A Bayesian model of natural language phonology: generating alternations from underlying forms

David Ellis
de@cs.brown.edu
Brown University
Providence, RI 02912

Abstract

A stochastic approach to learning phonology. The model presented captures 7-15% more phonologically plausible underlying forms than a simple majority solution, because it prefers "pure" alternations. It could be useful in cases where an approximate solution is needed, or as a seed for more complex models. A similar process could be involved in some stages of child language acquisition; in particular, early learning of phonotactics.

1 Introduction

Sound changes in natural language, such as stem variation in inflected forms, can be described as phonological processes. These are governed by a constraint hierarchy as in Optimality Theory (OT), or by a set of ordered rules. Both rely on a single lexical representation of each morpheme (i.e., its underlying form), and context-sensitive transformations to surface forms. Phonological changes often affect segments near morpheme boundaries, but can also apply over an entire prosodic word, as in vowel harmony.

It does not seem straightforward to incorporate context into a Bayesian model of phonology, although a clever solution may yet be found. A standard way of incorporating conditioning environments is to treat them as factors in a Gibbs model (Liang and Klein, 2007), but such models require an explicit calculation of the partition function. Unless the rule contexts possess some kind of locality, we don't know how to compute this partition function efficiently. Some context could be captured by generating underlying phonemes from an n-gram model, or by annotating surface forms with neighborhood features. However, the effects of autosegmental phonology and other long-range dependencies (like vowel harmony) cannot be easily Bayesianized.

1.1 Related Work

In the last decade, finite-state approaches to phonology (Gildea and Jurafsky, 1996; Beesley and Karttunen, 2000) have effectively brought theoretical linguistic work on rewrite rules into the computational realm. A finite-state approximation of optimality theory (Karttunen, 1998) was later refined into a compact treatment of gradient constraints (Gerdemann and van Noord, 2000).

Recent work on Bayesian models of morphological segmentation (Johnson et al., 2007) could be combined with phonological rule induction (Goldwater and Johnson, 2004) in a variety of ways, some of which will be explored in our discussion of future work. Also, the Hierarchical Bayes Compiler (Daume III, 2007) could be used to generate a model similar to the one presented here, but less constrained[1] which makes correspondingly more random, less accurate predictions.

1.2 Dataset

As we describe the model and its implementation in this and subsequent sections, we will refer to a sam-

[1] Recent updates to HBC, inspired by discussions with the author, have addressed some of these limitations.

Proceedings of the Tenth Meeting of the ACL Special Interest Group on Computational Morphology and Phonology, pages 12–19,
Columbus, Ohio, USA June 2008.

ple dataset (in Figure 1), consisting of a paradigm[2] of verb stems and person/number suffixes. The head of each row or column is an /underlying/ form, which in 3rd person singular is a phonologically null segment (represented as /ø/). In [surface] forms, the realization of each morpheme is affected by phonological processes. For example, in the combination of /tietä/ + /vat/, the result is [tietä+vät], where the 3rd person plural /a/ becomes [ä] due to vowel harmony.

1.3 Bayesian Approach

As a baseline model, we select the most frequently occurring allophone as the underlying form. Our goal is to outperform this baseline using a Bayesian model. In other words, what patterns in phonological processes can be inferred with such a statistical model? This simple framework begins learning with the assumption that the underlying forms are faithful to the surface (i.e., without considering markedness or phonotactics).

We model the generation of surface forms from underlying ones on the segmental (character) level. Input is an inflectional paradigm, with tokens of the form `stem+suffix`. Morphology is limited to a single suffix (no agglutination), and is already identified. Each character of an underlying stem or suffix (u_i) generates surface characters (s_{ij}) in an entire row or column of the input.

To capture the phonology of a variety of languages with a single model, we need constraints from linguistically plausible priors (universal grammar). We prefer that underlying characters be preserved in surface forms, especially when there is no alternation. It is also reasonable that there be fewer underlying forms (phonemes) than surface forms (phones, phonetic inventory), to account for allophones. We expect to be able to capture a significant subset of phonological processes using a simple model (only faithfulness constraints).

