepsilon)T}\right)^{\left(\frac{1}{(K+1/\varepsilon)T}\right) d + d^2 + d + 1}}} \tag{4.43}$$

$$= C \prod_{i=1}^{K} \sqrt{((K + 1/\varepsilon)T)^{\left(\frac{1}{(K+1/\varepsilon)T}\right) d + d^2 + d + 1}} \tag{4.44}$$

$$= C \sqrt{((K + 1/\varepsilon)T)^{\frac{K}{(K+1/\varepsilon)T} d + (d^2 + d + 1)K}} \tag{4.45}$$

$$= C \left(((K + 1/\varepsilon)T)^{\frac{K}{K+1/\varepsilon} d + (d^2 + d + 1)TK} \right)^{\frac{1}{2T}} \tag{4.46}$$

$$\leq C \left(((K + 1/\varepsilon)T)^{d + (d^2 + d + 1)TK} \right)^{\frac{1}{2T}} \tag{4.47}$$

$$= C \left((K + 1/\varepsilon)T \right)^{\frac{d}{2T} + (d^2 + d + 1)\frac{K}{2}} \tag{4.48}$$

□

[12] To show this, we must show that $x^{\alpha x + \beta}$ increases monotonically with $x > 0$ when $\alpha = d, \beta = d^2 + d + 1$, or equivalently, that $\log(x^{\alpha x + \beta}) = (\alpha x + \beta) \log x$ increases monotonically. Its derivative $\alpha(\log x + 1) + \frac{\beta}{x}$ is positive everywhere, since the unique minimum of the derivative (at $x = \beta/\alpha$) is positive for any $\beta \geq \alpha > 0$.

Proposition 3. *For any fixed K and any $\varepsilon > 0$ and $T > 0$, the function $\pi(\boldsymbol{\xi})$ may be normalized into a probability density over $\Omega(K)$, which can be used as a prior over K-vowel systems.*

Proof. We wish to show that the following normalizing constant is finite:

$$\int \cdots \int \pi(\boldsymbol{\xi}) \tag{4.49}$$

The nested integrals integrate over all K-vowel systems $\boldsymbol{\xi}$, by integrating over the parameter spaces of the Gaussians $G_1, \ldots, G_K$ as well as the vector $\mathbf{w} = (w_1, \ldots, w_k)$ of mixture weight parameters. If $\varepsilon = 0$, then the overall ease of the system, $1/\mathit{effort}(\boldsymbol{\xi})$, is a weighted harmonic mean of the e_i ease values. In the more general case when $\epsilon \geq 0$, let $w_i' = \frac{w_i + \varepsilon}{Z}$, where $Z = 1 + K \cdot \varepsilon$, so that $\mathbf{w}'$ is again a distribution. This allows us to express $1/\mathit{effort}(\boldsymbol{\xi})$ as a a weighted harmonic mean divided by $Z \geq 1$. We use the fact that the weighted harmonic mean of positive quantities (4.53) is upper-bounded by their weighted geometric mean (4.54):[13]

$$\pi(p) = \left(\frac{rate(\boldsymbol{\xi})}{effort(\boldsymbol{\xi})}\right)^{\frac{1}{T}} \tag{4.50}$$

$$= \left(rate(\boldsymbol{\xi}) \cdot \frac{1}{\sum_{i=1}^{K} \frac{w_i+\varepsilon}{e_i}}\right)^{\frac{1}{T}} \tag{4.51}$$

$$= \left(rate(\boldsymbol{\xi}) \cdot \frac{1}{Z} \cdot \frac{1}{\sum_{i=1}^{K} \frac{w'_i}{e_i}}\right)^{\frac{1}{T}} \tag{4.52}$$

$$\leq \left(rate(\boldsymbol{\xi}) \cdot \frac{1}{\sum_{i=1}^{K} \frac{w'_i}{e_i}}\right)^{\frac{1}{T}} \tag{4.53}$$

$$\leq \left(rate(\boldsymbol{\xi}) \cdot \prod_{i=1}^{K} e_i^{w'_i}\right)^{\frac{1}{T}} \tag{4.54}$$

$$= rate(\boldsymbol{\xi})^{\frac{1}{T}} \cdot \prod_{i=1}^{K} e_i^{w'_i/T} \tag{4.55}$$

Integrating this quantity over the possible $G_1, \ldots, G_K$ remains difficult because the first factor depends jointly on all of $G_1, \ldots, G_K$. However, by replacing

[13]Why didn't we define the overall ease using the weighted geometric mean in the first place? Because it seemed natural to define overall effort (like overall rate) to be an average over the individual phonemes weighted by their use. It also seemed natural for the overall communicative efficiency of the system to be measured as a rate/effort ratio: that is, how much information can be communicated per quantum of effort? Our resulting definition has the virtue that it is invariant to the size of the communicative units. For example, suppose that phoneme bigrams are compositional: bigram ij has joint probability $w_i w_j$ and total effort $\frac{1}{e_i} + \frac{1}{e_j}$. Then if we analyzed the system of K^2 bigrams, the rate and effort per bigram would double the rate and effort per unigram, yielding the same communicative efficiency ratio. This property is a starting point for analyzing systems that are not fully compositional, where for example j has higher probability and sloppier pronunciation when it follows i. Natural languages often achieve efficiency in this way—a kind of negotiation between phonology and phonetics.

that factor with a constant upper bound, we decouple the integrals, allowing us to use the distributive property:

$$\int_{G_1} \cdots \int_{G_K} \pi(\boldsymbol{\xi}) \tag{4.56}$$

$$\leq \int \cdots \int rate(\boldsymbol{\xi})^{\frac{1}{T}} \cdot \prod_{i=1}^{K} e_i^{w'_i/T} \tag{4.57}$$

$$\leq \int \cdots \int (\log_2 K)^{\frac{1}{T}} \cdot \prod_{i=1}^{K} e_i^{w'_i/T} \tag{4.58}$$

$$= (\log_2 K)^{\frac{1}{T}} \cdot \prod_{i=1}^{K} \int_{G_i} e_i^{w'_i/T} \tag{4.59}$$

$$\leq (\log_2 K)^{\frac{1}{T}} \cdot C((K + {}^{1}\!/\!{}_{\varepsilon})T)^{\frac{d}{2T} + \frac{K}{2}(d^2+d+1)} \tag{4.60}$$

To prove the final result, we have to additionally integrate over all vowel systems $\mathbf{w} \in \Delta^{K-1}$. Because Eq. (4.60) is a finite upper bound on the integrand, the integral over all $\Omega(K)$ is at most the volume of the simplex Δ^{K-1} times that upper bound:

$$\int \cdots \int \pi(\boldsymbol{\xi}) \leq \mu(\Delta^{K-1}) \cdot (\log_2 K)^{\frac{1}{T}} \cdot C((K + {}^{1}\!/\!{}_{\varepsilon})T)^{\frac{d}{2T} + \frac{K}{2}(d^2+d+1)} \tag{4.61}$$

where $\mu(\Delta^{K-1})$ is the measure of the simplex. This proves the result. $\square$

Let us define the volume of the $(K-1)$-simplex by $\mu(\Delta^{K-1}) \stackrel{\text{def}}{=} \left(\frac{\sqrt{K}}{(K-1)!}\right)^{\lambda}$. Taking $\lambda = 1$ gives the volume under Lebesgue measure. For other values of λ, this definition in effect scales the measure of any set of K-vowel systems by a factor of $\left(\frac{\sqrt{K}}{(K-1)!}\right)^{\lambda-1}$. This is equivalent to a reparameterization of the mixture weights (see section 6.3) so that they range over a smaller simplex (if

$\lambda > 1$).

