Nat Lang Linguist Theory (2012) 30:859–896
DOI 10.1007/s11049-012-9169-1

Information theoretic approaches to phonological structure: the case of Finnish vowel harmony

John Goldsmith · Jason Riggle

Received: 25 July 2009 / Accepted: 30 March 2012 / Published online: 26 May 2012

Abstract This paper offers a study of vowel harmony in Finnish as an example of how information theoretic concepts can be employed in order to better understand the nature of phonological structure. The probability assigned by a phonological model to a corpus is used as a means to evaluate how good such a model is, and information theoretic methods allow us to determine the extent to which each addition to our grammar results in a better treatment of the data. We explore a natural implementation of autosegmental phonology within an information theoretic perspective, and find that it is empirically inadequate; that is, it performs more poorly than a simple bigram model. We extend the model by means of a Boltzmann distribution, taking into consideration both local, segment-to-segment, relations and distal, vowel-to-vowel, relations, and find a significant improvement. We conclude with some general observations on how we propose to revisit other phonological questions from this perspective.

Keywords Information theory · Learning · Vowel harmony

1 Introduction

1.1 Information theoretic phonology

Vowel harmony has been a constant concern of phonologists since Trubetzkoy's *Grundzüge der Phonologie* 1939/1968, because it has something to interest everyone.

J. Goldsmith · J. Riggle (✉)
University of Chicago, Chicago, IL, USA
e-mail: jriggle@uchicago.edu

J. Goldsmith
e-mail: jagoldsmith@uchicago.edu

Vowel harmony is widespread without being anywhere near universal; it is often phonetically motivated and yet, just as often, not entirely phonologically regular. Where it is found, vowel harmony describes the phonotactics of the language, governing the choice of vowels that appear within a morpheme and the choice of distinctive vowels over morpheme boundaries within words. In many cases, furthermore, vowel harmony appears to provide *prima facie* evidence of the active role played by distinctive features in natural language.

Our goal in this paper is to explore the role vowel harmony plays as a phonotactic using information theoretic models. Such models provide phonologists with remarkably powerful quantitative tools for analysis starting from very few empirical assumptions. Such models can, indeed, be understood as empiricist models of phonological material, in the sense that the generalizations that emerge can be perfectly well understood as inhering in the data, rather than being the result of inferences that we make after the fact about the hidden nature of the device (the human brain) that generated the data in question (the phonological representations of various utterances). The question that we pose in this paper is whether such information theoretic models can be extended to the treatment of vowel harmony systems, and if so, whether non-local effects can be discovered and modeled in phonological data.

The central idea in this approach is that virtually all rational thought about empirical observations—of any sort—can be recast as a pair of measurable hypotheses: a fully explicit statement of the hypothetical model responsible for 'generating the data,' using the tools of computer science, and a precise statement of how likely the model predicts the observations to be.[1] The goal is to find a model that is simultaneously (relatively) simple and a good predictor of the observed data.

In Sect. 2, we define a class of probabilistic phonological models, and sketch some reasons for believing that these types of models are the most suitable for describing phonotactics. Despite their obvious utility and ubiquity in other fields, these models are not at present the norm in mainstream linguistics; we believe they should be, and some of the general reasons for this have been discussed elsewhere by Goldsmith (2007a, 2007b).

In Sects. 3 and 4, we explore a powerful aspect of the probabilistic framework, which is that it allows us to algorithmically compare alternative probabilistic models of a given set of data, and empirically test which is the superior model by comparing the probability assigned to the same set of data. To illustrate our proposal, we explore the vowel harmony system of Finnish, and present a probabilistic model within which it is possible to discover non-local dependencies among Finnish vowels and to represent these dependencies with probabilistic grammars.

In developing the Finnish model, we first illustrate a well-known limitation of probabilistic models based on pairs of adjacent segments, which is that they cannot capture non-local phenomena. We then propose to overcome this limitation by augmenting the model with an autosegmental vowel tier on which the harmonizing vowels are adjacent, and hence their interaction is local. Surprisingly, this move does

[1] This tradition of probabilistic analysis comes from encoding theory and works like those of Shannon and Weaver (1949), Shannon (1951), Solomonoff (1959a, 1959b, 1964, 1997), Rissanen (1989), or Li and Vitányi (1997), to mention a few of the most important.

not improve the analysis of Finnish vowel harmony. The failure arises from the fact that our augmented model isolates the vowels on the vowel tier and thereby occludes consonant-to-vowel and vowel-to-consonant patterns in the Finnish data. This leads us to develop a somewhat more sophisticated probabilistic model using a Boltzmann distribution to allow both segment-to-segment and vowel-to-vowel effects to be simultaneously modeled. This result has implications for how we should account for local and non-local patterns simultaneously and how we should understand autosegmental models of vowel harmony systems.