1.4 Pure Generators

Our model has an advantage over the baseline in its preference for "purity" in underlying forms. Each underlying segment should generate as few distinct surface segments as possible: if it generates non-alternating (identical) segments, it will be less likely to generate an alternation in addition. This means that when two segments alternate, the underlying form should be the one that appears less frequently in other contexts, irrespective of the majority within the alternation.

In the first stem of our Finnish verb conjugation (Figure 1), we see a [t,d] alternation (a case of consonant gradation), as well as unalternating [t]. If we isolate three of the surface forms where /tietä/ is inflected (1st person singular, and 3rd person singular and plural), and consider only the dental segments in the stem of each, we have two underlying segments. Here, we use question marks to indicate unknown underlying segments.

/??/ [dt] [tt] [tt]

In this subset of the data, the reasonable candidate underlying forms are /t/ and /d/. These two compete to explain the observed data (surface forms). The nature of the prior probability distribution determines whether the majority is hypothesized for each underlying form, so /t/ produces both alternating and unalternating surface segments, or /d/ is hypothesized as the source of the alternation (and /t/ remains "pure"). In a Bayesian setting, we impose a sparse prior over underlying forms conditioned on the surface forms they generate.

If u_2 is hypothesized to be /t/, the posterior probability of u_1 being /t/ goes down:

$$P(u_1 = /t/|u_2 = /t/) < P(u_1 = /t/)$$

The probability of u_1 being the competitor, /d/, correspondingly increases:

$$P(u_1 = /d/|u_2 = /t/) > P(u_1 = /d/)$$

Even though the majority in this case would be /t/, the favored candidate for the alternating form was /d/. This happened because of how we defined the model's prior, in combination with the evidence that /t/ (assigned to u_2) generated the sequence of [t]. So selection bias prefers /d/ as the source of an ambiguous segment, leaving /t/ to always generate itself.

A similar effect can occur if there are both unalternating [t]'s and [d]'s on the surface, in addition to the [t,d] alternation. The candidate (/t/ or /d/) that is

[2]The paradigm format lends itself to analysis of word types, but if supplemented with surface counts, can also handle tokens.

Stem \ Suffix	/n/ (1s)	/t/ (2s)	/ø/ (3s)	/mme/ (1p)	/tte/ (2p)	/vat/ (3p)
/tietä/	[tiedä+n]	[tiedä+t]	[tietä+ä]	[tiedä+mme]	[tiedä+tte]	[tietä+vät]
/aiko/	[aiøo+n]	[aiøo+t]	[aiko+o]	[aiøo+mme]	[aiøo+tte]	[aiko+vat]
/luke/	[luøe+n]	[luøe+t]	[luke+e]	[luøe+mme]	[luøe+tte]	[luke+vat]
/puhu/	[puhu+n]	[puhu+t]	[puhu+u]	[puhu+mme]	[puhu+tte]	[puhu+vat]
/saa/	[saa+n]	[saa+t]	[saa+ø]	[saa+mme]	[saa+tte]	[saa+vat]
/tule/	[tule+n]	[tule+t]	[tule+e]	[tule+mme]	[tule+tte]	[tule+vat]
/pelkää/	[pelkää+n]	[pelkää+t]	[pelkää+ø]	[pelkää+mme]	[pelkää+tte]	[pelkää+vät]

Figure 1: Sample dataset (constructed by hand): Finnish verbs, with inflection for person and number.

generating fewer unalternating segments is preferred to explain the alternation. For example, if there were 1000 cases of [t], 500 [d] and 500 [t,d], we would expect the following hypotheses: $/t/ \rightarrow [t]$, $/d/ \rightarrow [d]$ and $/d/ \rightarrow [t, d]$. This is because one of the two candidates must be responsible for both unalternating and alternating segments, but we prefer to have as much "'purity" as possible, to minimize ambiguity.

With this solution, we still have 1000 pure $/t/ \rightarrow [t]$, and only the 500 $/d/ \rightarrow [d]$ are now indistinct from $/d/ \rightarrow [t, d]$. If we had selected /t/ as the source of the alternation, there would be only 500 remaining "pure" (/d/) segments, and 1500 ambiguous $/t/$. Our Bayesian model should prefer the less ambiguous ("purer") solution, given an appropriate prior.