Proposition 4. *For any* $\lambda \geq \frac{d^2+d+1}{2}$, *the function* $\pi(\boldsymbol{\xi})$ *in Proposition 2 may be normalized into a probability distribution over all of* $\bar{\Omega} = \sqcup_{K=1}^{\infty}\Omega(K)$. *This can be used as a prior over all vowel systems (of any size).* λ *may be regarded as a complexity penalty that discourages large* K a priori, *which enables the sum to converge.*

Proof. Our goal is to show the following infinite sum converges:

$$\sum_{K=1}^{\infty} \int \cdots \int \pi(\boldsymbol{\xi}) \leq \sum_{K=1}^{\infty} \mu(\Delta^{K-1}) \cdot (\log_2 K)^{\frac{1}{T}} \cdot C((K + 1/\varepsilon)T)^{\frac{d}{2T} + \frac{K}{2}(d^2+d+1)} \tag{4.62}$$

$$= \sum_{K=1}^{\infty} \underbrace{\left(\frac{\sqrt{K}}{(K-1)!}\right)^{\lambda}}_{\substack{\text{volume of the simplex} \\ \text{with complexity penalty}}} \cdot (\log_2 K)^{\frac{1}{T}} \cdot C((K + 1/\varepsilon)T)^{\frac{d}{2T} + \frac{K}{2}(d^2+d+1)} \tag{4.63}$$

We prove convergence by the ratio test, which tells us that an infinite series $\sum_{n=1}^{\infty} a_n$ converges if $\lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| < 1$. We work out the limit of the ratio below:

$$\lim_{K\to\infty} \frac{\left(\frac{\sqrt{K+1}}{K!}\right)^{\lambda} \cdot (\log_2(K+1))^{\frac{1}{T}} \cdot C((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K+1}{2}(d^2+d+1)}}{\left(\frac{\sqrt{K}}{(K-1)!}\right)^{\lambda} \cdot (\log_2 K)^{\frac{1}{T}} \cdot C((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.64}$$

$$= \lim_{K\to\infty} \frac{\left(\frac{\sqrt{K+1}}{K!}\right)^{\lambda} \cdot (\log_2(K+1))^{\frac{1}{T}} \cdot ((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}}{\left(\frac{\sqrt{K}}{(K-1)!}\right)^{\lambda} \cdot (\log_2 K)^{\frac{1}{T}} \cdot ((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.65}$$

$$= \lim_{K\to\infty} \frac{\left((K-1)!\sqrt{K+1}\right)^{\lambda} \cdot (\log_2(K+1))^{\frac{1}{T}} \ ((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)+{}^1/_2(d^2+d+1)}}{\left(K!\sqrt{K}\right)^{\lambda} \ (\log_2 K)^{\frac{1}{T}} \ ((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.66}$$

$$= \lim_{K\to\infty} \frac{\left(\sqrt{K+1}\right)^{\lambda} \cdot (\log_2(K+1))^{\frac{1}{T}} \ ((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)+{}^1/_2(d^2+d+1)}}{\left(K\sqrt{K}\right)^{\lambda} \cdot (\log_2 K)^{\frac{1}{T}} \ ((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.67}$$

$$= \left(\underbrace{\lim_{K\to\infty} \frac{\sqrt{K+1}}{\sqrt{K}}}_{=1}\right)^{\lambda} \cdot \lim_{K\to\infty} \cdot \frac{(\log_2(K+1))^{\frac{1}{T}} \ ((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)+{}^1/_2(d^2+d+1)}}{(K)^{\lambda} \cdot (\log_2 K)^{\frac{1}{T}} \ ((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.68}$$

$$= \underbrace{\lim_{K\to\infty} \frac{(\log_2(K+1))^{\frac{1}{T}}}{(\log_2 K)^{\frac{1}{T}}}}_{=1} \cdot \lim_{K\to\infty} \frac{((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)+{}^1/_2(d^2+d+1)}}{(K)^{\lambda} \cdot ((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}+\frac{K}{2}(d^2+d+1)}} \tag{4.69}$$

$$= \underbrace{\lim_{K\to\infty} \frac{((K+1+{}^1/_\varepsilon)T)^{\frac{d}{2T}}}{((K+{}^1/_\varepsilon)T)^{\frac{d}{2T}}}}_{=1} \cdot \lim_{K\to\infty} \frac{((K+1+{}^1/_\varepsilon)T)^{\frac{K}{2}(d^2+d+1)+{}^1/_2(d^2+d+1)}}{(K)^{\lambda} \cdot ((K+{}^1/_\varepsilon)T)^{\frac{K}{2}(d^2+d+1)}} \tag{4.70}$$

$$= \lim_{K\to\infty} \frac{((K+1+{}^1/_\varepsilon)T)^{\frac{(d^2+d+1)}{2}}}{(K)^{\lambda}} \cdot \left(\underbrace{\lim_{K\to\infty} \frac{(K+1+{}^1/_\varepsilon)^K}{(K+{}^1/_\varepsilon)^K}}_{=e}\right)^{\frac{(d^2+d+1)}{2}} \tag{4.71}$$

$$= e^{\frac{(d^2+d+1)}{2}} \cdot \underbrace{\lim_{K\to\infty} \frac{((K+1+{}^1/_\varepsilon)T)^{\frac{(d^2+d+1)}{2}}}{(K)^{\lambda}}}_{=0} \tag{4.72}$$

$$= 0 \tag{4.73}$$

where Eq. (4.73) follows because $\lambda > \frac{(d^2+d+1)}{2}$ by assumption. Thus, Eq. (4.63) converges. $\square$

Chapter 5

Conclusion

The ideas in the thesis operate at two levels. At the higher level, the core contribution of the thesis is the idea that we should use generative modeling, an idea from machine learning, to address questions in computational typology. The general idea is that we can train a model on as many samples of languages as we can get and see how well we can generalize to new languages. Of course, it is necessary to reiterate the large number of caveats, discussed in Chapters 2–4, that go with this idea as a research program. First, languages are not i.i.d. Languages have areal and genetic relationships. Second, the languages for which we have data is a very biased sample. For historical reasons, we are more likely to have access to resources for languages that are spoken in opulent countries (North American, Western Europe and East Asia). Nevertheless, despite these practical limitations, the research program as a whole is cogent. Indeed, it is just as cogent as the aims of the generative grammarians, who seek to divine a universal grammar from the samples of language at hand. This thesis represents a small step in the direction of merging generative modeling and generative grammar. And, indeed, for some, our focus on modeling no more than de-contextualized vowel formants may seem like a truly tiny step. However, more than anything, this thesis represents a multi-year effort to take seriously the enterprise of modeling vowel formants.

Given the context of the above, we review what this thesis has achieved. The thesis starts with the basic premise that we should use generative modeling to approach the problem of explaining the vowel inventories of the languages of the world. Each chapter relaxes assumptions made by the previous model. Chapter 2 is the first chapter on modeling vowel systems. The content of the chapter was published at ACL 2017. The idea is to build a probabilistic model over sets of phonemes (encoded as IPA symbols) in a large database of languages. The phonemes are grounded in observed acoustic measurements by taking the average format measurements for those IPA symbols across languages. The work successfully demonstrated that a generative model imbued with a domain-specific prior, specifically a determinantal point process that naturally encodes dispersion, was able to outperform less informed baselines. The model also learned a universal scorer that was able to transform formant space non-linearly into a perceptual space with the goal of modeling how humans may perceive vowels. In the experiments, the non-linear transformation is performed by a learned neural network. Despite the empirical success of the method, many of the assumptions the model made are simply untrue. The primary problematic assumption is that we can talk about universal formants for IPA symbols. This assumption is wrong on two fronts. First, IPA symbols are chosen on a per-language basis. It does not make sense to talk about the /a/ in one language being the same as the /a/ in another.