1.2 Probabilistic models

Probabilistic models are in essence quantitative models of evidence; they are ideally suited for a domain such as phonology, in which the goal is to determine what aspects of the data are due to structure (either of a sort already understood, or of a sort that remains to be determined) and what aspects are not.

Probabilistic models offer the possibility of a style of model evaluation in which success is quantifiable in terms of the entirety of the available data without the need for putative exceptions to be ignored. As we will see in the examples analyzed in this paper, there are two primary characteristics of empirical probabilistic studies. First, the data which they seek to account for is generally quite large, and, in particular, is not selected or filtered to suit the specific needs of the analysis being evaluated—ideally, it is a corpus that has been independently collected. Second, describing the data in probabilistic terms provides a clear quantitative measure of the degree to which the model predicts the patterns in the data.

A probabilistic model begins with a theoretical statement of what the sample space is, what the universe of possibilities is, and a measure of how much of the probability of that sample space is to be attributed to the observed data. If the observed data filled all of the sample space (so that its probability were 1.0), then the model would be claiming (absurdly, in most cases) that the observations were an empirical necessity and could never have been anything other than what they were. In actual cases, the probability assigned to the observations will be quite small, but the discrepancies of the probabilities assigned to the data by different models will nonetheless constitute a clear statement about how 'accidental' each model takes the data to be. In general, all other things being equal, we prefer an analysis in which the patterns in the data are seen to be as minimally accidental as possible. In this context, what this means is that we want to find the model by virtue of which we can assign the highest probability to the data.

We wish to underscore the fact that probabilistic linguistic models are thoroughly structure-dependent; no such model can be developed without a clear understanding of the structure that it proposes to find in the data. Probabilistic models in linguistics have sometimes been associated with skepticism about the existence of abstract structure, but this perspective is not inherent to the logic of a probabilistic model (and we are not skeptical about the existence of such structure). Nonetheless, there is a reason that lurks behind this perception: probabilistic models extract considerably more information about a body of data than non-probabilistic models.

As we will see below, when faced with a model of phonological data in which each representation is little more than a sequence of phonemes, a probabilistic model

is able to extract a considerable amount of information without making use of features or hierarchical structure. Probabilistic models are not inimical to structure, far from it—they offer an explicit and quantitative measure of how much improvement models with additional structure give us, by comparison with less structured models.

The probabilistic approach to linguistic description consists essentially of noting the way in which the basic elements of a language (phonemes, features, units of all sorts) depart from equiprobability: to what extent is the average utterance in language X a *departure* from a random compilation of sounds from its inventory of sounds? We are able to measure how well a given model fits the data from a language in terms of its ability to quantify the degree to which the average utterance is indeed a departure from randomness.

Along the way, numerical values representing probabilities will be assigned to the basic elements of the model and to each way of combining them (regardless of whether each possible combination is actually present in the language). In this way, probabilistic analyses implement a straightforward model of markedness in terms of deviation from what is typical in a language.

2 Basics of probability

2.1 Distributions

A discrete probabilistic model such as we will consider here consists of a set U_1, called the *sample space*, in which each element is associated with a value between 0 and 1, which is its *probability*. In this work, our sample space will consist of an infinite number of simple phonological representations. We will generally refer to the function that associates an element to its probability as pr() and note that, in order to be well-formed, the sum of the probabilities associated with each of the elements in the sample space of the function must be 1. A function such as pr() that assigns non-negative numbers to the members of a set so that the sum of these values is 1.0 is said to be a *distribution* over that set.

$$(1) \qquad \mathrm{pr} : U_1 \rightarrow [0, 1] \qquad \sum_{x \in U_1} \mathrm{pr}(x) = 1$$

Note that one can have different probability models that are based on the same sample space but assign different probabilities to the individual elements.

