Emulating Language Acquisition with Stochastic Gradient Descent: A New Approach to Modeling Phonotactics

Master’s Thesis

Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Graduate Program in Computational Linguistics
Professor James Pustejovsky, Advisor

In Partial Fulfillment
of the Requirements for the Degree

Master of Arts
in
Computational Linguistics

by
Frederic Jason Freyer

May 2017
Acknowledgments

I am grateful, first and foremost, to my advisor, James Pustejovsky, who took on this project despite the busiest of academic schedules and gave consistently insightful advice every week nonetheless.

Second, I must acknowledge my friends and colleagues at Cobalt Speech and Language, during an internship for whom I first had the idea for this thesis as a way to detect errors in entries in pronunciation dictionaries.

My appreciation also goes out, in no particular order, to:

- My classmates and professors in the Computational Linguistics Program at Brandeis, who—in addition to being really interesting people—deserve credit for keeping me interested in phonology throughout a professional degree program
- Patrick Jones, who advised and collaborated with me in my first independent research, which helped me develop lasting theoretical interests
- My family, especially my parents, without whose support through these many years I could never have been so successful
- My friends, especially my roommates, who have encouraged my through my thesis-writing, even though I am sure the word “phonotactics” still doesn’t mean that much to most of them!
Abstract

Emulating Language Acquisition with Stochastic Gradient Descent: A New Approach to Modeling Phonotactics

A thesis presented to the Graduate Program in Computational Linguistics

Graduate School of Arts and Sciences
Brandeis University
Waltham, Massachusetts

by Frederic Jason Freyer

Every language has a “feel” to it: a nebulous set of *phonotactic constraints* characterizing which combinations of sounds are favored, disfavored and prohibited in the formation of words.

I present a phonotactic learning system that achieves strong performance in modeling gradience in phonotactic judgments by combining a natural class-based approach (following Albright 2009) with a learning algorithm that focuses more strongly than past models on emulating human acquisition.

It has long been recognized (*e.g.* [26]) that phonotactic restrictions in languages are not binary, but rather represent a full spectrum between complete acceptability and unacceptability. Several experiments have verified that English speakers prefer, for example, /p tô/ to /f tô/ in syllable onsets, though both are legal; and /m tô/ to /vm tô/, though both are illegal.

Previous approaches to the computational modeling of phonotactics have been notably successful at learning hard constraints, but less so at learning gradient judgments. [15, 20, 2]

Learning in the present model is done by stochastic gradient descent, in which every word to which the model is exposed (representing a word that a learner hears) very slightly nudges upward acceptability values for features extracted from the word.
This kind of model represents a much more restricted learning environment than past models have used: the model only has access to one word at a time, and does only basic arithmetic calculations. I show that it is possible to substantially replicate phonotactic acceptability judgments—crucially including gradience—despite these restrictions, and using only about 1 million words of training data.

1 million words represent only a month or two of infant speech exposure, further suggesting that it is conceivable for babies to effectively learn phonotactics at a younger age than has been established in existing literature.

Finally, I illustrate how the feature values learned by such a model can be used to compute phonotactic similarity between languages, a useful typological measure.
Contents

Chapter 1. Introduction 1
  1.1. Goal of this Thesis 1
  1.2. Phonotactics 2
  1.3. Past Research on Probabilistic Phonotactics 4
  1.4. Overview 8

Chapter 2. Approach 9
  2.1. The Form of the Phonotactic Function 9
  2.2. The Feature Set 12
  2.3. The Learning Algorithm 18
  2.4. Local Summary 25

Chapter 3. Results 27
  3.1. The English Model 27
  3.2. Results 31
  3.3. Discussion 37

Chapter 4. Conclusion 40
  4.1. Future Directions 40

Bibliography 42
CHAPTER 1

Introduction

In this thesis, I will present a novel approach to using machine learning to model both the acquisition and intuitions of human phonotactics. I will further argue that stochastic gradient descent, the algorithm used, is a promising tool for use in other similar applications with implications for linguistic theory.

1.1. Goal of this Thesis

Computational linguistics is, broadly, the use of computers for language-related applications. As a hybrid field, it draws on both computer science and linguistics for inspiration in research methods. As a field with strong ties to industry, computational linguistics ranges from highly theoretical to highly practical, encompassing all levels of theoretical and computational sophistication.

Due to the prominence of natural language processing in industry, a large majority of contemporary research in computational linguistics, even within universities, has a very practical character. The best-researched subfields are those with the clearest immediate applications: machine translation, automated speech recognition, sentiment analysis and others. The best-researched languages are also the best-resourced: primarily English, with little substantial work done on the vast majority that are not United Nations-official.

Even beyond the tasks being researched, the practical nature of the field shows through in the methods and standards applied to new research. Contemporary work very often involves applying a new corpus, a new learning algorithm or a new feature to a precisely defined existing task, where results can be evaluated against large bodies of prior work. Success is
then defined in the tenths of a percent by which the new system outperforms its predecessors using the accepted evaluation metric.

While not denying the great value of this style of work—which has resulted in great progress in its target tasks, and trained a generation of highly skilled researchers—I endeavor in this thesis to depart from it.

My interest, and my goal as a researcher in computational linguistics, is to explore ways in which computation, and specifically machine learning, may be employed to advance our understanding of the language faculty: how linguistic knowledge of all kinds is acquired, stored and used.

Machine learning, properly used, has immense promise as a tool to implement and verify linguistic theories; to make generalizations about language data; and to imitate language-learning processes. Yet too often, researchers in computational linguistics—among the few with significant understanding of both theoretical linguistics and computer science—are drawn away from such work by the norms of the field.

My goal is thus both academic and idealistic. Academically, I intend to illustrate a novel approach to modeling language acquisition with machine learning, and to argue for its broader research implications. Idealistically, I hope to show how productive computational linguistic research may be done outside of the practical, evaluation-oriented norm that has defined the field in recent decades.

1.2. Phonotactics

Phonotactics is a subfield of phonology that examines the constraints on word formation in a language, including restrictions on phoneme sequences, syllable structure, and stress and tone patterns.

Research has shown that infants become aware of their native language’s phonotactics within months of birth. [8, 23] Speakers are very sensitive to these patterns, and when
presented with made-up words are quite consistent in their judgments of whether made-up words fit into the patterns of their language. In the case of English, \textit{clup} /klʌp/ is obviously fine: an “accidental gap” in the English lexicon that could easily mean something but happens not to. On the other hand, \textit{tvek} /tvɛk/ is very much not because English does not allow /tv/ as an onset consonant cluster.

Each language has its own specific phonotactic constraints. All languages have some restraints on allowable consonant clusters: English does not allow nasal consonants to precede other sounds in syllable onsets, for example, though such onsets are commonplace in many Bantu languages. (The Luganda word for “dog” is \textit{mbwa}.) Conversely, Bantu languages commonly prohibit word-final consonants, but English does not. Languages with stress or tone usually also have restrictions on which units of words may carry which kinds of suprasegmental markings.

In some cases, we can generalize about phonotactics across many languages. For example, consonants in word and syllable onsets are very often ordered such that sounds with greater vibration of vocal folds occur later (so onsets like /kl/ are far more common than /lk/). This and most other generalizations, however, have many glaring exceptions.

Speakers’ phonotactic judgements are also far finer than simple acceptance and rejection. When a language borrows a foreign word, for example, we can easily see at least four outcomes:

- The word is borrowed essentially unaltered (\textit{i.e.} by simply mapping each phone to the closest equivalent) and sounds fully native, e.g. \textit{chef} /ʃɛf/ from French.
- The word is borrowed essentially unaltered, and though it does not violate any rules also does not sound quite like a native word, e.g. \textit{perestroika} /ˈpɛrəstrɔɪkə/ from Russian.
• Though the word contains sound sequences that do not occur in native words in the borrowing language, it is borrowed unchanged, e.g. *Vladimir* /vlædɪmɪr/ from Russian.

• Speakers of the borrowing language add, remove or reorder phonemes to correct illegal sequences, e.g. *Sbarro* /sɔˈbaɾɔ/ from Italian (with an epenthesized schwa).

Indeed, past studies have shown that speakers of a language can consistently distinguish not only between words that are possible and impossible, [10] but within the class of possible words between those that consist of common and uncommon combinations of sounds. [30]

The problem we address in this paper is training a machine learning algorithm to make such judgements about the phonotactic acceptability of sequences of sounds.