Furthermore, they are often chosen on a pragmatic basis: If a language has a single sound in a region of vowel space then the most common IPA symbol for that region will be used. For example, even if the sound is closer to /ae/, /a/ would be used for simplicity. Second, even if many languages have a phoneme that is closer to /a/ than to any other phoneme, averaging their formants across languages loses information that we ought to try to explain. Each language has its own system and the formants that it uses for /a/ may reflect the other vowels in the language.

Chapter 3 (second modeling chapter) attempts to remedy several of the problematic assumptions made above. First and foremost, the newer model no longer constitutes a distribution over discrete symbols that stand in for the phonetics, but rather over phonetic space itself. The model assumes a set of discrete prototypes, each of which has a corresponding Gaussian distribution. To generate the mean formant values for a language's vowels, we first sample a set of Gaussians from this universal prior and then we sample the (F1, F2) pairs from the Gaussians themselves before running them through a diffeomorphic neural network. The non-linear transformation allows for a similar warping between underlying perceptual space and observed acoustic space. To encode dispersion, we again make use of a determinantal point process, but this time the process is over sets of Gaussians rather than over sets of phonemes. This model represents a qualitative improvement over the previous model in that we are modeling a different type of data: Rather than modeling sets of discrete symbols, we are modeling sets of pairs of real numbers. However, the modeling paradigm still relies on modeling sets of vowel types, rather than vowel tokens. This is inconvenient as acoustic measurements of a vowel type may be ill-defined. The way the acoustic of a given vowel is realized depends on many factors; for instance, vowels are pronounced differently depending on what sounds are near them, i.e. coarticulation. Vowels may also have different allophones. The formant values used in this work stem from phonetic studies where the phonetician attempts to sample the formants of an idealized version of the vowels, rather than extracting them from large corpora of natural speech.

Chapter 4 (the third and final modeling chapter) focuses on a new type of model that is capable of modeling vowel tokens, rather than vowel types. Rather than coming up with a parametric distribution, as in the previous two models, this model relies on functionalist principles. The idea is to construct a universal prior over an abstract vowel space using our intuition of what might make a good vowel system. The prior over abstract vowel space takes a simple form with no more than three learnable parameters and says a good vowel system should have a good rate–effort ratio. The numerator, rate, describes how efficiently a vowel system can comment the phoneme id through the acoustic. The denominator, effort, describes how hard the vowel system is for a human to pronounce. The model also features, as the previous two, a universal non-linear mapping, modeled as a learnable neural network, from underlying abstract vowel space to an observed acoustic vowel space. We find that this prior is able to explain sequences of acoustic measurements better than baselines. This model concludes the modeling arc of the thesis, but more work should be continued.

Future work should focus on relaxing more problematic assumptions. First, a proper model of acoustic vowel tokens needs to be able to model co-articulation, allophonic variation as well as speaker variation. Much of the variance that is seen in the acoustic measurements will be predictable due to regular phonetic and phonological processes. The model in Chapter 5 does not attempt to explain this variance or show how such an explanation might reduce the rate and effort of the communicative system. Next, the broader aim of modeling an entire phonetic system should be considered. It is important to move beyond vowels and also model consonants. This, however, is a much harder endeavour. In contrast to vowels, which have clear measurable quantities, consonants are notably harder to describe phonetically (Lindblom and Maddieson, 1988).[1] Beyond acoustics, modeling vowel articulation is a necessary next direction along with modeling other aspects of linguistic systems: Phonology, morphology and syntax as well as other aspects of language should also be on the horizon. While the topic of the thesis was focused, we do hope it inspires future work on these topics—perhaps from the author himself. Indeed, as eluded to in the Chapter 4, functionalist accounts go hand in hand with an evolutionary account of language. A good question is whether we can learn both together, perhaps using the evolutionary tree of languages?

[1] Of course, some aspects of the acoustics of vowels, e.g. length, tone and creaky voice, require more than a few formants.

References

Ackley, David H., Geoffrey E. Hinton, and Terrence J Sejnowski (1985). “A learning algorithm for Boltzmann machines”. In: *Cognitive Science* 9.1, pp. 147–169.

Affandi, Raja Hafiz, Emily Fox, and Ben Taskar (2013). “Approximate Inference in Continuous Determinantal Processes”. In: *Advances in Neural Information Processing Systems*, pp. 1430–1438.

Affandi, Raja Hafiz, Alex Kulesza, and Emily B. Fox (2012). “Markov Determinantal Point Processes”. In: *UAI*, pp. 26–35.

Becker-Kristal, Roy (2006). “Predicting vowel inventories: The dispersion-focalization theory revisited”. In: *The Journal of the Acoustical Society of America* 120.5, pp. 3248–3248.

Becker-Kristal, Roy (2010). “Acoustic typology of vowel inventories and Dispersion Theory: Insights from a large cross-linguistic corpus”. PhD thesis. University of California, Los Angeles.

Boersma, Paul (1998). “Functional Phonology: Formalizing the interactions between articulatory and perceptual drives”. PhD thesis. University of Amsterdam.

Boersma, Paul and David Weenink (2019). *Praat: Doing Phonetics by Computer. Computer program, version 6.0. 46*.

Bronson, Barnard Sawyer (2004). *Phonetically Based Phonology*. Cambridge University Press.

Bybee, Joan (2003). *Phonology and Language Use*. Vol. 94. Cambridge University Press.

Chistovich, Ludmilla A. and Valentina V. Lublinskaya (1979). “The ‘center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli”. In: *Hearing Research* 1, pp. 185–195.

Chomsky, Noam and Morris Halle (1965). “Some controversial questions in phonological theory”. In: *Journal of Linguistics* 1.2, pp. 97–138.

Comrie, Bernard, Matthew S. Dryer, David Gil, and Martin Haspelmath (2013). "Introduction". In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Cooper, Franklin S., Pierre C. Delattre, Alvin M. Liberman, John M. Borst, and Louis J. Gerstman (1952). "Some Experiments on the Perception of Synthetic Speech Sounds". In: *The Journal of the Acoustical Society of America* 24.6, p. 597.

Cooper, Gregory F. (1990). "The Computational Complexity of Probabilistic Inference Using Bayesian Belief Networks". In: *Artificial Intelligence* 42.2-3, pp. 393–405.

Cotterell, Ryan and Jason Eisner. "Probabilistic Typology: Deep Generative Models of Vowel Inventories". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Cotterell, Ryan and Jason Eisner (2018). "A Deep Generative Model of Vowel Formant Typology". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, pp. 37–46.

Crothers, John (1978). "Typology and universals of vowel systems in phonology". In: *Universals of Human Language* 2, pp. 95–152.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22.

Deng, Li and Douglas O'Shaughnessy (2003). *Speech Processing: A Dynamic and Optimization-Oriented Approach*. CRC Press.

Dresher, B. Elan (2011). "The Phoneme". In: *The Blackwell Companion to Phonology*. Wiley Online Library. Chap. 11. ISBN: 9781444335262.

Elsner, Micha, Sharon Goldwater, Naomi Feldman, and Frank Wood (2013). "A Joint Learning Model of Word Segmentation, Lexical Acquisition, and Phonetic Variability". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 42–54.