Our task will be to consider various distributions over linguistic structures and to figure out which is the best distribution and how it can be calculated. In reasoning about these models, it is important to bear in mind the fact that the sum of an infinite series of positive values can be finite, such as $0.9 + 0.09 + 0.009 + \cdots = 0.\bar{9} = 1$. Though this seems counter-intuitive to some, $0.\bar{9}$ and 1.0 are merely different representations of the same number (see Courant and Robins 1941: 64 or Lewin 2003: Chap. 12 for more discussion of this fact). Even though the set of possible linguistic structures in our sample space is infinite, the sum of the probabilities associated with these structures will always equal 1, even in cases where the function pr(·) assigns non-zero probability to each one.

2.2 Strings and their simplest model

Let us start by developing a probabilistic model for strings of symbols. In the present work, the symbols will represent phonemes but they could also correspond to feature bundles, autosegmental representations, etc. We begin with a finite set of symbols, A, referred to as the *alphabet*. The notation A^+ denotes all strings (sequences) of one or more symbols drawn from A. We add a special symbol, # not present in A, to represent the word boundary.[2] We then define a *word* as any finite sequence of one or more symbols that ends with #. Given this definition, a *word-set* S is a subset of the set of all possible words: $S \subseteq (A^+\#)$. Similarly, a *word-list* or *corpus* C is an element of the set of all possible sequences of words, $C \in (A^+\#)^*$. In this work we will be presenting analyses based on sets of words, so the definitions that we give in this section will be for word-sets.

One of the simplest questions that can be asked about a set of words is how often any given single symbol, or *unigram*, appears. For a unigram a, we will write $Count(a)$ to indicate the total number of times that a occurs in all the words in the set. For each symbol $a \in A \cup \{\#\}$, the unigram model induced from a word-set S assigns a probability to a that represents its frequency in the word-set. That is:

$$\text{(2)} \qquad \text{pr}(a) = \frac{Count(a)}{|S|}$$

where $|S|$ equals the total number of symbols in all the words in S.

For a word $w \in S$, we use the notation $w_{[n]}$ to refer to the n-th symbol in the string (i.e. $w_{[1]}$ is the first symbol, $w_{[2]}$ the second, and so on). Given a word w, the *unigram probability* of w, denoted $\text{pr}(w)$, is defined as the product of the probabilities of the segments comprising the word. For a set of words S, the product of the probabilities of the words is denoted $\text{pr}(S)$. These are given in (3a) and (3b):

$$\text{(3)} \qquad \text{a.} \quad \text{pr}(w) = \prod_{i=1}^{|w|} \text{pr}(w_{[i]}) \qquad \text{b.} \quad \text{pr}(S) = \prod_{w \in S} \text{pr}(w)$$

where $|w|$ denotes the number of symbols in w (i.e. the length of the word). Note that in (3a), pr is a probability distribution over all possible words. When evaluating a set of words as in (3b), pr is a distribution over all sets of size k for any given $k > 0$, but pr is not a distribution over all sets of all sizes. This is appropriate in cases, such as the one at hand, where one is evaluating different models of the same word-set S because k is fixed. This very simple model independently scrutinizes each of the segments of the word without any regard for their order or configuration.

In many cases, the probability computed by a model is the product of a number of distinct factors; because $x \times y = \log(x) + \log(y)$ we can interpret the probability

[2] This allows us to assess average word length and to refer to segments at word edges as being adjacent to # in the same way that they are adjacent to their segmental neighbors. Like the symbols for phonemes in A, the symbol # is associated with a probability and can condition the probability of its neighbors. Thus, in what follows, we will refer to # as a phoneme (though it is, in many ways, a different kind of abstract object than a consonant or a vowel).

Table 1 Top and bottom five words and phonemes by (average) *plog*

Rank	Orthography	Phonemes	*Avg. plog*
1	a	ə	3.11
2	an	ən	3.44
3	to	tə	3.47
4	and	ənd	3.80
5	eh	ɛ́	3.88
63,200	geoid	ǰíɔ̌y̌d	7.40
63,201	Cesare	čězárě	7.40
63,202	Thurgood	θə́gɤ̌d	7.47
63,203	Chenoweth	čɛ́nɔ̌wěθ	7.49
63,204	Qureshey	kəréšě	7.54