2 Model

We will use boldface to indicate vectors, and subscripts to identify an element from a vector or matrix. The variable $\mathbf{N}(\mathbf{u})$ is a vector of observed counts with the current underlying form hypotheses. The notation we use for a vector $\mathbf{u}$ with one element i removed is $\mathbf{u}_{-i}$, so we can exclude the counts associated with a particular underlying form by indicating that in the parenthesized variable (i.e., $\mathbf{N}(\mathbf{u}_{-4})$ is all the counts except those associated with the fourth underlying form). $N_i(\mathbf{u})$ is the number of times character i is used as an underlying form, and $N_{ij}(\mathbf{u})$ is the number of times character i generated surface character j.

The priors over surface $\mathbf{s}$ and underlying $\mathbf{u}$ segments in Figure 2 are captured by Dirichlet priors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which generate the multinomial distributions $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, respectively (see Figure 3). The prior over underlying form encourages sparse solutions, so $\beta_u < 1$ for all u. The prior over surface form given underlying encourages identity mapping, $/x/ \rightarrow [x]$, so $\alpha_{xx} > 1$, and discourages different segments, $/x/ \rightarrow [y]$, so $\alpha_{xy} < 1$ for all $x \neq y$.

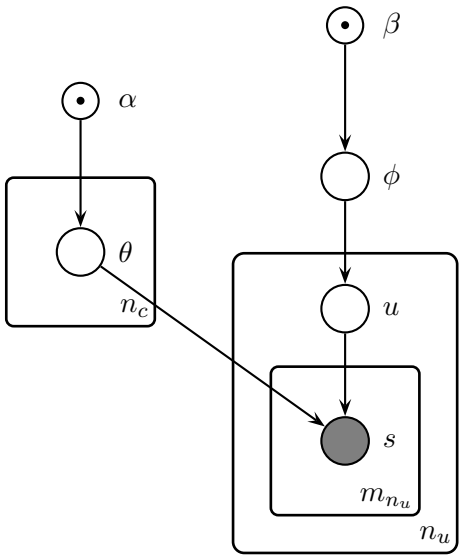

Figure 2: Bayesian network: $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are vectors of hyperparameters, and $\boldsymbol{\theta}_i$ (for $i \in \{1, \ldots, n_c\}$) and $\boldsymbol{\phi}$ are distributions. $\mathbf{u}$ is a vector of underlying forms, generated from $\boldsymbol{\phi}$, and $\mathbf{s}_i$ (for $i \in n_u$) is a set of observed surface forms generated from the hidden variable u_i according to $\boldsymbol{\theta}_i$

Phones and phonemes are drawn from a set of characters (e.g., IPA, unicode) C used to represent them. ϕ_i is the probability of a character (C_i for $i \in n_c$) being an underlying form, irrespective of current alignments or its position in the paradigm. θ_{ij} is the conditional probability of a surface char-

$$
\begin{array}{lcll}
\boldsymbol{\theta}_c & | & \boldsymbol{\alpha} & \sim \text{DIR}(\boldsymbol{\alpha}), c = 1, \ldots, n_c \\
\boldsymbol{\phi} & | & \boldsymbol{\beta} & \sim \text{DIR}(\boldsymbol{\beta}) \\
u_i & | & \boldsymbol{\phi}_i & \sim \text{MULTI}(\boldsymbol{\phi}_i), i = 1, \ldots, n_u \\
s_{ij} & | & u_i, \boldsymbol{\theta}_{u_i} & \sim \text{MULTI}(\boldsymbol{\theta}_{u_i}), i = 1, \ldots, n_u, \\
 & & & \quad j = 1, \ldots, m_i
\end{array}
$$

Figure 3: Model parameters: n_c is # different segments, n_u is # underlying segments

acter ($s_{kn} = C_j$ for $j \in n_c$, $n \in m_k$) given the underlying character it is generated from ($u_k = C_i$ for $i \in n_c$, $k \in n_u$), which is determined by its position in the paradigm.

In our Finnish example (Figure 1), if $k = 1$, we are looking at the first underlying character, which is /t/ (from /tietä/), so assuming our character set is the Finnish alphabet, of which 't' is the 20th character, $u_1 = C_{20} =$ t. It generates the first character of each inflected form (1st, 2nd, 3rd person, singular and plural) of that stem, so $m_1 = 6$, and since there is no alternation $s_{1n} =$ t (for $n \in \{1, \ldots, 6\}$). Given the phonologically plausible (gold) underlying forms, the probability of /t/ is $\phi_{20} = 7/41$.