### 1.3. Past Research on Probabilistic Phonotactics

The calculation of phonotactic probabilities, or acceptability scores, has been termed “probabilistic phonotactics.” [31] The literature is populated mostly by simple mathematical models correlated with observational studies (1.3.1), with some notably more theoretical and sophisticated models coming later (1.3.3).

#### 1.3.1. Early Work.

Early work on probabilistic phonotactics consisted of psycholinguistic studies that seeking to account for variance in test subjects’ performance on tasks like remembering and repeating nonce words, judging “wordlikeness” of nonce words, ability to remember names and spoken word recognition. Several test the relative effects of “phonotactics,” defined as effects of phone adjacency; and “lexical neighborhood density,” which is some variation on the number of known words within a short edit distance of the input nonce word.

There are numerous such studies, of which I will cite only a few here. Jusczyk et al. (1994) find that 9-month-old infants’ attention is held longer by sequences of phonotactically likely
CHAPTER 1. INTRODUCTION

non-words than by less likely ones, but that 6-month-olds show no such pattern. [23] Coady and Aslin (2004) find that young children (ages 2 and 3) are more likely to accurately repeat nonce words if they are phonotactically likely. [14] Storkel and Rogers (1999) find that 10- and 13-year-old children are more likely to remember phonotactically likely words after several minutes pass than unlikely ones, but that 7-year-olds are not. [29]

Vitevich and Luce (1999), among a number of others, report that test subjects’ repetition of nonce words is directly correlated with phonotactic probability, but inversely correlated with lexical neighborhood density. [31] They theorize that when a non-word is too close to a real word, the real word interferes with the ability to remember it, even though independently of that likely sound combinations are easier to process than unlikely ones. Bailey and Hahn (2001) find independent contributions of phonotactics and lexical neighborhoods on adults’ wordlikeness judgments. [4] They find lexical neighborhoods to be the more important of the two, and theorize that the significant amount of variation in judgments that cannot be accounted for by either measure is likely a result of imperfect measurement of lexical neighborhoods.

In all of the above-cited studies, the model used to calculate phonotactic probability of a phone sequence is very simplistic: a probability is assigned to each diphone (adjacent pair of sounds) in the input word based on how often it occurs in a dictionary, which may or may not be weighted by frequency. The probability of the full sequence is then the product of the probabilities of its component diphones. Sometimes the probability of each phone occurring in its specific position (word-initial, -medial or -final) is also multiplied in. I will scrutinize this and the other approaches below in more detail in chapter 2.

Another, somewhat more sophisticated model was given by Coleman and Pierrehumbert (1997), who calculated single probabilities for full onsets and full rhymes, as well as position relative to word edge (as above) and the stress pattern of the word. [15] They experimented
with calculating full-word probability as the product of the components, the logarithm of
the product of the components, and the least and most likely components. All but the last
gave statistically significant results. I return to this specific choice later.

1.3.2. A Note About Lexical Neighborhoods. Past studies acknowledge that lexi-
cal neighborhood density and phonotactic probability are necessarily highly correlated, but
nevertheless find they have complementary effects on acceptability judgments. It is my opin-
ion that (a) lexical neighborhoods are largely the same as phonotactics in this application;
(b) their complementary impact is due in large part to the inadequacy of the phonotac-
tic probability calculation; and (c) by restricting their test sets to words that are easily
pronounceable and monosyllabic (even CVC), past studies have excluded large classes of
judgments for which phonotactics are a far more plausible explanation.

It takes only a simple thought experiment to show that phonotactics must play a signifi-
cant role in acceptability: consider nonce words *vlat* /ˈvlæt/ and *tkan* /ˈtkæn/, which should
have decent lexical neighborhood density scores by any measure (edit distance of 1 to *flat*
and *vat* in the first case, and *scan* and *can* in the second). These words, however, are clearly
unacceptable English due to their onset clusters.

Clearly, therefore, phonotactics plays a role, and is thus worth studying both computa-
tionally and otherwise. The present study focuses solely on emulating phonotactics, and
as such does not concern itself past this subsection with the question of whether lexical
neighborhood density may also contribute to some judgments.

1.3.3. Natural Classes and More Powerful Machine Learning. A significant later
study by Hayes and Wilson (2008) takes an entirely different and much more linguistically
and computationally sophisticated approach to calculating phonotactic acceptability. [20] In-
puts in their system are evaluated according to a system of weighted constraints, with higher
weights representing greater penalties. The total penalty, which is inversely proportional to acceptability as discussed here, is the sum of the weights of all violated constraints.

The constraints used are sequences of natural classes of sounds. For example, a possible constraint *[-strid][+cons] in English would penalize non-strident sounds followed by non-glides, thus lowering the acceptability of bad onsets like /dl/ and /θl/.

The use of phonologically solid natural classes is a major step forward. Also significant is their application of maximum entropy learning to derive the constraint set. Because the number of possible constraints in their specification is so large, they first prune the constraints to include only the least-violated and most general according to a reference lexicon. They then optimize these feature weights using iterative gradient ascent based on the likelihood of the reference data according to the current weights at each step.

Hayes and Wilson report successfully applying this approach to learn patterns of English onset clusters, Shona vowel harmony and the full phonotactics of Wargamay.

A later paper by Albright (2009) follows Hayes and Wilson in the use of natural classes, but considers acceptability to be the sum of the probabilities of all natural classes to which a subsequence belongs, with the goal that the greater number of components will better capture gradient acceptability. [2] He calculates these probabilities according to lexical frequencies, and on the same English consonant cluster data reports lower overall accuracy but more consistent performance showing gradience across attested and unattested clusters. (Hayes and Wilson did not effectively capture varying judgments within attested clusters because they learned no constraints on things seen in the data.)

In a somewhat different vein, a dissertation by Heinz (2007) gives detailed analyses of a number of strategies for learning phonotactics, focused on their mathematical implications for the space of learnable formal grammars. [22] While it does not address gradient acceptability,
it does contain a number of relevant points that will be discussed as appropriate in the next chapter.

1.4. Overview

In Chapter 2, I will develop my theoretical approach to phonotactics, including the form of the function for calculating a word’s acceptability (2.1); the features of the word used as input to the function (2.2); and the learning algorithm (2.3).

In Chapter 3, I will discuss the implementation of the learning framework described in Chapter 2, using English phonotactics as a case study. I will illustrate the algorithm’s performance in learning English phonotactic constraints and in replicating acceptability judgments.

In Chapter 4, I will discuss future directions and implications of the work, and conclude.
CHAPTER 2

Approach

In this chapter, I will lay out my theoretical approach to modeling phonotactics. Drawing on the literature reviewed in 1.3, there are three notable and largely separate problems to address in building a system to judge phonotactic acceptability:

(1) How should phonotactic acceptability be calculated so as to reflect gradient judgments, given all necessary weights or probabilities?

(2) What features of an input word should contribute to a phonotactic acceptability judgment?

(3) How should a system learn those features?

The sections below will address each of these questions in turn, developing an approach to the problem in contrast with those of previous researchers.

2.1. The Form of the Phonotactic Function

In its most abstract form, the goal of probabilistic phonotactics is to define a phonotactic function \( \phi_L(w) \) that returns a numerical phonotactic acceptability score for an input word \( w \) in language \( L \). The right side of the equation, though, may take a number of different forms.

2.1.1. Output of the Function. It is worth briefly discussing the form the output of the function should take. A lot of linguistic literature thinks in terms of binary judgments: “yes,” this is correct, or “no,” it is not. As I have already established, however, linguistic
judgments in general—and phonotactic judgments in particular—are gradient. Within the
sets of acceptable and unacceptable sequences, there are still clear levels of acceptability.

The output of the function $\phi$ should therefore not be binary, but rather a numerical
value. (I see no reason for the final output to be more complex than a single value, though
multiple sub-results may be produced during the calculation.)

The range of the function is theoretically irrelevant, as long as it allows for arbitrary
specificity. For simplicity’s sake, I will assume that the range of outputs should be between
0 and 1, with 0 being the least acceptable and 1 being the most.

2.1.2. The Function as a Sum or Product of Probabilities. Early research incor-
porating probabilistic phonotactics calculated phonotactic acceptability as either the pro-
duct, sum or mean of the probability of each part of the input word: $[4, 8, 14, 23, 29, 31, 28]$

$$\phi_L(w) = \prod_{0<i \leq |w|} p(w_i)$$
$$\phi_L(w) = \sum_{0<i \leq |w|} p(w_i)$$
$$\phi_L(w) = \frac{\sum_{0<i \leq |w|} p(w_i)}{|w|}$$

where $p$ is a function returning the probability of a word component $w_i$. Note that I take no
position (yet) on what $w_i$ should represent—a phone, a sequence of phones or otherwise—for
now examining only whether these frameworks are appropriate for calculating acceptability.