Rank	Phoneme	*Plog*
1	#	2.30
2	ə	3.92
3	n	4.10
4	t	4.17
5	s	4.61
50	ɔ̌	11.79
51	ě	12.76
52	ɤ̌	14.30
53	aw̌	14.35
54	ɔy̌	15.91

assigned to a form as the sum of the logarithms of these factors. Since $\log(x)$ is negative for $0 < x < 1$, the logs of probabilities are often multiplied by -1 to yield what is referred to as *inverse log probability*; we propose a simpler neologism, the *positive log probability*, or *plog*, for short. Thus (3) can be recast with plogs as in (4).

$$\text{(4)} \quad \text{a.} \quad plog(w) = -1 \sum_{i=1}^{|w|} \log \text{pr}(w_{[i]}) \qquad \text{b.} \quad plog(S) = -1 \sum_{w \in S} \log \text{pr}(w)$$

The average plog of a word w or word-set S can be calculated as in (5a) or (5b).

$$\text{(5)} \quad \text{a.} \quad -\frac{1}{|w|} \sum_{i=1}^{|w|} \log \text{pr}(w_{[i]}) \qquad \text{b.} \quad -\frac{1}{|S|} \sum_{w \in S} \log \text{pr}(w)$$

Insofar as expectedness is the opposite of complexity (a basic premise of coding theory), the average plog, as calculated in (5a), encodes the average complexity of the phonemes comprising the word. If we calculate this figure for all the words of our vocabulary and sort them in light of this figure, the words with the smallest value will be the words largely composed of high frequency phonemes, and the words with the largest values will be words composed largely of low frequency phonemes.

In Table 1 we illustrate the range of average plogs from the top five and the bottom five of a sample of 63,204 English words along with the plogs of the frequencies of the top and bottom five of 54 English phonemes. The data combines a modified version of the CMU English lexicon weighted by word frequencies based on counts from the Brown corpus. The particular transcriptions that appear may raise some eyebrows, but we have used their transcription throughout, though we have used here American phonetic symbols rather than the Darpabet.

If one takes the (not completely uncontroversial) position that markedness is correlated with frequency, then the plogs in this table would be seen as roughly quantitative estimates of various segments' markedness.[3]

2.3 Linear structure: bigram model

Unigram models describe the basic frequency of phonemes. Much of the phonological structure of languages, however, involves conditions on sequences of phonemes, which goes beyond the descriptive purview of unigram models. The natural way to encode this information is to use a bigram model, which is to say, to use as the probability for a given phoneme its probability in a given context.

One of the simplest models along these lines conditions the probability of a phoneme on its left-hand neighbor in the word. Because the initial segment of a word, $w_{[1]}$, does not have a left-neighbor, it is conventional to define $w_{[i<1]}$ as the boundary symbol #. Informally speaking, the *conditional probability* of phoneme b immediately following a, where a is the left-neighbor or # if b is word-initial, is calculated as in (6):

$$\text{(6)} \qquad \text{pr}(b \mid a) = \frac{Count(ab)}{Count(a)}$$

where $Count(ab)$ denotes the number of times that b occurs in context a in the wordset and $Count(a)$ denotes the number of times that context a occurs.[4]

A considerable advantage comes now from using logarithms: it allows us to easily express what the advantage is of the bigram model over the unigram model. The change in the log probability computed under the unigram and the bigram models is precisely equal to another quantity of particular interest, the *mutual information*, defined as in (7).

$$\begin{aligned} \text{(7)} \qquad MI(a;b) &= \log \frac{\text{pr}(ab)}{\text{pr}(a)\,\text{pr}(b)} = \log \text{pr}(ab) - \log \text{pr}(a) - \log \text{pr}(b) \\ &= -plog(ab) + plog(a) + plog(b) \end{aligned}$$

If $\text{pr}(ab) = \frac{Count(ab)}{|S|}$ is the probability of the pair ab and $\text{pr}(a)\,\text{pr}(b)$ is the product of the symbol's individual probabilities, then the mutual information between a and b is the log of the ratio of these quantities. The probability of a joint event, such as the sequence ab, is equal to the product of the individual probabilities just in case the two events are independent of each other (this being the definition of independence), so the ratio here takes the value 1 just in case the two events are independent.

If the probability sequence of the phonemes is greater than the product of the individual probabilities, then the structure involved in the model being explored pulls

[3] In constraint-based models other than Optimality Theory (Prince and Smolensky 1993/2004) that allow violability but eschew strict domination such as, e.g., Pater et al. (2007), Hayes and Wilson (2008), or Goldwater and Johnson (2003) one could view the unigram model as setting up a constraint against each segment, and weighting the violation of constraint $^*a$ by the value $plog(a)$.