"Eynard–Mehta theorem, Schur process, and their pfaffian analogs". In: ().

Ferguson, T. S. (1973). "A Bayesian Analysis of Some Nonparametric Problems". In: *Annals of Statistics* 1, pp. 209–230.

Fitch, W. Tecumseh, Marc D. Hauser, and Noam Chomsky (2005). "The evolution of the language faculty: Clarifications and implications". In: *Cognition* 97.2, pp. 179–210.
Flanagan, James L. (1957). "Estimates of the Maximum Precision Necessary in Quantizing Certain "Dimensions" of Vowel Sounds". In: *The Journal of the Acoustical Society of America* 29.4, pp. 533–534.
Fry, D.B., Arthur S. Abramson, Peter D. Eimas, and Alvin M. Liberman (1962). "The Identification and Discrimination of Synthetic Vowels". In: 5.4, pp. 171–189.
Garellek, Marc and Patricia Keating (2011). "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec". In: *Journal of the International Phonetic Association* 41.2, pp. 185–205.
Gartrell, Mike, Ulrich Paquet, and Noam Koenigstein (2016). "Low-Rank Factorization of Determinantal Point Processes for Recommendation". In: *CoRR* abs/1602.05436.
Gelman, Andrew, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin (2013). *Bayesian Data Analysis*. CRC Press.
Geman, Stuart and Donald Geman (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741.
Gordon, Matthew K. (2016). *Phonological Typology*. Oxford.
Gottfried, Terry L. and Traci L. Suiter (1997). "Dynamic information in the identification of Mandarin Chinese vowels and tones". In: *Journal of Phonetics* 25, pp. 207–231.
Green, Peter J. (1995). "Reversible jump Markov chain Monte Carlo computation and Bayesian model determination". In: *Biometrika* 82.4, pp. 711–732.
Hastings, W. K. (1970). "Monte Carlo Sampling Methods Using Markov Chains and Their Applications". In: *Biometrika* 57.1, pp. 97–109.
Hayes, Bruce P. (1999). "Phonetically Driven Phonology: The Role of Optimality Theory and Inductive Grounding". In: *Functionalism and Formalism in Linguistics: Volume I: General Papers* 41, p. 243.
Hayes, Bruce P. (2011). *Introductory Phonology*. John Wiley & Sons.
Hinton, Geoffrey E. and Terry J. Sejnowski (1986). "Learning and Relearning in Boltzmann Machines". In: *Parallel Distributed Processing*. Ed. by David E. Rumelhart and James L. McClelland. Vol. 2. MIT Press. Chap. 7, pp. 282–317.

House, Arthur S. and Grant Fairbanks (1953). “The Influence of Consonant Environment upon the Secondary Acoustical Characteristics of Vowels”. In: *The Journal of the Acoustical Society of America* 25.1, pp. 105–113.

Huber, Marco F., Tim Bailey, Hugh Durrant-Whyte, and Uwe D. Hanebeck (2008). “On entropy approximation for Gaussian mixture random vectors”. In: *IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*. IEEE, pp. 181–188.

Ising, Ernst (1925). “Beitrag zur Theorie des Ferromagnetismus”. In: *Zeitschrift für Physik A Hadrons and Nuclei* 31.1, pp. 253–258.

Jakobson, Roman (1941). *Kindersprache, Aphasie und allgemeine Lautgesetze*. Suhrkamp Frankfurt aM.

Jebara, Tony, Risi Kondor, and Andrew Howard (2004). “Probability Product Kernels”. In: *Journal of Machine Learning Research* 5, pp. 819–844.

Kingma, D. P. and M. Welling (2014). “Auto-Encoding Variational Bayes”. In: *International Conference on Learning Representations (ICLR)*.

Kulesza, Alex and Ben Taskar (2011). “Learning Determinantal Point Processes”. In: *UAI*, pp. 419–427.

Kulesza, Alex and Ben Taskar (2012). “Determinantal Point Processes for Machine Learning”. In: *Foundations and Trends® in Machine Learning* 5.2–3, pp. 123–286.

Ladefoged, Peter (2001). *Vowels and Consonants: An Introduction to the Sounds of Languages*. Wiley-Blackwell.

Ladefoged, Peter and Keith Johnson (2014). *A Course in Phonetics*. Nelson Education.

Ladefoged, Peter and Ian Maddieson (1996). *The Sounds of the World's Languages*. Oxford University Press.

Lehiste, Ilse and Gordon E. Peterson (1961). “Transitions, Glides, and Diphthongs”. In: *The Journal of the Acoustical Society of America* 33.3, pp. 268–277.

Levine, Richard A. and George Casella (2001). “Implementations of the Monte Carlo EM algorithm”. In: *Journal of Computational and Graphical Statistics* 10.3, pp. 422–439.

Liljencrants, Johan and Björn Lindblom (1972). “Numerical Simulation of Vowel Quality Systems: The Role of Perceptual Contrast”. In: *Language* 48.4, pp. 839–862.

Lindblom, B. E. F. and M. Studdert-Kennedy (1967). “On the Role of Formant Transitions in Vowel Recognition”. In: *The Journal of the Acoustical Society of America* 42.4, pp. 830–843.

Lindblom, Björn (1986). “Phonetic Universals in Vowel Systems”. In: *Experimental Phonology*, pp. 13–44.
Lindblom, Björn and Ian Maddieson (1988). “Phonetic Universals in Consonant Systems”. In: *Language, Speech, and Mind*, pp. 62–78.
Liu, Dong C. and Jorge Nocedal (1989). “On the limited memory BFGS method for large scale optimization”. In: *Mathematical Programming* 45.1-3, pp. 503–528.
Liu, Qiang, Jian Peng, Alexander T. Ihler, and John W. Fisher III (2015). “Estimating the Partition Function by Discriminance Sampling”. In: *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 514–522.
Macchi, Odile (1975). “The coincidence approach to stochastic point processes”. In: *Advances in Applied Probability*, pp. 83–122.
Maddieson, Ian (2013). “Vowel Quality Inventories”. In: *The World Atlas of Language Structures Online*. Ed. by Matthew S. Dryer and Martin Haspelmath. Leipzig: Max Planck Institute for Evolutionary Anthropology.
Marrs, Alan D. (1998). “An application of reversible-jump MCMC to multivariate spherical Gaussian mixtures”. In: *Advances in Neural Information Processing Systems*, pp. 577–583.
Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller (1953). “Equation of state calculations by fast computing machines”. In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.
Mortensen, David R., Siddharth Dalmia, and Patrick Littell (2018). “Epitran: Precision G2P for Many Languages”. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. Miyazaki, Japan: European Language Resources Association (ELRA).
Neal, Radford M. and Geoffrey E. Hinton (1998). “A view of the EM algorithm that justifies incremental, sparse, and other variants”. In: *Learning in Graphical Models*. Springer, pp. 355–368.
Nearey, Terrance M and Michael Kiefte (2003). “Comparison of Several Proposed Perceptual Representations of Vowel Spectra”. In: *Proceedings of the XV^{th} International Congress of Phonetic Sciences* 1, pp. 1005–1008.
Newmeyer, Frederick J. (2005). *Possible and Probable Languages: A Generative Perspective on Linguistic Typology*. Oxford University Press.
Petersen, Kaare Brandt, Michael Syskind Pedersen, et al. (2008). “The Matrix Cookbook”. In: *Technical University of Denmark* 7, p. 15.