There are problems with each of these formulations. If we use the sum or the product
of components, we find the undesirable behavior of strongly correlating phonotactic accept-
ability with word length: a direct correlation with the sum, and an inverse correlation with
the product.
Neither of these is a reasonable claim to make about a phonotactic function that we claim to be cross-linguistically applicable. If there is a correlation between length and acceptability, it is inverse: most languages do have a loose upper limit on word length. Even that general tendency, however, is highly language-specific: English tolerates much longer words than Mandarin, but is far more restrictive than an agglutinative language like Finnish. To argue that acceptability declines with length, and does so evenly across languages, is clearly false.

Some researchers recommend controlling for word length by taking the mean, $z$-score or geometric mean of the feature values. [28, 4] While these adjustments are certainly an improvement over a simple sum, there is another glaring problem with a sum of features that they do not resolve.

Intuitively, when a language prohibits a certain sound sequence—that is, when its phonotactic probability is low enough for speakers to reject it outright—it is impossible for a word containing that sequence to be phonotactically acceptable. The problem with considering a word’s total acceptability to be the sum of its parts is that, even when adjusted for length, this formulation implies that one bad component can be ameliorated by enough good components, which is clearly wrong. If an English word begins with /mb/, it doesn’t matter how many inoffensive sequences follow it; the word is not going to be acceptable.

### 2.1.3. The Function as a Minimum of Probabilities

A more consistent formulation would say that a word’s acceptability reflects only the least acceptable part, i.e. the minimum of the probabilities of its parts:

$$\phi_L(w) = \min\{p(w_i) : 1 < i \leq |w|\}$$

Thus, once an English speaker hears that initial /mb/, the rest of the word is irrelevant (unless another part is even worse).
Curiously, Coleman and Pierrehumbert (1997) report that in their study, a single “illegal” sequence is not actually enough to make a word universally unacceptable. Some of their subjects reported that the word *mrupation* /məˈrʌpəˈneɪn/, for example, would be an acceptable English word. Consistent with this finding, they found that the logarithm of the product of the component probabilities was a better predictor than the probability of the lowest component. (They note that this finding contradicts their own intuitions as well as foundational assumptions of generative phonology.)

Without seeing their full data (which is not included in the paper), it is difficult to judge how significant or solid this finding is. There are reasons to be skeptical, however, primarily because the only example given (*mrupation*) clearly contains a common English morpheme (-ation).

Acceptability judgments are conventionally tested on nonce words because of the presumed strong bias speakers have toward the acceptability of known items in their lexicon, which presumably extends to bound morphemes. I would suggest that when a nonce word is morphologically well-formed, even if phonotactically ill-formed, the risk of bias in subjects is significant. For example, if the prompt asks for some variation on “Could this be an English word?,” subjects may answer relative to morphology rather than phonotactics and corrupt the data.

In spite of Coleman and Pierrehumbert’s unexpected finding, I endorse the minimum-probability equation as the most intuitively plausible for calculating phonotactic acceptability. I will nevertheless test multiple formulations in chapter 3.

### 2.2. The Feature Set

In this section, I will go into more detail about the probability function $p$: the kinds of input it takes, and the kinds of features it extracts from each input.
2.2.1. Components of the Word. Most papers hitherto published have divided words into diphones (pairs of adjacent sounds) and calculated probabilities for each of those. So if the input word is *cactus* /ˈkæktəs/, the form of *w* is something like:

⟨/kæ/,/æk/,/kt/,/te/,/es/⟩

Previous studies have gotten away with this level of complexity by restricting their test data to monosyllables, but it is not hard to find cases of phonotactic restrictions that they cannot capture:

- English restricts the size of consonant clusters to 4, and all syllables must have a nucleus: A vowelless word like /ststs/ will have fine diphone probabilities, but is invalid English because it lacks a vowel and so many consecutive consonants are not allowed.
- English appears to disprefer sequences of /sC₁VC₁/, with identical consonants in the onset after /s/ and in the coda; except when C₁ is /t/ (*state*, *stout*, etc.), there are no words like /spæp/ or /slul/. [6] Diphones could not capture this tendency because of the space between the elements involved.
- Stress, and autosegmental concepts generally, involve awareness beyond the adjacent phone. For example, many languages have regularly penultimate stress, or vowel harmony rules. These necessarily require awareness beyond the single adjacent phone.
- Syllable boundaries are relevant to phonotactics: English allows clusters like /n.st₁/ and /t₁/ even though it prohibits them tautosyllabically.

To address these issues, I propose the following scheme:
CHAPTER 2. APPROACH

(1) Add word and syllable boundaries to the original input. (Both can be represented as /#/.) *Cactus* is then /#ˈkæk#təs#/; yielding the following input:

⟨/#/k/, '/kæ', '/æk/', /k#/ , /#t/, /tə/, /əs/, /s#/)⟩

(2) Include not only diphones, but all \( n \)-phones for \( 1 \leq n \leq 4 \). Most known phonological alternations operate within a three-phoneme window (Kenstowicz 1994, cited in [22]); we extend to four because we treat # as its own phoneme. This is very similar to the strategy taken in [20]. The uniphones are unlikely to have much effect on a model that only cares about the minimum feature, but I include them for completeness and because they are not a very large added load on the model. *Cactus* is now:

⟨
/#/ , /k/ , '/kæ' , /k/ , /'æk' , /k#/ , /#t/ , /tə/ , /əs/ , /s#/ ,
/#/k/ , '/kæ' , '/æk' , /k#/ , /#t/ , /tə/ , /əs/ , /s#/ ,
/#'kæ' , /'kæk' , /'æk#' , /k#t/ , /#tə/ , /təs/ , /əs#/ ,
/#'kæk' , /'kæk#' , /'æk#'t/ , /#təs/ , /təs#/⟩

(3) Add a separate autosegmental tier that includes only vowels; this will allow for better learning of vowel harmony, as well as stress and tone—once features are extracted, as discussed in 2.2.2 below. Within this tier, also give \( n \)-phones for \( 1 \leq n \leq 4 \) (though note that for \( n = 1 \) the features are necessarily redundant). The autosegmental tier for *cactus* might be written /#-ˈæ-ə-#/; meaning that the following new features would be added to the above list:

⟨#/ˈæ/ , /ˈæ-ə/ , /ə-#/ , /#-ˈæ-ə/ , /ˈæ-ə#/ , /#-ˈæ-ə-#/⟩
2.2.2. **Features Extracted from Phones.** Most past computational phonotactic systems have stopped here, and calculated probabilities for the sequences themselves. This is fine for making quick comparisons between observed sequences, but it has two significant weaknesses:

1. It may account well for gradient acceptability of attested sequences, but will not account for gradient acceptability of unattested sequences. Most English speakers would agree that /vl/ is significantly more acceptable as a syllable onset than /vk/, but as both of those onsets are all but unattested, the probability of each will be measured very close to 0.

2. Accuracy aside, the probabilities are not very theoretically interesting. Probabilities that reference only specific phonemes or phoneme sequences do not generalize (hence point 1), and cannot find patterns. If English prohibits /vk/, /zk/, /ðk/ and /ʒk/ as onsets, simple probabilities may replicate such behavior but will not learn the more interesting pattern, which is that /v/, /z/, /ð/ and /ʒ/ are all voiced fricatives. (The actual pattern is yet more general, but requires broader data to demonstrate.)

These issues can both be resolved by representing each phone not simply as a phone, but as a set of natural classes to which the phone belongs, as in [2] and [20].

Phonological theory holds that humans interpret and represent sounds in their heads as a series of “distinctive features,” which represent characteristics of the production of the sound. Research has shown that sets of sounds that are affected by the same alternation can generally be defined as a *natural class* because a certain set of distinctive features can define that group of sounds and no others in the language.

Distinctive features are therefore well-suited for capturing phonotactics because of their fundamental ability to generalize. If a feature-based framework sees a certain behavior affecting /v/ and /ʒ/, but not /s/ or /b/, the system has enough information to learn that
this behavior affects voiceless fricatives, which may be represented as [+cont +voice], and
generalize that it should affect /z/ and /ð/ as well—even if specific supporting data are
absent.

(In the chapter 3, I describe the specific model I constructed for English phonotactics
according to the framework developed here. See 3.1.3 for a description of the feature set
used for English in that model.)