On the other hand, $k = 33$ identifies the 3rd person singular /ø/, which inflects each of the seven stems, so $m_{33} = 7$. Since we need our alphabet to identify a null character, we'll give it index zero (i.e., $u_{33} = C_0 = \emptyset$). For each of the (underlying, surface) alignments in this alternation (caused by vowel gemination), we can identify the probability in $\boldsymbol{\theta}$. For 3rd person singular [tietä+ä], where $s_{33,1} = C_{28} =$ ä, the conditional probability $\theta_{0,28} = 1/7$.

The prior hyperparameters can be understood as follows. As β_i gets smaller, an underlying form u_k is less likely to be C_i. As α_{ij} gets smaller, an underlying $u_k = C_i$ is less likely to generate a surface segment $s_{kn} = C_j \; \forall n \in m_k$. In our experiments, we will vary $\alpha_{i=j}$ (prior over identity map from underlying to surface) and $\alpha_{i \neq j}$.

Our implementation of this model uses Gibbs sampling (c.f., (Bishop, 2006), pp 542-8), an algorithm that produces samples from the posterior distribution. Each sample is an assignment of the hidden variables, $\mathbf{u}$ (i.e., a set of hypothesized underlying forms). Our sampler initializes $\mathbf{u}$ from a uniform distribution over segments in the training data, and resamples underlying forms in a fixed order, as input in the paradigm. Rather than reestimate $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ at each iteration before sampling from $\mathbf{u}$, we can marginalize these intermediate probability distributions in order to ease implementation and speed convergence.

Our search procedure tries to sample from the posterior probability, according to Bayes' rule.

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

$$P(\mathbf{u}, \mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\alpha}) \propto P(\mathbf{u}|\boldsymbol{\beta}) P(\mathbf{s}, \mathbf{u}|\boldsymbol{\alpha})$$

Each of these probabilities is drawn from a Dirichlet distribution, which is defined in terms of the multivariate Beta function, C. The prior $\boldsymbol{\beta}$ added to underlying counts $\mathbf{N}(\mathbf{u})$ forms the posterior Dirichlet corresponding to $P(\mathbf{u}|\boldsymbol{\beta})$. In $P(\mathbf{s}|\mathbf{u}, \boldsymbol{\alpha})$, each $\boldsymbol{\alpha}_i$ vector is supplemented by the observed counts of (underlying, surface) pairs $N(\mathbf{s}_i)$.

$$P(\mathbf{u}, \mathbf{s}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \frac{C(\boldsymbol{\beta} + N(\mathbf{u}))}{C(\boldsymbol{\beta})} \prod_{c=1}^{n_c} \frac{C(\boldsymbol{\alpha}_c + \sum_{i:u_i=c} N(\mathbf{s}_i))}{C(\boldsymbol{\alpha})}$$

The collapsed update procedure consists of resampling each underlying form, u, incorporating the prior hyperparameters $\boldsymbol{\alpha}, \boldsymbol{\beta}$ and counts N over the rest of the dataset. The relevant counts for a candidate k being the underlying form u_i are $N_k(\mathbf{u}_{-i})$ and $N_{ks_{ij}}(\mathbf{u}_{-i})$ for $j \in m_i$. $P(u_i = k|\mathbf{u}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta})$ is proportional to the probability of generating $u_i = k$, given the other $\mathbf{u}_{-i}$ and all s_{ij} (for $j \in m_i$), given $\mathbf{s}_{-i}$ and $\mathbf{u}_{-i}$.

$$P(u_i = c|\mathbf{u}_{-i}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \frac{N_c(u_{-i}) + \boldsymbol{\beta}_c}{n - 1 + \boldsymbol{\beta}_\bullet} \frac{C(\boldsymbol{\alpha} + \sum_{i' \neq i: u_{i'}=c} N(s'_i) + N(s_i))}{C(\boldsymbol{\alpha} + \sum_{i' \neq i: u_{i'}=c} N(s'_i))}$$