[4] More formally, $\text{pr}(b \mid a)$ is the probability that a certain function will take on the value a or ab (see Cover and Thomas 1991: Chap. 2 for a thorough exposition).

Table 2 English words ranked by average plog in the bigram model

Rank	Orthography	Phonemes	*Avg. plog*$_2$
1	the	ðə	1.93
2	hand	hǽnd	2.15
12,640	plumbing	plʌ́mɨ̌ŋ	3.71
12,642	Friday	fraýdɨ̌	3.71
25,281	tolls	tólz	4.01
25,282	recorder	rɨ̌kórdə̆˞	4.01
37,922	overburdened	óvə̆˞bə̆˞dənd	4.32
37,923	Australians	ɔ̆streýlyənz	4.32
50,563	retire	rɨ̌taýr	4.75
50,564	poorer	púrə̆˞	4.75
63,200	eh	ɛ́	9.07
63,201	Oahu	óáhŭ	9.21

the two events together, while if the probability of the phonemes together is less than the product, the structure at hand is responsible for them repelling each other, so to speak. By taking the logarithm of this ratio, we translate attraction to a positive value, repelling to a negative value, and independence to a zero value. (When we are calculating this quantity for particular symbols, the term *pointwise mutual information* is often used, and then the term *mutual information* is used to describe the average pointwise mutual information as we average over all pairs of elements, each pair weighted by its probability.)

Just as important, the mutual information is exactly the difference between the unigram and bigram models' log probability. This is shown in (8).

$$
\begin{aligned}
(8) \qquad \sum_{i=1}^{|w|} \log \mathrm{pr}(w_{[i]} \mid w_{[i-1]}) &= \sum_{i=1}^{|w|} \log \frac{\mathrm{pr}(w_{[i]}\, w_{[i-1]})}{\mathrm{pr}(w_{[i-1]})} \\
&= \sum_{i=1}^{|w|} \log \mathrm{pr}(w_{[i]}) + \sum_{i=1}^{|w|} \log \frac{\mathrm{pr}(w_{[i]}\, w_{[i-1]})}{\mathrm{pr}(w_{[i-1]})\,\mathrm{pr}(w_{[i]})} \\
&= \sum_{i=1}^{|w|} \log \mathrm{pr}(w_{[i]}) + MI(w_{[i-1]};\ w_{[i]})
\end{aligned}
$$

For a concrete illustration we return to our English word list from Table 1. Our English data set contains 54 phonemes, and thus there are $54^2 = 2{,}916$ possible bigrams. Consider, in Table 2, the way that the bigram model enriches the evaluation of the English data by taking two-word slices at six points along the ranking of all 63,000 words according to their average bigram plog.

With the bigram model, we obtain a set of parameters that describe the phonological well-formedness (in terms of 'typicality') to a second order degree of detail. If there are P phonemes in the language, then the number of parameters for the uni-

gram and bigram models together is $P + P^2$. Each setting of values (weights) for the parameters assigns a probability to a corpus, and the degree of success acheived by a set of parameters with weightings can be measured by that probability: the higher the probability, the more successful the characterization.[5]

The reader should bear in mind the following tension: when presented with probabilistic analyses of this sort regarding linguistic data, there is often a tendency to interpret it as, in effect, an implicit argument against the necessity for structure going beyond that which is used by the model (i.e. purely linear and symbolic structure). The reason for this may be that one can get quite striking results on the basis of simple quantitative methods that do not incorporate structure other than the most superficial. It might not be surprising if some were to interpret these results as a sign that linguistic structure need not be incorporated in the next generation of phonological analysis, and that what is necessary is more mathematics instead. Such a conclusion would be hasty, and just as surely wrong. We believe that the right way to think about it is that any account of phonological representations will include statements about segmental inventory and linear position in the formation of morphemes and words, and if meaningful generalizations can be extracted with simple modeling of this data, then we should identify what that information is. But that there *is* more structure than linear and quantitative structure is not in any sense challenged by the material we describe here, as the second part of this paper, on vowel harmony, attempts to show.