Peterson, Gordon E. and Harold L. Barney (1952). “Control methods used in a study of the vowels”. In: *The Journal of the Acoustical Society of America* 24.2, pp. 175 –184.
Pinker, Steven and Ray Jackendoff (2005). “The faculty of language: what’s special about it?” In: *Cognition* 95.2, pp. 201 –236.
Povey, Daniel, Arnab Ghoshal, Gilles Boulianne, Lukáš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovský, Georg Stemmer, and Karel Veselý (2011). “The Kaldi Speech Recognition Toolkit”. In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.
Reichart, Roi and Anna Korhonen (2013). “Improved Lexical Acquisition through DPP-based Verb Clustering”. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Sofia, Bulgaria: Association for Computational Linguistics.
Richardson, S. and P. Green (1997). “On Bayesian Analysis of Mixtures with an Unknown Number of Components”. In: *Journal of the Royal Statistical Society B* 59, pp. 731–792.
Roark, Brian (2001). “Explaining Vowel Inventory Tendencies via Simulation: Finding a Role For Quantal Locations and Formant Normalization”. In: *NELS*. Vol. 31. 2, pp. 419–434.
Robert, Christian P. and George Casella (2005). *Monte Carlo Statistical Methods*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. ISBN: 0387212396.
Rosner, Burton S and John B Pickering (1994). *Vowel Perception and Production*. Oxford University Press.
Salesky, Elizabeth, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner (2020). “A Corpus for Large-Scale Phonetic Typology”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4526–4546.
Schwartz, Jean-Luc, Christian Abry, Louis-Jean Boë, Nathalie Vallée, and Lucie Ménard (2005). “The Dispersion–Focalization Theory of vowel systems”. In: *The Journal of the Acoustical Society of America* 117.4, pp. 2422–2422.
Schwartz, Jean-Luc, Louis-Jean Boë, Nathalie Vallée, and Christian Abry (1997). “The Dispersion–Focalization Theory of Vowel Systems”. In: *Journal of Phonetics* 25.3, pp. 255–286.
Shepard, Roger N. (1987). “Toward a Universal Law of Generalization for Psychological Science”. In: *Science* 237.4820, pp. 1317–1323.

Silverman, Daniel, Barbara Blankenship, Paul Kirk, and Peter Ladefoged (1995). “Phonetic structures in Jalapa Mazatec”. In: *Anthropological Linguistics* 37.1, pp. 70–88.
Stark, Henry and John Woods (2011). *Probability, Statistics, and Random Processes for Engineers*. Pearson.
Stevens, Kenneth N. (1989). “On the Quantal Nature of Speech”. In: *Journal of Phonetics* 17, pp. 3–45.
Stevens, Kenneth N. (2000). *Acoustic Phonetics*. MIT Press.
Sussman, Harvey M. (1990). “Acoustic correlates of the front/back vowel distinction: A comparison of transition onset versus “steady state””. In: *The Journal of the Acoustical Society of America* 88.1, pp. 87–96.
Taylor, Wilson L (1953). “Cloze Procedure: a New Tool for Measuring Readability”. In: *Journalism and Mass Communication Quarterly* 30.4, p. 415.
Valiant, Leslie G. (1979). “The Complexity of Computing the Permanent”. In: *Theoretical Computer Science* 8.2, pp. 189–201.
Van Lieshout, M. N. M. (2000). *Markov Point Processes and their Applications*. London: Imperial College Press.
Velupillai, Viveka (2012). *An Introduction to Linguistic Typology*. John Benjamins Publishing.
Volkovs, Maksims and Richard S. Zemel (2012). “Efficient sampling for bipartite matching problems”. In: *Advances in Neural Information Processing Systems*, pp. 1313–1321.
Wallach, Hanna, Ian Murray, Ruslan Salakhutdinov, and David Mimno (2009). “Evaluation methods for topic models”. In: *International Conference on Machine Learning (ICML)*, pp. 1105–1112.
Wei, Greg C. G. and Martin A. Tanner (1990). “A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms”. In: *Journal of the American Statistical Association* 85.411, pp. 699–704.
Whalen, D.H. and Andrea G. Levitt (1995). “The Universality of Intrinsic F_0 of Vowels”. In: *Journal of Phonetics* 23, pp. 349–366.
Wiesner, Matthew, Oliver Adams, David Yarowsky, Jan Trmal, and Sanjeev Khudanpur (2019). “Zero-Shot Pronunciation Lexicons for Cross-Language Acoustic Model Transfer”. In: *Proceedings of IEEE Association for Automatic Speech Recognition and Understanding (ASRU)*.
Yee, Connie, Nathan Keane, and Liang Zhou (2015). “Modeling and Characterizing Social Media Topics Using the Gamma Distribution”. In: *EVENTS*, pp. 117–122.

Zhang, Zhihua, Kap Luk Chan, Yiming Wu, and Chibiao Chen (2004). “Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm”. In: *Statistics and Computing* 14.4, pp. 343–355.

Ryan D. Cotterell

Contact Information	Department of Computer Science ETH Zürich CAB H 39.2 Universitätsstrasse 6 8092 Zürich Switzerland	*mobile:* +41 76 583 25 89 *email:* ryan.cotterell@inf.ethz.ch *www:* rycolab.io

Education

Johns Hopkins University — 2019
Ph.D. in Computer Science
Advisor: Jason Eisner

Ludwig-Maximilians-Universität München — 2014-2016
Visiting Ph.D. Student
Advisor: Hinrich Schütze

Johns Hopkins University — 2017
M.S.E. in Computer Science
Advisor: Chris Callison-Burch

Johns Hopkins University — 2013
B.A. in Cognitive Science (General and Deparmental Honors)g
Minor: Linguistics
Advisor: Colin Wilson

Employment History

ETH Zürich — 2020–
Assistant Professor in the Department of Computer Science

University of Cambridge — 2018–2020
Lecturer (≈Assistant Professor) in the Department of Computer Science and Technology

Institutional Responsibilities

I am a member of the graduate admissions committee for the MSc and direct doctorate programs at ETH Zürich.

Supervision of PhD Researchers

Paula Czarnowska *(Cambridge; coadvised with Ann Copestake)*	Fall 2018–
Tiago Pimentel *(Cambridge)*	Fall 2019–
Ran Zmigrod *(Cambridge)*	Fall 2019–
Marinela Parović *(Cambridge; coadvised with Anna Korhonen)*	Fall 2019–
Clara Meister *(ETH Zürich)*	Spring 2020–
Jennifer C. White *(Cambridge; coadvised with Simone Teufel)*	Fall 2020–
Rowan Hall Maudslay *(Cambridge; coadvised with Simone Teufel)*	Fall 2020–
Niklas Stöhr *(ETH Zürich)*	Fall 2020–
Karolina Stańczak *(Copenhagen through ELLIS; coadvised with Isabelle Augenstein)*	Fall 2020–
Tianyu Liu *(ETH Zürich; coadvised with Mrinmaya Sachan)*	Fall 2020–
Yuchen Jiang *(ETH Zürich; coadvised with Mrinmaya Sachan)*	Fall 2020–
Eleftheria Tspidi *(ETH Zürich)*	Spring 2021–
Jiaoda Li *(ETH Zürich; coadvised with Mrinmaya Sachan)*	Fall 2021–
Afra Amini *(ETH Zürich)*	Fall 2021–

Teaching Activities I have taught a graduate class on NLP (252-3005-00L Natural Language Processing) at ETH Zürich in Fall 2020 and Spring 2021. I taught at similar class at the University of Cambridge in Fall 2019.