Suppose we were updating this sampler running on the Finnish verb inflections. If we had all segments as in Figure 1, but wanted to resample u_{31} (1st person singular /n/), we would consider the counts N excluding that form (i.e., under $\mathbf{u}_{-31}$). The prior for /n/, β_{14}, is fixed, and there are no other occurrences, so $N_{14}(\mathbf{u}_{-31}) = 0$. Another potential underlying form, like /t/, would have higher unconditioned posterior probability, because of the counts

(7, in this case) added to its prior from $\boldsymbol{\beta}$. Then, we have to multiply by the probability of each generated surface segment (all are [n], so $7 * P([n]|c, \boldsymbol{\alpha})$ for a given hypothesis $u_{31} = c$).

We select a given character $c \in C$ for u_{31} from a distribution at random. Depending on the prior, /n/ will be the most likely choice, but other values are still possible with smaller probability. The counts used for the next resampling, $N(\mathbf{u}_{-31})$, are affected by this choice, because the new identity of u_{31} has contributed to the posterior distribution. After unbounded iterations, Gibbs sampling is guaranteed to converge and produce samples from the true posterior (Geman and Geman, 1984).

3 Evaluation

This model provides a language agnostic solution to a subset of phonological problems. We will first examine performance on the sample Finnish data (from Figure 1), and then look more closely at the issue of convergence. Finally, we present results from larger corpora [3].

3.1 Finnish

Output from a trial run on Finnish verbs (from Figure 1) follows, with hyperparameters $\alpha_{ij}\{100 \iff i = j, 0.05 \iff i \neq j\}$ and $\beta_i = \{0.1\}$.

In the paradigm (a sample after 1000 iterations), each [sur+face] form is followed by its hypothesized /under/ + /lying/ morphemes.

[tiedä+n] : /tiedä/ + /n/
[tiedä+t] : /tiedä/ + /t/
[tietä+ä] : /tiedä/ + /ä/
[tiedä+mme] : /tiedä/ + /mme/
[tiedä+tte] : /tiedä/ + /tte/
[tietä+vät] : /tiedä/ + /vät/
[aiøo+n] : /aiøo/ + /n/
...
[pelkää+vät] : /pelkää/ + /vat/

With strong enough priors (faithfulness constraints), our sampler often selects the most common surface form aligned with an underlying segment. Although [vat] is more common than [vät], we choose the latter as the purer underlying form. So /a/ is always [a], but /ä/ can be either [ä] or [a].

[3] 2.8 million word types from Morphochallenge2007 (Kurimo et al., 2007)

3.2 Convergence

Testing convergence, we run again on the sample data (Figure 1), using $\alpha_{ij} = 0.1$ when $i \neq j$ and 10 when $i = j$ and $\boldsymbol{\beta} = 0.1$, starting from different initializations, we get the same solution.

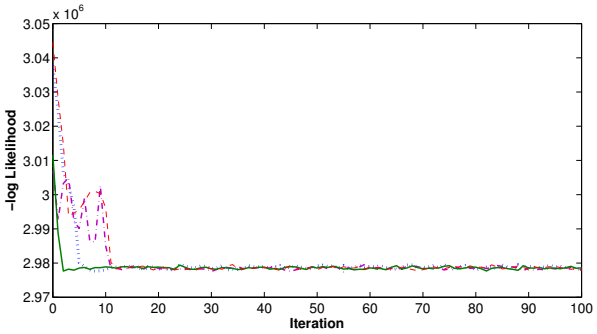

Figure 4: Posterior likelihood at each of the first 100 iterations, from 4 runs (with different random seeds) on 10% of the Morphochallenge dataset ($\alpha_{i \neq j} = 0.001, \alpha_{i=j} = 100, \beta = 0.1$), indicating convergence within the first 15 iterations.

To confirm that the sampler has converged, we output and plot trace statistics at each iteration, including marginal probability, log likelihood, and changes in underlying forms (i.e., variables resampled). If the sampler has converged, there should no longer be a trend (consistent slope) in any of these statistics (as in Figure 4).

Examining the posterior probability of each selected underlying form reveals interesting patterns that also help explain the variation. In the above run, the ambiguous segments (with surface alternations) were drawn from the distributions (with improbable segments elided) in Figure 5.