2.4 The problem of sparse data

The problem posed by sparse data is how to treat all the structures that occur rarely or not at all in the training data. Thus far, we have been using what are known as maximum likelihood estimates (MLE) in our models. Using MLE, the probability assigned to structure a, $\mathrm{pr}(a) = \mathit{Count}(a)/|S|$ is essentially its frequency. This approach provides the tightest fit between the parameters (i.e. probability estimates) in a model and the data with which it is trained. Consequently, any structures (phones, n-grams, whatever) that are not observed in the training data will be assigned zero probability and thus treated as true grammatical impossibilities. It can be the case, however, that the missing structures are accidental gaps in the training data. When a model erroneously treats an accidental gap as a systematic gap the model is said to have *over-fit* the training data.

If the goal is the construction of a generative model, MLE probabilities are usually avoided because they yield models that are 'brittle' in the sense that the occurrence

[5]One striking characteristic of probabilistic phonology of the 1950s (e.g., Cherry et al. 1953; Belevitch 1956; etc.), compared with what we attempt to do here (or Coleman and Pierrehumbert 1997), is the focus in that early work on average values over an entire corpus. The clearest example of this is the emphasis on calculating the entropy of a language under various models. The entropy is the weighted average of the inverse log frequency, and each word in the lexicon contributes to its computation in proportion to the word's frequency in the language. By contrast, we are not only interested in these ensemble averages, we are also interested in how some words (or subgroups of words) differ from other words, although we have not emphasized that in this paper. The most striking relevance of probabilities at the level of individual words involves the selection of the appropriate form of a suffix in a vowel harmony system, in which we typically find (as we find in Finnish) two forms of the suffix, one corresponding to each of the harmonic feature values. Selection of the correct harmonic feature value in a suffix corresponds to selection of the suffix allomorph that maximizes the phonological probability of the word (stem plus suffix).

of a zero-probability element in a form nullifies all other distinctions (i.e. any pair of words containing zero probability elements have the same probability, zero, regardless of any other distinctions between them). This problem has been extensively studied in statistical natural language processing, and it has been approached with a wide range of sophisticated solutions that go by the general name of *smoothing* techniques.

One of the most basic smoothing strategies is to use Laplace's Law in a scheme that adds one to all counts by initializing each count to one when computing frequencies. This is a specific instance of a more general strategy of adding λ, called Lidstone's Law:

$$\text{pr}(s) = \frac{Count(s) + \lambda}{N + B\lambda}, \tag{9}$$

where N is the total number of instances of structures like s, and B is the number of possible kinds of structures like s. When $\lambda = 0$ this formula is simply the maximum likelihood estimator; this gives the best fit for the training data but reserves no probability for unseen events. When $\lambda = 1$ we are using what is usually referred to as Laplace's Law, which corresponds conceptually to a uniform Bayesian prior over the possible structures. When $\lambda = 1/2$ we are using what is usually called the Jeffreys-Perks Law (though Perks more strongly advocated $\lambda = 1/|T|$ where T is the set of types). The value $\lambda = 1/2$ is also referred to as expected likelihood estimation (ELE) and is the most commonly used fixed value for λ in language modeling. There are many strategies for calculating optimal values for λ in given contexts and, more generally, many other strategies for calculating the amount of probability to reserve for unseen events (see Manning and Schütze 2000, Chap. 6 for an overview and Good 1980 for a thorough discussion of the development of many of these ideas). In our general presentation of models in the sections that follow we will use MLE probabilities. However, whenever we compare alternative models we will use ELE $\lambda = 1/2$ in smoothing the probabilities. Smoothing is useful in comparing alternative models to evaluate not only their ability to fit the data but also their tendency to over-fit the data. In Sect. 4.4 we will discuss an alternative to smoothing whereby minimization of model complexity is used to avoid over-fitting.

3 Finnish

In this section, we consider information theoretic approaches to patterns whose scope is larger than pairs of adjacent segments. For this study, we will use vowel harmony in Finnish. Vowel harmony presents a type of phonological pattern that simple bigram models miss, but that any algorithm designed to act like a human phonologist ought to detect. In vowel harmony, vowels exhibit a high degree of mutual information, but because they can be separated by varying numbers of consonants, this information is hidden from bigram models.