Academic Service **Senior Area Chair:** EACL 2021 (Phonology, Morphology and Word Segmentation), NAACL 2021 (Phonology, Morphology and Word Segmentation), EMNLP 2021 (Linguistic Theories, Cognitive Modeling and Psycholinguistics)
Area Chair: NAACL 2019 (Machine Learning), ACL 2019 (Phonology, Morphology and Word Segmentation), ACL 2021 (Phonology, Morphology and Word Segmentation), EMNLP 2020 (Linguistic Theories, Cognitive Modeling and Psycholinguistics)
Journal Reviewer: TACL (2017, 2018), *Computational Linguistics* (2017, 2015), *Computer Speech and Language* (2017), *Natural Language Engineering* (2017)
Conference Reviewer: ACL (2021, 2020, 2019, 2018, 2017, 2016), EMNLP (2021, 2020, 2019, 2018, 2017, 2016), NIPS (2018), NAACL (2021, 2019, 2018, 2016), EACL (2017), CoNLL (2018), COLING (2018, 2016), ICLR (2018), LREC (2018), AAAI (2016 secondary)
Workshop Reviewer: Workshop on Analyzing and Interpreting Neural Networks for NLP (EMNLP 2018), Student Research Workshop (ACL 2018), Relevance of Linguistic Structure in Neural NLP (ACL 2018), Workshop on New Forms of Generalization in Deep Learning and Natural Language Processing (NAACL 2018), Ethics in NLP (NAACL 2018, EACL 2017), Workshop on Deep Structured Prediction (ICML 2017), Subword and character level models in NLP (NAACL 2018, EMNLP 2017), SIGMORPHON (EMNLP 2018, ACL 2016), Multilingual and Cross-lingual Methods in NLP (NAACL 2016)
Poster Session Chair: EMNLP (2017)
Other: SIGTYP Secretary, SIGMORPHON Officer At-Large (2nd Election), CLSP Happy Hour (and Karaoke) Coordinator

Organization of Workshops

1. SIGMORPHON 2019: 16th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (co-hosted at ACL 2019), SIGMORPHON 2020: 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (co-hosted at ACL 2020), SIGMORPHON 2021: 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology (co-hosted at ACL 2021)
2. First workshop on Typology for Polyglot NLP (co-hosted at ACL 2019), SIGTYP 2020: The 2nd Workshop on Computational Research in Linguistic Typology (co-hosted at ACL 2020), The 3rd Workshop on Research in Computational Typology and Multilingual NLP (co-hosted at NAACL 2021)

Scholastic Achievements and Awards

Honorable Mention for Best Paper at EMNLP (Meister et al., 2020)	2020
Nomination for Best Paper at ACL (Pimentel et al., 2019)	2019
Best Paper at ACL (Cotterell and Eisner, 2017)	2017
Outstanding Paper at EACL (Cotterell et al., 2017)	2017
Runner-up for Best Short Paper at NAACL (Cotterell et al., 2016)	2016
Honorable Mention for Best Short Paper at EMNLP (Müller et al., 2015)	2015
Facebook Ph.D. Fellowship	2018
Frederick Jelinek Fellowship	2017
National Defense Science and Engineering Fellowship (NDSEG)	2016-2018
DAAD Long-term Research Grant, Germany	2015-2016
Fulbright Research Grant, Germany	2014-2015
George M. L. Sommerman Engineering Graduate Teaching Assistant Award Finalist	2014
Computer Science Department Outstanding Teaching Assistant	2014
Cognitive Science Undergraduate Research Award	2013
Provost's Undergraduate Research Award	2012

Major Scientific Achievements

I have over 50 peer-reviewed publications in top-tier NLP conferences (defined as NAACL, ACL, EMNLP according to csrankings.org). My *h*-index is 26 according to Google scholar, which lists a total of 2185 citations as of March 30, 2021. A complete accounting of my scholarly activity, including recent invited talks, may be found at my lab's website rycolab.io. Below I have listed my award winning publications and my invited talks for completeness, which I consider part of my major scientific achievements.

Selected Publications

1. Ryan Cotterell, Nanyun Peng, and Jason Eisner. 2015. Modeling word forms using latent underlying morphs and phonology. *Transactions of the Association for Computational Linguistics (TACL)*, 3:433–447.

2. Clara Meister, Ryan Cotterell, and Tim Vieira. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2173–2185, Online. Association for Computational Linguistics

3. Tiago Pimentel, Arya D. McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics

4. Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics

5. Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 759–765, Valencia, Spain. Association for Computational Linguistics

6. Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 664–669, San Diego, California. Association for Computational Linguistics

7. Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with LEMMING. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics

Invited Talks

1. TBD
   University of Wolverhampton (May 2021)

2. Two New Insights into Beam Search
   ELLIS NLP 2021 (Feburary 2021)

3. Systematicity as Information
   Universität Zürich (December 2020), University of Geneva (December 2020), Moscow State University (November 2020)

4. Probabilistic Typology: Generating Languages from Scratch
   Athens Summer School (August 2019)

5. Mitigating Gender Bias in Morphologically Rich Languages
   22nd International Conference Text, Speech and Dialogue (Ljubljana; September 2019), University of Sheffield (April 2019), ETH Zürich (March 2019), New York University–Abu Dhabi (March 2019), Uppsala University (March 2019)

6. The Past Tense Debate on Steriod: Results from the CoNLL–SIGMORPHON Shared Task 2018 University of Colorado Boulder (January 2019)

7. Linguistics ∩ Machine Learning
Universität Zürich (November 2018), ETH Zürich (November 2018), Yale University (April 2018), Carnegie Melon University (April 2018), University College London (March 2018), University of Cambridge (March 2018)

8. Musings on Linguistic Complexity and Typology
University of Pennsylvania (November 2018), University of Copenhagen (February 2018)

9. Probabilistic Typology: Deep Generative Models of Vowel Inventories
Stony Brook University (November 2018), Universität Hamburg (January 2018), Massachusetts Institute of Technology (November 2017), Yale University (October 2017), New York University (August 2017), University of Edinburgh (January 2019), Northwestern University (January 2019), University of Chicago (January 2019)

10. Neural Weighted Finite-State Machines
First Workshop on Subword and Character Level Models in NLP at EMNLP (September 2017)

11. Neural Graphical Models over Strings
Universität Heidelberg (May 2017)

12. Neural String-Valued Graphical Models
Schloss Dagstuhl (January 2017)

13. Graphical Models over Strings
University of Alberta (October 2016), Johns Hopkins University (September 2016)

14. Modeling Word Forms Using Latent Underlying Morphs and Phonology
Universität Tübingen (July 2016)

15. Modeling Word Forms Using Latent Underlying Morphs and Phonology
Xerox Research Centre Europe (February 2016), Priberam Labs (September 2015)

16. A Probabilistic Approach to Synchronic Phonology
Institut für Phonetik und Sprachverarbeitung, LMU München (November 2014)

Publication List

Journal Publications

1. Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damian Blasi, and Hanna Wallach. 2021. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *Transactions of the Association for Computational Linguistics*

2. Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2021. Efficient computation of expectations under spanning tree distributions. *Transactions of the Association for Computational Linguistics*

3. Edoardo M. Ponti, Ivan Vulić, Ryan Cotterell, Marinela Parovic, Roi Reichart, and Anna Korhonen. 2021. Parameter space factorization for zero-shot learning across tasks and languages. *Transactions of the Association for Computational Linguistics*