Importantly, the generalizing ability of distinctive features should be able to extend
to reflecting gradient judgments of unattested sequences. Albright shows such an effect
in [2], where he proposes that speakers judge the acceptability of a sequence as the sum
of probabilities of the natural classes to which each sound belongs that occur in similar
environments.

So when an English speaker hears a word beginning with /vl/, they should make a number
of implicit judgments, including the following:

(1) /vl/ is almost never attested in syllable onsets.
(2) Voiced fricatives do not occur in consonant clusters.
(3) However, fricatives followed by approximants are common.
(4) Labials followed by approximants are similarly acceptable.
(5) Even labial fricatives followed by /l/ are substantially attested.

These judgments together should still leave /vl/ ranked well behind conventional conso-
nant clusters, but nevertheless should add up to a point at which it does not sound terrible.
A similar exercise with /bk/ would yield almost no substantially attested features beyond
“consecutive obstruents,” so we would expect—in line with our intuition—a much lower
score.
2.2.3. The Form of the Featurizing and Probability Functions. I follow Albright’s theory, and propose that we define a featurizing function $f$ as follows:

$$f(w_i) = \prod_{j=1}^{|w_i|} f'(w_{ij})$$

where $w_i$ is some $n$-phone substring of a word $w$ (and therefore $w_{ij}$ is a single phoneme) and $f'$ returns a set of natural classes to which that phoneme belongs:

$$f'(w_{ij}) = \{c : c \in C \land w_{ij} \in c\}$$

where $C$ is the set of natural classes in the language. What $f(w_i)$ returns, then, is the Cartesian product—a set of ordered tuples—of the natural classes that contain each of its component phonemes.

The probability function $p$, can then be phrased like this:

$$p(w_i) = \frac{\sum_{j \in f(w_i)} p'(j)}{|f(w_i)|}$$

where $p'$ returns the probability of a single feature. I choose to average the features partly to keep the final value between 0 and 1, and partly to control for the number of features (which will not be constant for all $n$-phone inputs, particularly between different values of $n$).

2.2.4. Alternate or Additional Features. It is an open question just what kind of information people store about the sounds they hear, and though I believe the feature set described in 2.2.2 is intuitive and should cover most phonotactic tendencies I can think of, I do not intend to make a stubborn claim that its scope is either necessary or sufficient for describing all languages.
Indeed, Heinz clearly articulates several long-distance and suprasegmental alternations that cannot be described by any \( n \)-gram models. For example, long-distance sibilant harmony in Navajo requires any two sibilants in the same word to be homorganic, regardless of the number of phonemes separating them; [22, p. 28–29] and Murik requires stress on the first heavy syllable of a word, or the first syllable if none are heavy. [22, p. 164–165]

Because these patterns apply to arbitrarily long distances, the maximum window for an \( n \)-gram would have to be set, post-hoc, based on the maximum word length in the language in order to have any reasonable shot at learning them. This is (1) cheating, and (2) incredibly inefficient, especially in light of the truism that a vast majority of phonological rules operate within a very small phoneme window.

Heinz is probably correct in arguing that long-distance alternations should be considered a different class of phonotactics than local alternations. As such, in the feature framework I use here, they deserve to be encoded with features that are not phrased as \( n \)-grams. I have not included such features in the model over the course of this work, but I fully acknowledge that my own feature set needs refinements and extensions to handle the full range of phonotactics observed in human language.

2.3. The Learning Algorithm

Finally, having worked through the other parts of the problem, we are left with the question of how to obtain the actual probabilities of each individual feature (\( p' \) in 2.2.3).

Emulating human acquisition of linguistic structures is an interesting challenge for machine learning for a number of reasons.

2.3.1. The Projection Problem. First, an adequate system must learn to classify inputs without observing any negative samples. The usual approach to learning classification problems like “Does this sound like a word?” or “Is this grammatical?” is to train using
substantial sets of data that fall into each relevant category: in this case, well-formed and ill-formed sequences of sounds. The classifier can then learn which characteristics of the inputs are most strongly characteristic of phonotactic acceptability and unacceptability.

Children, however, learn every part of their language from positive samples: researchers have observed that parents rarely correct even obvious speech errors, implying that the overwhelming majority of data on which learners base their linguistic generalizations are simply correct examples of language in use. This issue is known as the Projection Problem, and has for decades been a part of the central quandary in linguistics known as the Poverty of the Stimulus: the general idea that the linguistic data with which learners are presented is not in itself sufficient to perfectly derive the structures that are ultimately learned. This is used as a key argument in favor of a Universal Grammar that exists independently of outside stimuli. [5, 11, 12]

With respect to this point, it is helpful to observe that phonotactic judgments (like any linguistic judgments) are not binary, but rather a full spectrum. The problem is not classification, then, but rather regression, and we assume that the indicators of goodness and badness are in fact the same. We can therefore measure the presence and absence of features as evidence of their goodness and badness, and do not need specific negative evidence.

The analogue to Universal Grammar in this application comes in the feature set that we choose. I assume here that humans are innately able to segment speech into individual sounds, and to process those sounds into certain articulatory features. This is not a very difficult claim to make; indeed, it appears to be self-evidently true. I further claim that they can separate syllables on their own, which may not be quite as clear but is supported by studies of pattern recognition by infants (for example, [25]).
CHAPTER 2. APPROACH

It is far harder to make such solid claims about innate abilities relevant to acquisition of other linguistic skills (syntax, for example), but at least in the case of phonotactics we can rest comfortably on awareness of sounds as our primary assumption.

2.3.2. Accessibility of Stimuli. Aside from a lack of negative stimuli, we need to recognize that each stimulus a child hears is ephemeral, heard once and perhaps remembered for a bit but usually quickly forgotten except for any impact it has on the language faculty.

I argue that this means it is unreasonable for a learning algorithm to assume that a child has access to a full, on-demand corpus of stimuli from which to learn. Now, most research on probabilistic phonotactics has focused on imitating final intuitions, and makes no claim of accurately reflecting the true learning process. Replicating the learning as well as the result is an important part of my goal, however, so this is a standard I will hold myself to.

2.3.3. Past Approaches. Almost all works hitherto cited have calculated feature probabilities as simple frequencies (of diphones, features, etc.): the number of words in a lexicon or corpus of text that satisfy the feature divided by the total number of words. \[31, 15, 2, \text{among others}\]

Hayes and Wilson’s maximum entropy model uses a more complex calculation. They first prune their total possible feature set to include only the most accurate and most general constraints according to a corpus. They then train the weights of the constraints by gradient ascent to maximize the likelihood of the corpus. That is, at each step, weights are nudged upward or downward based on whether higher or lower weights will result in a better fit to the corpus, until they reach the maximum achievable corpus likelihood. \[20, \text{p. 383–395}\]

(Note that Hayes and Wilson’s constraint weights are not probabilities, per se, nor is the final output of their phonotactic function. Their function’s output is a weighted sum of constraint violations, and as such is inversely proportional to acceptability. I have used the
term “probability” to refer to it in my discussion to emphasize the similar role that it plays in the calculation.)

Both of these methods of deriving probabilities (or weights, as it were) rely on the presence of a full corpus. The first group, which calculates standard Bayesian probabilities, presumably do not do so because they intend to claim that children maintain an internal record of everything they have heard, against which they calculate probabilities. They rather hypothesize (reasonably, I might add) that simple frequency calculations will be usefully proportional to whatever is actually learned.

In the case of Hayes and Wilson, there is an explicit claim that “the learner has access to a large and representative set of observed forms drawn from the target language.” [20, p. 385] Subsequent descriptions of the training make clear that each iteration of the gradient ascent requires—or at least greatly benefits from—a comparably large corpus, whereas I would argue that it is unreasonable to expect a child in critical acquisition stages to have more than a couple of sentences on hand at any given time.

2.3.4. An Allegory. Picture a large barn with a sod floor and a leaky roof. Every time it rains, water drips onto the floor through the holes in the roof; sunlight also shines through the holes when it is sunny. As the barn falls into disuse, plants start to sprout in the patches of sod that lie directly beneath holes in the roof. They grow first, and most densely, in the areas beneath the largest holes: these are the places that see the most water, and the most consistent sun. With time, shoots come up where there is even a small drip. After long enough, the foliage on the floor presents a pattern that matches the pattern of holes in the roof.

This is approximately how I conceptualize human phonotactic learning: rainfall and sunshine are stimuli, and the holes in the roof represent the true, observed phonotactic
grammar. Each plant is some feature value: those that grow highest are the ones that receive the most stimuli, which is to say the most strongly attested.