We expect this model to maximize the probability of either the "majority" solution or a solution demonstrating selection bias. We compare likelihood of the posterior sample with that of a "phonologically plausible" solution (in which underlying forms are determined by referring to formal linguistic accounts of phonological derivation) and a "majority solution" (see Figure 6 for a log-log plot, where lower is more likely).

The posterior sample has optimal likelihood with each parameter setting, as expected. The majority parse is selected with $\alpha_{i \neq j} = 0.5$ With lower values of $\alpha_{i \neq j}$, the "phonologically plausible" parse is

u_4=/d/ s_4=[d,d,t,d,d,t]

$P(u_i = c)$	$\approx$
d	0.99968
t	0.00014

u_8=/k/ s_8=[ø,ø,k,ø,ø,k]
(same behavior as u_{12})

$P(u_i = c)$	$\approx$
ø	0.642
k	0.124

u_{33}=/e/ s_{33}=[ä,o,e,u,ø,e,ø]

$P(u_i = c)$	$\approx$
ä,o,u	0.0029
ø	0.215
a	0.0003
e	0.297

Figure 5: Resampling probabilities for alternations, after 1000 iterations.

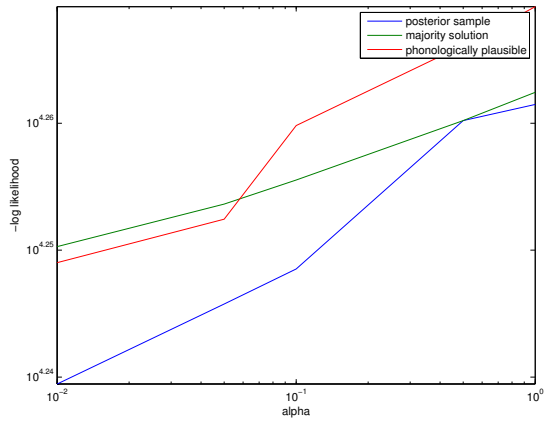

Figure 6: Parse likelihood

more likely than the majority. However, the sampler does not converge to this solution, because in this [t,d] alternation, the "phonologically plausible" solution identifies /t/, but neither selection bias nor majority rules would lead to that with the given data.

3.3 Morphologically segmented corpora

In our search for appropriate data for additional, larger-scale experiments, we found several viable alternatives. The correct morphological segmentations for Finnish data used in Morphochallenge2007 (Kurimo et al., 2007) provide a rich and varied set of words, and are readily analyzable by our sampler. Rather than associating each surface form with a position in the paradigm, we use the annotated morphemes.

	Majority	Bayesian
types	50.84	69.53
tokens	65.23	72.11

Figure 7: Accuracy of underlying segment hypotheses.

For example, the word *ajavalle* is listed in the corpus as follows:

ajavalle aja:ajaa|V va:PCP1 lle:ALL The word is segmented into a verb stem, 'aja' (drive), a present participle marker 'va', and the allative suffix ("for"). Each surface realization of a given morpheme is identified by the same tag (e.g., PCP1). However, in this corpus, insertion and deletion are not explicitly marked (as they were in the paradigm, by ø). Rather than introduce another component to determine which segments in the form were dropped, we ignore these cases.

The sampling algorithm proceeds as described in section 2. To run on tokens (as opposed to types), we incorporate another input file that contains counts from the original text (*ajavalle* appeared 8 times). The counts of each morpheme's surface forms then reflect the number of times that form appeared in any word in the corpus.

3.3.1 Type or Token

In Finnish verb conjugation, 3rd person (esp. singular) forms have high frequency and tend to be unmarked (i.e., closer to underlying). In types, unmarked is a minority (one third), but incorporating token frequency shifts that balance, benefiting the "majority learner." Among noun inflections, unmarked has higher frequency in speech, but marked tokens may still dominate in text. We might expect that it is easier to learn from tokens than types, in part because more data is often helpful.

Testing on half of the Morphochallenge 2007 Finnish data (1M word types, 5M morph types, 17.5M word tokens, 48M morph tokens), we ran both our Bayesian model and a majority solver on the morphological analyses, and compared against phonologically plausible (gold) underlying forms. Results are reported in Figure 7.