4. Tiago Pimentel, Brian Roark, and Ryan Cotterell. 2020. Phonotactic complexity and its trade-offs. *Transactions of the Association for Computational Linguistics*

5. Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. Best-first beam search. *Transactions of the Association for Computational Linguistics*

6. Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342

7. Ryan Cotterell and Hinrich Schütze. 2018. Joint semantic synthesis and morphological analysis of the derived word. *Transactions of the Association for Computational Linguistics*, 6:33–48

8. Christo Kirov and Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6:651–665

Conference Publications

1. Irene Nikkarinen, Tiago Pimentel, Damián Blasi, and Ryan Cotterell. 2021. Modelling the unigram distribution. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics

2. Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics

3. Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2021. On finding the K-best non-projective dependency trees. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics

4. Clara Meister, Martina Forster, and Ryan Cotterell. 2021. Determinantal beam search. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics

5. Clara Meister, Stefan Lazov, Isabelle Augenstein, and Ryan Cotterell. 2021. On the interpretability of sparse attention. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics

6. Clara Meister and Ryan Cotterell. 2021. Analyzing the statistical tendencies of language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics

7. Jennifer C. White and Ryan Cotterell. 2021. Examining the inductive bias of neural language models with artificial languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics

8. Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2021. Higher-order derivatives of weighted finite-state machines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics

9. Jennifer C. White, Tiago Pimentel, Naomi Saphra, and Ryan Cotterell. 2021. A non-linear structural probe. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics

10. Josef Valvoda, Tiago Pimentel, Niklas Stoehr, Ryan Cotterell, and Simone Teufel. 2021. What about the precedent: An information-theoretic analysis of common law. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics

11. Rowan Hall Mauslay and Ryan Cotterell. 2021. Do syntactic probes probe syntax? Experiments with Jabberwocky probing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics

12. Tiago Pimentel, Brian Roark, Søren Wichmann, Ryan Cotterell, and Damián Blasi. 2021. Finding concept-specific biases in form–meaning associations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics

13. Tiago Pimentel*, Irene Nikkarinen*, Kyle Mahowald, Ryan Cotterell, and Damián Blasi. 2021. How (non-)optimal is the lexicon? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics

14. Tiago Pimentel, Ryan Cotterell, and Brian Roark. 2021. Disambiguatory signals are stronger in word initial positions. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics

15. Martina Forster, Clara Meister, and Ryan Cotterell. 2021. Searching for search errors in neural morphological inflection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics

16. Shijie Wu, Mans Hulden, and Ryan Cotterell. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics

17. Paula Czarnowska, Sebastian Ruder, Ryan Cotterell, and Ann Copestake. 2020. Morphologically aware word-level translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics

18. Clara Meister, Tim Vieira, and Ryan Cotterell. 2020. If beam search is the answer, what was the question? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

19. Tiago Pimentel, Naomi Saphra, Adina Williams, and Ryan Cotterell. 2020. Pareto probing: Trading off accuracy for simplicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

20. Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. 2020. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

21. Jun Yen Leung, Guy Emerson, and Ryan Cotterell. 2020. Investigating cross-linguistic adjective ordering tendencies with a latent-variable model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

22. Tiago Pimentel, Rowan Hall Maudslay, Damián Blasi, and Ryan Cotterell. 2020. Speakers fill semantic gaps with context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

23. Ran Zmigrod, Tim Vieira, and Ryan Cotterell. 2020. Please mind the root: Decoding arborescences for dependency parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

24. Francisco Vargas Palomo and Ryan Cotterell. 2020. Exploring the linear subspace hypothesis in gender bias mitigation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

25. Arya D. McCarthy, Adina Williams, Shijia Liu, David Yarowsky, and Ryan Cotterell. 2020. Measuring the similarity of grammatical gender systems by comparing partitions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

26. Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. Information-theoretic probing for linguistic structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

27. Rowan Hall Maudslay, Josef Valvoda, Tiago Pimentel, Adina Williams, and Ryan Cotterell. 2020. A tale of a probe and a parser. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

28. Adina Williams, Tiago Pimentel, Arya McCarthy, Hagen Blix, Eleanor Chodroff, and Ryan Cotterell. 2020. Predicting declension class from form and meaning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

29. Elizabeth Salesky, Eleanor Chodroff, Tiago Pimentel, Matthew Wiesner, Ryan Cotterell, Alan W Black, and Jason Eisner. 2020. A corpus for large-scale phonetic typology. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

30. Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of english than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

31. Alexander Erdmann, Micha Elsner, Shijie Wu, Ryan Cotterell, and Nizar Habash. 2020. The paradigm discovery problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

32. Clara Meister, Elizabeth Salesky, and Ryan Cotterell. 2020. Generalized entropy regularization or: There's nothing special about label smoothing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

33. Arya D. McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, Timofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal morphology. In *Proceedings of the Twelfth International Conference on Language Resources and Evaluation*. European Language Resources Association

34. Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. It's all in the name: Mitigating gender bias with name-based counterfactual data substitution. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics

35. Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damian Blasi, and Hanna Wallach. 2019. Quantifying the semantic core of gender systems. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5734–5739, Hong Kong, China. Association for Computational Linguistics

36. Edoardo Maria Ponti, Ivan Vulić, Ryan Cotterell, Roi Reichart, and Anna Korhonen. 2019. Towards zero-shot language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2900–2910, Hong Kong, China. Association for Computational Linguistics

37. Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 974–983, Hong Kong, China. Association for Computational Linguistics

38. Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics

39. Sabrina Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics

40. Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. Uncovering typological implications with belief nets. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3924–3930, Florence, Italy. Association for Computational Linguistics

41. Damian Blasi, Ryan Cotterell, Lawrence Wolf-Sonkin, Sabine Stoll, Balthasar Bickel, and Marco Baroni. 2019. On the distribution of deep clausal embeddings: A large cross-linguistic study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3938–3943, Florence, Italy. Association for Computational Linguistics

42. Shijie Wu and Ryan Cotterell. 2019. Exact hard monotonic attention for character-level transduction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1530–1537, Florence, Italy. Association for Computational Linguistics

43. Tiago Pimentel, Arya McCarthy, Damian Blasi, Brian Roark, and Ryan Cotterell. 2019. Meaning to form: Measuring systematicity as information.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1751–1764, Florence, Italy. Association for Computational Linguistics

44. Shijie Wu, Ryan Cotterell, and Timothy J. O'Donnell. 2019. Measuring morphological irregularity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5117–5126, Florence, Italy. Association for Computational Linguistics

45. Ran Zmigrod, Sabrina Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender bias in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics

46. Alexander M. Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics

47. Chaitanya Malaviya, Shijie Wu, and Ryan Cotterell. 2019. A simple joint model for improved contextual neural lemmatization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 1517–1528, Minneapolis, Minnesota. Association for Computational Linguistics

48. Shijia Liu, Adina Williams, Hongyuan Mei, and Ryan Cotterell. 2019. On the idiosyncrasies of the Mandarin Chinese classifier system. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 4100–4106, Minneapolis, Minnesota. Association for Computational Linguistics

49. Johannes Bjerva, Yova Kementchedjhieva, Ryan Cotterell, and Isabelle Augenstein. 2019. A probabilistic generative model of linguistic typology. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 1529–1540, Minneapolis, Minnesota. Association for Computational Linguistics