2.3.5. Stochastic Gradient Descent. Stochastic gradient descent is a simple but powerful algorithm for learning parameter weights. The parameter vector is iteratively compared to feature vectors representing single, randomly selected instances from a training set, and the values of active features are incremented proportionally to the difference between the observed feature values in the training instance and the current, hypothesized values in the parameter vector. [9] The following equation represents the incrementing of a single weight

$$\lambda_k = \lambda_k + \gamma(t_k - \lambda_k)$$

where $\lambda$ is the full parameter vector, $t$ is the vectorized training instance and $\gamma$ is a “learning rate.” The learning rate is traditionally set to a very small decimal value like 0.0001 to ensure that parameter weights do not move too far in response to outliers. The full algorithm is characterized in Algorithm 1.

**Algorithm 1:** Stochastic Gradient Descent

```plaintext
for $t \in \text{training data}$ do
  for $1 \leq k \leq n$ do
    $\lambda_k \leftarrow \lambda_k + \gamma(t_k - \lambda_k)$
  end
end
```

Stochastic gradient descent is very intuitively appropriate for modeling phonotactic learning due to its the natural analogy it makes to the presentation of speech stimuli to a child. Every word that the child hears is one step in the loop, and represents one increment to the parameter vector. The parameters are analogous to probabilities in other learning frameworks.
Another appealing quality of this learning model is that it is very passive: the child is not actively searching for the best set of weights, as appears to be implied by the gradient ascent in [20]. The adjustment of parameter weights is such a simple calculation that it is easy to construe as a subconscious response.

2.3.6. Differences from Stochastic Gradient Descent. Due to the nature of phonotactic learning and the constraints on the problem set out in previous sections, we will make several adjustments to stochastic gradient descent as standardly used:

(1) The parameter vector $\lambda$ should be initialized to contain all zeros, indicating that a child starts off not accepting any sequences in its grammar (but is capable of learning to, of course). In other applications, it is common to assign randomly chosen initial weights.

(2) Similarly, when vectorizing a training instance, all active features should be assigned values of 1—which is to say that the occurrence of a feature in a word is a single data point suggesting that that feature is acceptable. This generalization, combined with the previous point, considerably restricts the calculations; it guarantees that all feature values will remain between 0 and 1, and further means that adjusting the feature weights will mean increasing them in every case (albeit by different magnitudes depending on the current weight).

(3) Gradient descent algorithms generally terminate when they have reached a gradient of 0, i.e. when feature values cannot improve anymore. In the case of stochastic gradient descent, common practice is to measure performance of the model on an outside test dataset, and exit the loop if too many iterations pass without an improvement. (This avoids overfitting.) In the case of learning phonotactics, the concept of unrepresentative training data does not apply; so overfitting is not a concern. Moreover, on principle the training should go on as long as there are linguistic
stimuli, unless we want to define some arbitrary point at which a human’s linguistic knowledge somehow freezes.

(4) We should, however, acknowledge, that linguistic intuitions are at their most malleable at young ages, and that adults adapt far more slowly. It would therefore make sense to implement a learning rate that slowly declines proportionally to the number of stimuli experienced. This would allow a model to become very stable past a certain point despite continued exposure to stimuli that necessarily raise parameter values.

Taken together, these adjustments result in a learning algorithm that is very simple, but intuitively appears to emulate the most important characteristics of phonotactic acquisition.

2.3.7. A Note About Formal Language Theory. Phonological grammars are generally agreed to fit within the class of regular languages, i.e. those that can be represented with finite state machines and do not require sensitivity to non-adjacent states. Moreover, we observe that many regular languages exhibit behavior not seen in phonology (requiring certain phones to occur an even number of times, for example), so we can classify phonological grammars as subregular. [21]

Heinz discusses (in [22]) inductive learning algorithms for a number of subregular grammatical formalisms and their relevance to phonology. In each case, the learner returns a finite-state acceptor that ideally captures the patterns in the positive data presented to it.

I have not, so far, described the phonotactics learned by the system laid out in this chapter as a formal grammar, and this is partly by design. Consider the following points:

(1) A critical assumption of this framework is that phonotactic acceptability judgments are gradient rather than binary. As such, a finite-state machine representing acceptability judgments would need transition weights, and would need to output a numerical value determined by the path traversed.
(2) The \( n \)-gram languages are subregular, and so can be compiled to finite-state machines. Heinz discusses [22, p. 83–87] strategies for generalizing about \( n \)-gram languages to make more compact acceptors by merging states together. The presence of transition probabilities makes state-merging necessarily more complicated (because perfect overlap between \( n \)-grams will be less common), but heuristics could be adopted to mitigate this.

(3) Because all of the features I propose to include in a phonotactic model are \( n \)-gram features, the model effectively defines a regular language in which each string is returned with a numerical acceptability value.

(4) I propose stochastic gradient descent solely as a way of learning feature weights (from which transition probabilities in a finite-state machine could be derived); this should not affect strategies for building the acceptor itself.

(5) Stochastic gradient descent is equally capable of learning any set of feature weights. Recall the discussion in 2.2.4 of potential alternate features to include in the feature set; as long as these features represent regular expressions over a word, the status of the implicit grammar as representing a regular language is not affected.

Because all of the features I can reasonably see using are regular and therefore encodable with finite-state machines, I have not gone out of my way to address the formal-language implications of the various choices I make.

2.4. Local Summary

In this section, I have proposed a theoretical approach to modeling phonotactics that aims to emulate the learning of phonotactic intuitions as well as the intuitions themselves.

In 2.1, I argue that the phonotactic acceptability of a sequence of sounds should be the minimum of a series of judgments about subsequences, in contrast to other proposals that it be a sum of product of such judgments.
In 2.2, I propose that the separately judged subsequences of the word should be \( n \)-phones for \( 1 \leq n \leq 4 \) of two sequences: the full word including syllable boundaries, and only the vowels of the word (to encode suprasegmental features).

In 2.3, I propose learning feature weights using stochastic gradient descent, a very simple algorithm that allows for learning based only on positive data and in a manner consistent with children's exposure to language stimuli.

In the next chapter, I describe implementing such a model to learn English phonotactics, and the results obtained.
CHAPTER 3

Results

In the previous chapter, I described a model that differs from previous ones in a number of ways, namely:

- It harshly restricts the model’s access to stimulus information.
- The training consists only of very simple arithmetic, and assumes no further calculations on the part of the learner except extracting features from words.
- It learns a moderate set of features that are essentially analogous to frequencies, and are not weighted relative to each other.

In light of these points, the implicit hypothesis is that phonotactics can be learned in such a way: not relying on active probability calculations or iteratively training against a large corpus, not using complicated features or complicated math. As I will discuss in 3.3, I find notable support for these ideas with an English model, but testing reveals aspects of the hypothesis that require closer examination.

3.1. The English Model

Given the model described in the previous chapter, the following information is needed to create a phonotactic model of a language:

1. A corpus of text in the language
2. A function to convert words in the language to sequences of phones with syllable boundaries, whether a pronunciation dictionary or an actual mapping of graphemes to phones
(3) A mapping of phones to distinctive features

The corpus can be arbitrarily large; larger corpora will obviously provide more representative data, but will take longer to train. For practical purposes, it may be advisable to use a higher learning rate for small corpora, and for very large corpora one that declines on a schedule.

To test this model framework’s ability to capture phonotactic patterns, and its performance at predicting phonotactic acceptability judgments, I implemented a full model of American English phonotactics.

3.1.1. The Corpus. The Brown Corpus is a freely available corpus of American English texts in a variety of genres, compiled from texts published in 1961. It totals about 1.1 million tokens, including punctuation. Research shows that children hear millions of words every year, so this corpus is equivalent to at best a few months of true exposure—but should nevertheless be sufficient to make significant generalizations about the phonotactic structure of English. The corpus was accessed through the Natural Language Toolkit.

3.1.2. The Lexicon. Words in the Brown Corpus were matched with pronunciations in the C.M.U. Pronouncing Dictionary, a freely available phonetic dictionary with pronunciations for about 140,000 English words and proper nouns, representing fairly dialect-neutral American English.

The following adjustments were made to the original Arpabet transcriptions of words.