In the next two sections, we will explore more sophisticated bigram models that capitalize on the autosegmental idea that segments which are not adjacent in the surface string can be adjacent at another level of representation (i.e. on another tier). By

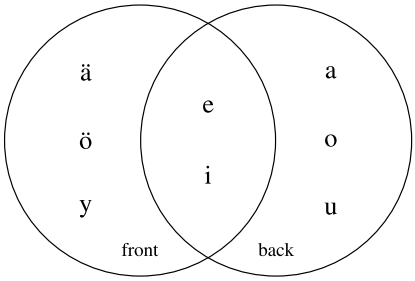

Fig. 1 Front, neutral, and back vowels in Finnish

allowing our model to try out various partitionings of the segments into groups that interact as if they were adjacent even when segments from a different group intervene, we are able to algorithmically discover something like an autosegmental vowel tier. In Finnish, separating the vowels from the consonants adds an extra level of representation where all vowels are adjacent and thereby renders their mutual information transparent to a bigram-based analysis over the separate tiers. Before getting into the specifics and results of the model, we will establish a baseline with unigram and bigram models for a Finnish corpus.

We remind the reader that Finnish contains eight vowels that are usually grouped into those that are *strictly front* $\{\ddot{a}, \ddot{o}, y\}$, those that are *strictly back* $\{a, o, u\}$, and those that are *neutral* $\{i, e\}$— this is illustrated in Fig. 1. The strictly front and the neutral vowels together comprise the *front vowels* of the language, and the strictly back and the neutral vowels comprise the *back vowels* of the language. A majority of words in Finnish are *harmonic*, which is to say, all the vowels of a given word come from either the front vowel set or the back vowel set.[6]

We begin with an analysis of unigrams and bigrams to establish a baseline against which models of harmony can be evaluated. For this case study, we used a word list containing 44,040 unique inflected Finnish words with initial and final word boundary symbols '#'. The orthography of Finnish is particularly helpful to our endeavor because it transparently encodes the relevant properties of the vowels.

3.1 A unigram model of Finnish

In Table 3 we present the counts, frequencies, and plogs for the unigrams in our Finnish corpus. Analysis of the unigrams gives us a fair amount of information about the corpus.

There are 510,174 unigrams in total, the sum of their positive log probabilities (*plogs*) is 2,088,530, and the average positive log probability per segment is 4.09. Thus the average positive log probability (i.e. the entropy under this model) of our corpus is 4.09 and the total cost (in terms of bits of information) for encoding our Finnish corpus given the unigram model is 2,088,530 bits.

This base-line entropy of 4.09 for the unigram model and base-line encoding cost of 2,088,530 bits is what we aim to improve upon with more articulated models expressed over bigrams and other kinds of enriched representations.

[6] See Kiparsky (1973), Ringen (1975/1988), and citations in Ringen and Heinämäki (1997).

Table 3 Counts and frequencies for unigrams in our Finnish corpus

Type	Count	Frequency	*plog*	Type	Count	Frequency	*plog*
a	56397	0.11000	3.18	r	13540	0.02650	5.24
i	50053	0.09810	3.35	v	11487	0.02250	5.47
t	47927	0.09390	3.41	p	9970	0.01950	5.68
#	44040	0.08630	3.53	y	9300	0.01820	5.78
s	38567	0.07560	3.73	h	9018	0.01760	5.82
e	37362	0.07320	3.77	j	7048	0.01380	6.18
n	35072	0.06870	3.86	d	3734	0.00732	7.09
l	28060	0.05500	4.18	ö	2989	0.00586	7.42
k	26064	0.05100	4.29	g	828	0.00162	9.27
u	25314	0.05000	4.33	b	580	0.00113	9.78
o	22097	0.04330	4.53	f	326	0.00063	10.6
ä	15102	0.02960	5.08	c	312	0.00061	10.7
m	14815	0.02900	5.11	w	118	0.00023	12.0

3.2 A bigram model of Finnish

Though a unigram model of Finnish captures some basic properties of the data, it can be significantly improved by widening the model's scope to include information from adjacent segments (as will be true in all natural languages). Incorporating bigrams into our model of the Finnish corpus will capture more of the structure that is present in the data and thus will assign a higher probability to the corpus.

Table 4 gives the mutual information for a small 8×8 fragment of the bigrams of Finnish. Recall from (7) that the MI for a bigram ab is $plog(a) + plog(b) - plog(ab)$. For the bigram ab which has a frequency of 0.0001 whose positive log is 13.29, the mutual information $MI(a; b) = -0.33$ is obtained by adding 3.18 to 9.78 (the plogs for a and b in Table 3) and then subtracting 13.29. The natural unit for quantifying mutual information is the *bit*. Mutual information tells us the increase or decrease in the cost of describing a segment in a particular environment, given our model.