The Bayesian estimate consistently outperformed the majority solution, and cases where the two differ could often be ascribed to the preference for "pure"

analyses.

4 Conclusion

We have described a model where surface forms are generated from underlying representations segment by segment. Taking this approach allowed us to investigate the properties of a Bayesian statistical learner, and how these can be useful in the context of sound systems, a basic component of language. Experiments with our implementation of a collapsed sampler have produced results largely confirming our hypotheses.

Without context, we can often learn about 60 to 80 percent of the mapping from underlying phonemes to surface phones. Especially with lower values of $\alpha_{i \neq j}$, closer to 0, our model does prefer pure alternations. Gibbs sampling tends to select the majority underlying form, particularly with $\alpha_{i \neq j}$ relatively high, closer to 1. So, a sparser prior leads us further from the baseline, and often closer to a phonologically plausible solution.

4.1 Directions

In future research, we hope to integrate morphological analysis into this sort of a treatment of phonology. This is a natural approach for children learning their first language. They intuitively discover phonotactics, and how it affects the prosodic shape of each word, as they learn meaningful units and compose them together. It is clear that many layers of linguistic information interact in the early stages of child language acquisition (Demuth and Ellis, 2005 in press), so they should also interact in our models. As discussed above, the present model should be applicable to analysis of language-learners' speech errors, and this connection should be explored in greater depth.

It might be interesting to predispose the sampler to select underlying forms from open syllables. That is, set $\boldsymbol{\alpha}$ to increase the probability of matching one of the surface segments if its context (feature annotations) includes a vocalic segment or a word boundary immediately following. The probability of phonological processes like assimilation could be similarly modeled, with the prior higher for choosing a segment that appears on the surface in a contrastive context (where it shares few features with neighboring segments).

If we define a MaxEnt distribution over Optimality Theoretic constraints, we might use that to inform our selection of underlying forms. In (Goldwater and Johnson, 2003), the learning algorithm was given a set of candidate surface forms associated with an underlying form, and tried to optimize the constraint weights. In addition to the constraint weights, we must also optimize the underlying form, since our goal is to take as input only observable data. Sampling from this type of complex distribution is quite difficult, but some approaches (e.g., (Murray et al., 2006)) may help reduce the intractability.

References

Kenneth R. Beesley and Lauri Karttunen. 2000. Finite-state non-concatenative morphotactics. In Lauri Karttunen, Jason Eisner, and Alain Thériault, editors, *SIGPHON2000, August 6 2000. Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology.*, pages 1–12.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August.

Hal Daume III. 2007. Hbc: Hierarchical bayes compiler.

Katherine Demuth and David Ellis, 2005 (in press). *Revisiting the acquisition of Sesotho noun class prefixes*. Lawrence Erlbaum.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6):721–741, Nov.

Dale Gerdemann and Gertjan van Noord. 2000. Approximation and exactness in finite state optimality theory.

Daniel Gildea and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics*, 22(4):497–530.

Sharon Goldwater and Mark Johnson. 2003. Learning ot constraint rankings using a maximum entropy model.

Sharon Goldwater and Mark Johnson. 2004. Priors in Bayesian learning of phonological rules. In *Proceedings of the Seventh Meeting of the ACL Special Interest Group in Computational Phonology*, pages 35–42, Barcelona, Spain, July. Association for Computational Linguistics.

Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater. 2007. Adaptor grammars: A framework for specifying compositional nonparametric Bayesian models. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 641–648. MIT Press, Cambridge, MA.

Lauri Karttunen. 1998. The proper treatment of optimality in computational phonology. In Lauri Karttunen, editor, *FSMNLP'98: International Workshop on Finite State Methods in Natural Language Processing*, pages 1–12. Association for Computational Linguistics, Somerset, New Jersey.

Mikko Kurimo, Mathias Creutz, and Ville Turunen. 2007. Overview of morpho challenge in clef 2007. In *Working Notes for the CLEF 2007 Workshop*, Budapest, Hungary.

Percy Liang and Dan Klein. 2007. Tutorial 1: Bayesian nonparametric structured models, June.

Iain Murray, Zoubin Ghahramani, and David MacKay. 2006. MCMC for doubly-intractable distributions. In *UAI*. AUAI Press.