50. Alexander Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Ryan Cotterell, and Isabelle Augenstein. 2019. Combining sentiment lexica with a multi-view variational autoencoder. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers),

pages 635–640, Minneapolis, Minnesota. Association for Computational Linguistics

51. Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics

52. Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, Trevor Cohn, and Jason Eisner. 2019. Contextualization of morphological inflection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long and Short Papers), pages 2018–2024, Minneapolis, Minnesota. Association for Computational Linguistics

53. Yova Kementchedjhieva, Sebastian Ruder, Ryan Cotterell, and Anders Søgaard. 2018. Generalizing Procrustes analysis for better bilingual dictionary induction. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 211–220, Brussels, Belgium. Association for Computational Linguistics

54. Sebastian Ruder*, Ryan Cotterell*, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 458–468, Brussels, Belgium. Association for Computational Linguistics

55. Shijie Wu, Pamela Shapiro, and Ryan Cotterell. 2018. Hard non-monotonic attention for character-level transduction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4425–4438, Brussels, Belgium. Association for Computational Linguistics

56. Lawrence Wolf-Sonkin*, Jason Naradowsky*, Sabrina J. Mielke*, and Ryan Cotterell*. 2018. A structured variational autoencoder for contextual morphological inflection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 2631–2641, Melbourne, Australia. Association for Computational Linguistics

57. Ryan Cotterell and Jason Eisner. 2018. A deep generative model of vowel formant typology. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (Long Papers), pages 37–46, New Orleans, Louisiana. Association for Computational Linguistics

58. Ryan Cotterell, Sabrina J. Mielke, Jason Eisner, and Brian Roark. 2018. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 536–541, New Orleans, Louisiana. Association for Computational Linguistics

59. Ryan Cotterell, Christo Kirov, Sabrina J. Mielke, and Jason Eisner. 2018. Unsupervised disambiguation of syncretism in inflected lexicons. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2 (Short Papers), pages 548–553, New Orleans, Louisiana. Association for Computational Linguistics

60. Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina Mielke, Arya McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. UniMorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 1868–1873, Miyazaki, Japan. European Language Resources Association

61. Ryan Cotterell and Georg Heigold. 2017. Cross-lingual, character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics

62. Ryan Cotterell, Ekaterina Vylomova, Huda Khayrallah, Christo Kirov, and David Yarowsky. 2017. Paradigm completion for derivational morphology. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 714–720, Copenhagen, Denmark. Association for Computational Linguistics

63. Ryan Cotterell and Kevin Duh. 2017. Low-resource named entity recognition with cross-lingual, character-level neural conditional random fields. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, volume 2 (Short Papers), pages 91–96, Taipei, Taiwan. Asian Federation of Natural Language Processing

64. Francis Ferraro, Adam Poliak, Ryan Cotterell, and Benjamin Van Durme. 2017. Frame-based continuous lexical semantics through exponential family tensor factorization and semantic proto-roles. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics*, pages 97–103, Vancouver, Canada. Association for Computational Linguistics

65. Ryan Cotterell and Jason Eisner. 2017. Probabilistic typology: Deep generative models of vowel inventories. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume

1 (Long Papers), pages 1182–1192, Vancouver, Canada. Association for Computational Linguistics

66. Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. One-shot neural cross-lingual transfer for paradigm completion. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 1993–2003, Vancouver, Canada. Association for Computational Linguistics

67. Christo Kirov, John Sylak-Glassman, Rebecca Knowles, Ryan Cotterell, and Matt Post. 2017. A rich morphological tagger for english: Exploring the cross-linguistic tradeoff between morphology and syntax. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 (Short Papers), pages 112–117, Valencia, Spain. Association for Computational Linguistics

68. Ryan Cotterell, John Sylak-Glassman, and Christo Kirov. 2017. Neural graphical models over strings for principal parts morphological paradigm completion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 (Short Papers), pages 759–765, Valencia, Spain. Association for Computational Linguistics

69. Ryan Cotterell, Adam Poliak, Ben Van Durme, and Jason Eisner. 2017. Explaining and generalizing skip-gram through exponential family principal component analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 (Short Papers), pages 175–181, Valencia, Spain. Association for Computational Linguistics

70. Ekaterina Vylomova, Ryan Cotterell, Timothy Baldwin, and Trevor Cohn. 2017. Context-aware prediction of derivational word-forms. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 (Short Papers), pages 118–124, Valencia, Spain. Association for Computational Linguistics

71. Arun Kumar, Ryan Cotterell, Lluís Padró, and Antoni Oliver. 2017. Morphological analysis of the Dravidian language family. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 2 (Short Papers), pages 217–222, Valencia, Spain. Association for Computational Linguistics

72. Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2017. Neural multi-source morphological reinflection. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 514–524, Valencia, Spain. Association for Computational Linguistics

73. Ryan Cotterell, Arun Kumar, and Hinrich Schütze. 2016. Morphological segmentation inside-out. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2325–2330, Austin, Texas. Association for Computational Linguistics

74. Katharina Kann, Ryan Cotterell, and Hinrich Schütze. 2016. Neural morphological analysis: Encoding-decoding canonical segments. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 961–967, Austin, Texas. Association for Computational Linguistics

75. Tim Vieira*, Ryan Cotterell*, and Jason Eisner. 2016. Speed-accuracy tradeoffs in tagging with variable-order CRFs and structured sparsity. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1973–1978, Austin, Texas. Association for Computational Linguistics

76. Ryan Cotterell, Hinrich Schütze, and Jason Eisner. 2016. Morphological smoothing and extrapolation of word embeddings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1 (Long Papers), pages 1651–1660, Berlin, Germany. Association for Computational Linguistics

77. Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 664–669, San Diego, California. Association for Computational Linguistics

78. Pushpendre Rastogi, Ryan Cotterell, and Jason Eisner. 2016. Weighting finite-state transductions with neural context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 623–633, San Diego, California. Association for Computational Linguistics

Workshop Publications

1. Johannes Bjerva, Elizabeth Salesky, Sabrina J. Mielke, Aditi Chaudhary, Giuseppe G. A. Celano, Edoardo M. Ponti, Ekaterina Vylomova, Ryan Cotterell, and Isabelle Augenstein. 2020. SIGTYP 2020 shared task: Prediction of typological features. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, Online. Association for Computational Linguistics

2. Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, , and Simone Teufel. 2020. Metaphor detection using context and concreteness. In *Proceedings of the Second Workshop on Figurative Language Processing*, Online. Association for Computational Linguistics

3. Arya D. McCarthy, Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2018. Marrying universal dependencies and universal morphology. In *Proceedings of the Second Workshop on Universal Dependencies*, pages 91–101, Brussels, Belgium. Association for Computational Linguistics

4. John Sylak-Glassman and Ryan Cotterell. 2016. Contrastive morphological typology and logical hierarchies. In *Proceedings of the 52nd Annual Meeting of the Chicago Linguistic Society*, Chicago, Illinois. Chicago Linguistic Society

Invited Publications

1. Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online. Association for Computational Linguistics

2. Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics

3. Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels, Belgium. Association for Computational Linguistics

4. Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, , and Mans Hulden. 2017. CoNLL–SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver, Canada. Association for Computational Linguistics

5. Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics

Unrefereed Manuscripts

1. Ryan Cotterell and Julia Kreutzer. 2018. Explaining and generalizing back-translation through wake-sleep. *arXiv preprint arXiv:1806.04402*

2. Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2018. On the diachronic stability of irregularity in inflectional morphology. *arXiv preprint arXiv:1804.08262*

3. Chandler May, Ryan Cotterell, and Benjamin Van Durme. 2016. Analysis of morphology in topic modeling. *arXiv preprint arXiv:1608.03995*