(1) ER (the nurse vowel) $\rightarrow$ AH R /ɔɪ/

(2) All diphthongs and long/tense vowels were converted to sequences of a lax vowel and a glide:¹

¹Phonetic diphthongs were originally separated to make assigning distinctive features more straightforward—is one to assign both +high and +low to the MOUTH vowel?—and then the pattern was extended to other tense vowels because diphthongs and tense vowels appear to form a natural class (e.g. not appearing before velar nasals), which would be captured more easily with consistent encoding.
CHAPTER 3. RESULTS

- \textit{AW} (MOUTH) $\rightarrow$ AA W /aw/
- \textit{AY} (PRICE) $\rightarrow$ AA Y /aj/
- \textit{EY} (FACE) $\rightarrow$ EH Y /ɛj/
- \textit{IY} (FLEECE) $\rightarrow$ IH Y /iʃ/
- \textit{OW} (GOAT) $\rightarrow$ AO W /ɔw/
- \textit{OY} (CHOICE) $\rightarrow$ AO Y /ɔj/
- \textit{UW} (GOOSE) $\rightarrow$ UH W /uʃ/

(3) Syllable boundary symbols (.) were added between syllables and at the ends of words. Syllabification followed the Maximum Onset Principle.\(^2\)

(4) Primary and secondary stress were merged, so AH2 became AH1, and so on.\(^3\)

3.1.3. Feature Set. Phones were featurized according to an underspecified, minimalist S.P.E.-style distinctive feature set, following Hayes and Wilson for consonant features and Stemberger (1992) for vowels.\(^4\) \([18, 20, 27]\) See Table 1 for consonant features and Table 2 for vowel features. Vowels are encoded as +syll, but are underspecified for ±cons, ±approx and ±son. (It was found that including these features caused too much confusion in the model between vowels and other sonorants.)

Following the Arpabet transcription standard already used by C.M.U., \([9]\) AH0 was considered the unstressed allophone of /ʌ/ AH1. In addition to the four features shown in Table 2, stress was added as a feature of the vowel, fully specified (so stressed vowels were encoded

---

\(^2\)As a sole exception, glides were considered strongly attached to preceding vowels, because it was found that the model did not respond well to V.{jw}V sequences.
\(^3\)Secondary stress’s phonemic status in English is dubious, and in any case it is marked so inconsistently in is dictionary that I judged the system would learn better with only one stress level. (This approach also lent itself well to a ±stress feature.)
\(^4\)I depart from Stemberger’s radically underspecified vowel system only to include –back as a specified feature for /æʃt/. Complete underspecification of /ɛ/—which is admittedly the main thrust of the paper—would mean no restriction applies to /ɛ/ that does not apply to all other vowels, and that front vowels could not be a natural class. I was skeptical enough to play it safe.
CHAPTER 3. RESULTS

as +stress, and unstressed vowels as –stress). The syllable boundary symbol was given only one feature: itself.

Using the features in Tables 1 and 2, I calculated all of the natural classes that exist over English phonemes according to this feature set. There are 3248, which I reduced to 132 by collapsing together those classes that are differently specified but have the same membership. (In each such redundant case, the specification for the class was determined by sorting the candidates first by number of features, then alphabetically.)

Table 3 shows the minimal natural classes to which /m/ belongs.

Table 1. Consonant features used in the English model

|     | p | t | tf | k | b | d | dʒ | g | f | θ | s | j | h | v | ʒ | z | m | n | η | l | i | j | w |
| syll| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| cons| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| approx| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| son| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| cont| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| nas| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| voice| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| spread| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| lab| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| cor| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| ant| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| strid| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| lat| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| dors| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| high| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |
| back| + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + | + |

Table 2. Vowel features used in the English model

<table>
<thead>
<tr>
<th></th>
<th>i</th>
<th>e</th>
<th>æ</th>
<th>ʌ</th>
<th>ɔ</th>
<th>ʊ</th>
</tr>
</thead>
<tbody>
<tr>
<td>syll</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>high</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>low</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>back</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>round</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
3.1.4. Training Conditions. The model was trained with a learning rate of 0.001. Given the size of the corpus, I did not deem it necessary to implement a decline in learning rate.

3.2. Results

In this section, I will report on a number of tests of the system described in 3.1. These tests fall into three basic categories: demonstrations of the model’s success in learning specific aspects of English phonotactics; examples of scores for representative words; and correlations with actual people’s phonotactic acceptability judgments, both from my own trial and from those reported in previous literature.

3.2.1. Learning English Onset Clusters. English consonant clusters are a good first test for a variety of reasons:

- They are phonotactically interesting.
- There are many logically possible clusters (the number of consonants cubed, assuming we limit ourselves to three consonants—as English does), and relatively few that are attested.
- There is clear gradience within the acceptability of unattested clusters.

<table>
<thead>
<tr>
<th>Natural Class</th>
<th>Membership</th>
</tr>
</thead>
<tbody>
<tr>
<td>-syll</td>
<td>{p, t, ts, k, b, d, dʒ, g, f, θ, s, ʃ, h, v, ŋ, z, ʒ, m, n, ŋ, l, ɹ, j, w}</td>
</tr>
<tr>
<td>+cons</td>
<td>{p, t, ts, k, b, d, dʒ, g, f, θ, s, ʃ, h, v, ŋ, z, ʒ, m, n, ŋ, l}</td>
</tr>
<tr>
<td>-approx</td>
<td>{p, t, ts, k, b, d, dʒ, g, f, θ, s, ʃ, h, v, ŋ, z, ʒ, m, n, ŋ, l}</td>
</tr>
<tr>
<td>+son</td>
<td>{m, n, ŋ, l, ɹ, j, w}</td>
</tr>
<tr>
<td>+son +cons</td>
<td>{m, n, ŋ, l}</td>
</tr>
<tr>
<td>+nas</td>
<td>{m, n, ŋ}</td>
</tr>
<tr>
<td>+lab</td>
<td>{p, b, f, v, m, w}</td>
</tr>
<tr>
<td>+lab +cons</td>
<td>{p, b, f, v, m}</td>
</tr>
<tr>
<td>+lab +son</td>
<td>{m, w}</td>
</tr>
<tr>
<td>+lab +nas</td>
<td>{m}</td>
</tr>
</tbody>
</table>

Table 3. Minimal natural classes containing /m/
• Several researchers have published their results on this task, and some data on native speakers’ acceptability judgments is available in the literature.

I calculated word-initial acceptability scores from the model first for all groupings of 1 and 2 consonants, excluding those ending with /j/. (/j/ is often parsed as part of the rhyme, rather than the onset, as in [13].) When ranked, the first 52 contained only two prohibited onsets (the not-so-infelicitous /hl/ and /hô/) and two marginally attested (/ʒ/, /sô/). That leaves 49, which represent all but one permitted onset (missing /ʃ¹/ , which it ranks at #86). See the top 100 in Table 4.

Clearly, the model is learning something right. Clearly, also, it is not perfect: English words starting with the marginal /sô/ (e.g. Sri Lanka), which the model thought was in the acceptable range, are often repaired to /ʃ¹/—which it ranked after /θm/ and before /dl/.

Interestingly, this does represent an exception of sorts: /ʃ/ is otherwise prohibited in onsets, except in marginal words like shlep and shtick.

I did a similar exercise with all 13,824 possible sequences of three consonants. The six permitted clusters—/stô/, /skô/, /spô/, /spl/, /skl/ and /skw/—were all in the top nine highest-ranked. Three illegal clusters were ranked above the worst legal one, all starting with /h/. It seems the model learned very effectively that voiceless fricatives are good for starting onsets—48 of the 50 highest-ranked triplets begin with one of /θ sʃ h/—and the lack of evidence for /h/, pretty much all relying on the +spread feature, was not enough to counterbalance.

Data from Scholes (1966) [26] provides a good test for acceptability of onsets. Scholes presented a class of seventh-graders with a number of monosyllabic nonce words, and asked for binary judgments on whether each could be a word. The words all had common rhymes, but the onsets were a deliberate mix of attested and unattested clusters. Words included /fkîp/ (3 “yes” votes out of 33), /klûn/ (33 of 33) and /sîn/ (15 of 33).
### Table 4. The model’s 100 most-acceptable onsets of 1 and 2 consonants. Bold-faced clusters are substantially attested in English.

<table>
<thead>
<tr>
<th>#</th>
<th>Cluster Score</th>
<th>#</th>
<th>Cluster Score</th>
<th>#</th>
<th>Cluster Score</th>
<th>#</th>
<th>Cluster Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>/p/</td>
<td>1.000</td>
<td>26</td>
<td>/f/</td>
<td>0.986</td>
<td>51</td>
<td>/j/</td>
</tr>
<tr>
<td>2</td>
<td>/w/</td>
<td>1.000</td>
<td>27</td>
<td>/d/</td>
<td>0.981</td>
<td>52</td>
<td>/t/</td>
</tr>
<tr>
<td>3</td>
<td>/st/</td>
<td>1.000</td>
<td>28</td>
<td>/sk/</td>
<td>0.981</td>
<td>53</td>
<td>/ŋ/</td>
</tr>
<tr>
<td>4</td>
<td>/n/</td>
<td>1.000</td>
<td>29</td>
<td>/fl/</td>
<td>0.980</td>
<td>54</td>
<td>/ŋ/</td>
</tr>
<tr>
<td>5</td>
<td>/k/</td>
<td>1.000</td>
<td>30</td>
<td>/ʃ/</td>
<td>0.980</td>
<td>55</td>
<td>/ŋ/</td>
</tr>
<tr>
<td>6</td>
<td>/b/</td>
<td>1.000</td>
<td>31</td>
<td>/ð/</td>
<td>0.977</td>
<td>56</td>
<td>/n/</td>
</tr>
<tr>
<td>7</td>
<td>/f/</td>
<td>1.000</td>
<td>32</td>
<td>/v/</td>
<td>0.976</td>
<td>57</td>
<td>/v/</td>
</tr>
<tr>
<td>8</td>
<td>/m/</td>
<td>1.000</td>
<td>33</td>
<td>/θ/</td>
<td>0.976</td>
<td>58</td>
<td>/p/</td>
</tr>
<tr>
<td>9</td>
<td>/l/</td>
<td>1.000</td>
<td>34</td>
<td>/kw/</td>
<td>0.975</td>
<td>59</td>
<td>/sw/</td>
</tr>
<tr>
<td>10</td>
<td>/h/</td>
<td>1.000</td>
<td>35</td>
<td>/bl/</td>
<td>0.973</td>
<td>60</td>
<td>/s/</td>
</tr>
<tr>
<td>11</td>
<td>/p/</td>
<td>1.000</td>
<td>36</td>
<td>/ð/</td>
<td>0.962</td>
<td>61</td>
<td>/š/</td>
</tr>
<tr>
<td>12</td>
<td>/t/</td>
<td>0.999</td>
<td>37</td>
<td>/s/</td>
<td>0.960</td>
<td>62</td>
<td>/hm/</td>
</tr>
<tr>
<td>13</td>
<td>/t/</td>
<td>0.999</td>
<td>38</td>
<td>/s/</td>
<td>0.958</td>
<td>63</td>
<td>/hn/</td>
</tr>
<tr>
<td>14</td>
<td>/t/</td>
<td>0.999</td>
<td>39</td>
<td>/s/</td>
<td>0.932</td>
<td>64</td>
<td>/sj/</td>
</tr>
<tr>
<td>15</td>
<td>/g/</td>
<td>0.999</td>
<td>40</td>
<td>/s/</td>
<td>0.931</td>
<td>65</td>
<td>/vl/</td>
</tr>
<tr>
<td>16</td>
<td>/k/</td>
<td>0.998</td>
<td>41</td>
<td>/s/</td>
<td>0.928</td>
<td>66</td>
<td>/dw/</td>
</tr>
<tr>
<td>17</td>
<td>/p/</td>
<td>0.998</td>
<td>42</td>
<td>/s/</td>
<td>0.923</td>
<td>67</td>
<td>/tl/</td>
</tr>
<tr>
<td>18</td>
<td>/d/</td>
<td>0.998</td>
<td>43</td>
<td>/hw/</td>
<td>0.916</td>
<td>68</td>
<td>/bw/</td>
</tr>
<tr>
<td>19</td>
<td>/b/</td>
<td>0.998</td>
<td>44</td>
<td>/z/</td>
<td>0.900</td>
<td>69</td>
<td>/fw/</td>
</tr>
<tr>
<td>20</td>
<td>/z/</td>
<td>0.998</td>
<td>45</td>
<td>/h/</td>
<td>0.897</td>
<td>70</td>
<td>/fn/</td>
</tr>
<tr>
<td>21</td>
<td>/s/</td>
<td>0.997</td>
<td>46</td>
<td>/tw/</td>
<td>0.895</td>
<td>71</td>
<td>/kn/</td>
</tr>
<tr>
<td>22</td>
<td>/kl/</td>
<td>0.996</td>
<td>47</td>
<td>/hl/</td>
<td>0.885</td>
<td>72</td>
<td>/ðj/</td>
</tr>
<tr>
<td>23</td>
<td>/g/</td>
<td>0.993</td>
<td>48</td>
<td>/gw/</td>
<td>0.883</td>
<td>73</td>
<td>/ft/</td>
</tr>
<tr>
<td>24</td>
<td>/sp/</td>
<td>0.990</td>
<td>49</td>
<td>/θw/</td>
<td>0.881</td>
<td>74</td>
<td>/fj/</td>
</tr>
<tr>
<td>25</td>
<td>/θ/</td>
<td>0.988</td>
<td>50</td>
<td>/ʒ/</td>
<td>0.881</td>
<td>75</td>
<td>/fp/</td>
</tr>
</tbody>
</table>

Figure 3.1 shows a scatterplot comparing the model’s prediction (on the x-axis) with Scholess recorded observation (the number of children accepting, on the y-axis). The correlation is high, with $r = 0.862$, and statistically significant with $p < 0.00001$.

Hayes and Wilson report significantly better performance on this dataset ($r = 0.946$). Albright trained a natural-class-based diphone model on the same data, whose correlation he measured with Kendall’s $\tau_c$: 0.602, less than Hayes and Wilson’s $\tau_c = 0.702$. [1] The
comparable $\tau_c$ figure for the present model is 0.610: a slight improvement on Albright’s model, but not nearly as good as Hayes and Wilson’s.

The models are not quite parallel. Not only are the feature sets different, but both past models trained exclusively on onset data from the C.M.U. Dictionary, and excluded what they considered “exotic” onsets (*i.e.* marginally attested, like /sf/ and /pw/). As such, the judgments they gave were solely for the onsets of Schole’s test words. I built my model using data for the whole language—which both previous papers had claimed resulted in a slight decrease in accuracy for them—and my model’s judgments represent full words.

Also importantly, both past models were trained on type frequencies (number of lexical items containing a sequence), rather than token frequencies (number of occurrences of a sequence in a corpus), claiming that this also gave slightly better performance. My model is not trained on frequencies, per se, but it learns analogues to token frequency. I trained an alternate version that did not adjust its weights except upon seeing new words, and contrary
CHAPTER 3. RESULTS

to their prediction found no significant difference in performance. \( r \) changed by less than 0.02.)

3.2.2. Predicting Wordlikeness Judgments. To further test the abilities of his feature-based diphone framework, Albright trains a general model to predict wordlikeness judgments. His test set is 88 judgments of monosyllabic nonce words collected for a previous study, not focusing on onsets so much as a general distribution across levels of phonotactic acceptability. [3] Subjects were asked to rate words on a scale from 0 to 6.

This time, his model somewhat outperformed mine. Albright achieved \( \tau_c = 0.453 \); my model gave \( \tau_c = 0.372, r = 0.599 \). One possible reason is the size of the corpus he used: he trained on frequencies in CELEX, which represent the then-18-million-word Cobuild corpus. It is conceivable that having effectively an order of magnitude more training data (with the resulting larger vocabulary and more precise frequency information) could produce such a difference. I trained an alternate model using only diphone features to test whether the feature set had an impact, and found that the smaller feature set performed slightly less well on this set \( (r = 0.5711) \). Figure 3.2 shows that the data here are less neatly judged than in 3.2.1.

Nevertheless, the correlation is highly significant \( (p < 0.00001 \) with both diphone and triphone features). Moreover, Albright picks out 15 words from the set that contain clear or borderline phonotactic violations, and my model ranks all but two of them among the 25 worst (out of 88). Those words are shown in Table 5.

3.2.3. Predicting Repetition Accuracy. Another useful (if small) dataset to work with is from Brown and Hildum (1956), which asked test subjects to listen to nonce words and then try to spell them, with the trial overseer clarifying ambiguous spellings by soliciting spoken pronunciations. [10] The set consists of 10 legal sequences and 10 illegal sequences,
Figure 3.2. Model (x-axis) performance on data from Albright and Hayes (2003)

<table>
<thead>
<tr>
<th>Word</th>
<th>Observed</th>
<th>Predicted</th>
<th>Model Rank</th>
<th>Violation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jwuz</td>
<td>2.68</td>
<td>0.286</td>
<td>88</td>
<td>*[JC]</td>
</tr>
<tr>
<td>plbomf</td>
<td>2.42</td>
<td>0.380</td>
<td>85</td>
<td>*tense vowel + non-coronal cluster</td>
</tr>
<tr>
<td>dwsod3</td>
<td>2.29</td>
<td>0.401</td>
<td>83</td>
<td>rhyme [sod3] unattested</td>
</tr>
<tr>
<td>pwaqdz</td>
<td>2.89</td>
<td>0.475</td>
<td>80</td>
<td>labial dissimilation</td>
</tr>
<tr>
<td>fulq</td>
<td>2.68</td>
<td>0.589</td>
<td>77</td>
<td>[lg#] unattested</td>
</tr>
<tr>
<td>klnq</td>
<td>2.58</td>
<td>0.601</td>
<td>75</td>
<td>[lg#] unattested</td>
</tr>
<tr>
<td>smelrq</td>
<td>2.58</td>
<td>0.607</td>
<td>73</td>
<td>*tense vowel + non-coronal cluster</td>
</tr>
<tr>
<td>spuuf</td>
<td>2.05</td>
<td>0.611</td>
<td>72</td>
<td>*[CrVr]</td>
</tr>
<tr>
<td>smiilθ</td>
<td>2.47</td>
<td>0.625</td>
<td>70</td>
<td>tense vowel + [Cθ] unattested</td>
</tr>
<tr>
<td>θwiiks</td>
<td>2.59</td>
<td>0.651</td>
<td>67</td>
<td>*tense vowel + non-coronal cluster</td>
</tr>
<tr>
<td>smimθ</td>
<td>2.06</td>
<td>0.662</td>
<td>65</td>
<td>tense vowel + [Cθ] unattested</td>
</tr>
<tr>
<td>snoks</td>
<td>3.00</td>
<td>0.679</td>
<td>64</td>
<td>*tense vowel + non-coronal cluster</td>
</tr>
<tr>
<td>θiipt</td>
<td>2.26</td>
<td>0.775</td>
<td>53</td>
<td>*tense vowel + non-coronal cluster</td>
</tr>
<tr>
<td>tnilb</td>
<td>2.63</td>
<td>0.553</td>
<td>78</td>
<td>[lb#] very rare</td>
</tr>
<tr>
<td>plbomnθ</td>
<td>2.26</td>
<td>0.815</td>
<td>49</td>
<td>tense vowel + [Cθ] unattested</td>
</tr>
</tbody>
</table>

Table 5. Words picked out from data in Albright and Hayes (2003) as having concrete phonotactic problems, displayed with comparisons between respondents’ wordlikeness judgments and the predictions of the current English phonotactic model.
and establishes (among other things) that people are more likely to mishear or incorrectly reproduce phonotactically illegal words.\(^5\)

As in other studies, the samples were all monosyllabic.\(^6\) My model almost perfectly ranked all “English non-words” above all “Non-English syllables.” The exceptions were /sp1b/, which was classified as a non-word (though a pretty strange one); and /sɛ̃it/, which got very good ratings because—as established above—the model really likes /s1/ as an onset. Overall, \(r\) for the 20 words was 0.717 \((p < 0.001)\), which rose to 0.762 \((p < 0.0001)\) when the independent variable was fit to a log scale. I do not have a good explanation of why it would be appropriate to predict a logarithm of these judgments, but not of the others (which are better as linear correlations), except that repetition accuracy is a different task than wordlikeness judgments, and perhaps the data behaves differently.

3.3. Discussion

The three datasets discussed in 3.2 provide ample support for the claim that the phonotactic model described in this thesis successfully learns a substantial amount of phonotactic information. It is somewhat outperformed by other models, but the degree of correlation is notable nonetheless given the constraints under which it operates: a moderate feature set, very simple math, relatively little data and tightly controlled access to the data that is available.

That said, in examining the errors the model makes, one thing is clear: it is very good at generalizing, but there will probably always be some cases where it overgeneralizes, and

\(^5\)The study consisted of separate trials of students with and without experience in linguistics, and found the former group considerably better at recognizing illegal sequences. I chose to predict the data from the naïve group.

\(^6\)Edwards et al. (2004) do include wordlikeness data for polysyllabic nonce words. \([16]\) I did not attempt to predict them with my model because their transcriptions did not include stress information—a key determiner of phonotactic acceptability—and many words therefore had multiple plausible pronunciations.
CHAPTER 3. RESULTS

it cannot establish a sufficiently wide range of feature weights to make specific exceptions when it needs to.

The mistakenly accepted onset clusters give abundant evidence for this. The model learned accurate generalizations that /s/ is acceptable in onset clusters but /ʃ/ is not: /ʃ/ is a member of 23 natural classes, five of which are defined in part by being –ant, which distinguishes it from /s/. The low values learned for features involving these items are sufficient to kill acceptability scores for any cluster containing /ʃ/. For a model to learn that /ʃ/ ≻ /s/, but /sl/ ≻ /ʃl/, it must be able to assign a very high weight to a few specific features, which this model cannot.

/h/ in onsets is a reversed example. /h/ is only in nine natural classes, and eight of them—+cont; –son; –approx; +cont, –voice; –syll; +cons; and –voice—all represent classes of sounds that work very well in onset clusters. Indeed, /s/ is a member of all eight. Although the remaining natural class, +spread, is unattested in clusters, that is not enough to significantly pull down the very high values of the rest. In other words, because /h/ is so underspecified, this model is almost incapable of learning rules that affect it alone, without other fricatives.

Hayes and Wilson’s model addresses this problem by learning specific constraints that are quite highly rated for this purpose. One of their constraints penalizes /h/ in clusters except before /w/; and another, phrased “*[+ant, +strid][–ant],” expressly penalizes /s/. [20, p. 397–398]

While I am not convinced that Hayes and Wilson’s model is a realistic portrayal of child learning (for reasons already stated), this advantage is clear, and future incarnations of the model under discussion should incorporate changes either to the feature structure or to the learning process in order to allow the model more freedom in assigning weights.
(Note, however, that with \textit{no weights at all}, this model was able to come reasonably close to Hayes and Wilson’s, which may suggest that the necessary changes are not so drastic.)
CHAPTER 4

Conclusion

In the preceding chapters, I have established and justified a novel approach to modeling phonotactics. I have demonstrated that learning phonotactics is possible despite restricting an algorithm to simple, passive calculations on one word at a time, and exposing it to only as many words as the average infant hears in its first month. I have not explored the psycholinguistic ramifications of the latter part of the claim, but it suggests that phonotactic awareness could be very far along at an earlier age than has been substantially tested in the literature.

I have also demonstrated that many of the model’s performance issues can be traced to a single weakness—a restriction on the range of weights that the algorithm can assign—which suggests the above claim is even more solid because the quantity of data is not the limiting factor.

4.1. Future Directions

4.1.1. Phonotactic Distance Between Languages. One outcome of the algorithm that I have not explored is the fact that it outputs a vector of weights corresponding to features whose identities are postulated to be language-universal. What this means is that, if the program were to be run on a variety of different languages, it would be fairly straightforward to calculate a Euclidean distance between the vectors for any two languages. Many features would exist in some languages but not others, so we would assume that those features had values of 0 wherever they were absent.
CHAPTER 4. CONCLUSION

Calculating such “phonotactic distance scores” has very interesting implications for linguistic typology. Many generalizations can be drawn about basic phonetic or phonological behaviors across languages, but this score would be uniquely positioned as at least a decent proxy for an aggregate comparator.

Even language-internally, the vectors output by this algorithm have a lot of potential to be theoretically interesting. The nuance that they claim to capture may go beyond affirming the existence of known phonotactic constraints, and actually expose new ones as well.

4.1.2. Using Phonotactic Scores to Motivate Phonology. Optimality Theory has perhaps been the elephant in the room through this whole discussion. [24] The more one studies phonotactics, the more one suspects that, to a large extent, phonotactics is a measure of the felicitousness of a sound sequence, and phonology is the tool for repairing sequences that do not meet a certain threshold. Given a phonotactic parameter vector like what is output by this function, one could write a simple program to propose changes to ill-formed words that improve their acceptability while staying within a short edit distance (and perhaps satisfying other faithfulness constraints). The devil is in the details, but the prospect is encouraging.
Bibliography

2. _____, *Feature-based generalisation as a source of gradient acceptability*, Phonology 26 (2009), no. 01, 9–41.