In Table 4, the base cost of describing a segment is taken to be plog of its unigram probability (also expressed in bits), so the MI directly encodes the increase or decrease in the expectation of a given segment in a particular environment. This is expressed in terms of how many fewer bits it takes to describe that segment in that environment. Consider the third row in Table 4. This row gives the MI for bigrams in which the first element is b. Here we see that the word boundary is less common immediately following a b than it is overall, so the cost of describing it (3.53 bits in Table 3) goes up by 3.06 bits. Conversely, b is relatively more common immediately following another b, so the cost of describing it (9.78 bits in Table 3) goes down by 4.30 bits. The clusters bf, bg, and bh are unattested in our corpus and thus, since log 0 is undefined, their plogs and MI are also undefined.[7]

[7] Because there are zero occurrences of f following b in the corpus the ML estimate of the probability of f in this position is zero. Leaving no bits/probability aside for f makes the description of the attested

Table 4 Mutual information among Finnish bigrams (∗ marks gaps)

	#	a	b	c	d	e	f	g	h	...
#	*	−1.04	2.54	2.00	−1.41	−0.97	2.40	0.46	1.62	
a	1.33	−0.19	−0.33	−0.30	−0.49	−4.79	−0.92	−0.63	0.03	
b	−3.06	1.06	4.30	3.08	−2.09	1.08	*	*	*	
c	−1.05	0.42	1.50	2.39	*	0.57	2.33	0.98	3.69	
d	−2.36	−0.55	1.08	*	−2.77	2.64	−1.25	−0.60	−6.04	
e	−0.65	−3.54	−1.41	−0.43	0.58	0.06	−1.99	−0.06	0.50	
f	−1.64	−0.26	*	*	−1.25	1.47	5.72	*	*	
g	−0.91	0.24	1.67	*	−0.60	1.34	0.92	1.16	−1.55	
h	−5.36	0.73	−3.36	*	3.39	0.90	−1.53	*	*	
...										

There are 510,174 segments in our Finnish corpus. The average unigram plog is 4.09 bits and concomitantly the sum of the unigram plogs for the whole corpus is 2,088,533 bits. The average MI among adjacent segments is 0.59 bits per bigram. This increase in the probability of each bigram is directly reflected in the sum of the positive logs of the conditional probabilities in the corpus, which is 1,780,261 bits.

3.3 Harmony and tiers

But what of vowel harmony? Thus far it has played no role in our description of Finnish. The challenge, as discussed in Sect. 1.2, is to formulate a representation for Finnish words under which potential connections among non-adjacent vowels can be described in the same way as the connections between adjacent segments. One particularly simple way to do this is to bifurcate the Finnish corpus so as to extract the vowels onto a separate tier that excludes the consonants. By selectively ignoring consonants we can obtain a vowel-only sub-corpus that exposes connections among non-adjacent vowels.

But it is appropriate to stop and ask what the epistemological basis is for using features: given the nature of what we are trying to do, should we allow ourselves to use them? Can a foundation be found for them that rests on probabilistic grounds?

The answer is yes. We outline here the proposal of Goldsmith and Xanthos (2006). If we ask the question, what partitioning of the segments of Finnish into two categories, C_1 and C_2, maximizes the probability of the data, given that each category assigns a probability distribution over the segments, and only two independent variables are allowed for transition probabilities (the probability of transition from C_1 to C_2, and the probability of transition from C_2 to C_1), the answer turns out to be: one category consists of all the vowels, and the other consists of all the consonants. This is a reflection of the fact that in any language where there is a preference for vowel-consonant alternation such a division of segments is very likely to maximize

elements smaller. However, if the model were applied to future data in which an f occurred in this context the model would not be able to recognize/represent it at all.

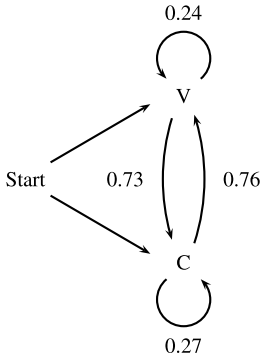

Fig. 2 A vowel/consonant HMM

the probability of a corpus subject to the constraint that the probability is computed by a two-state first order Markov model.[8]

There is a well-known algorithm for hidden Markov models (HMMs) that determines the optimal parameters for the emission and transition parameters that maximizes the probability of the data, and this algorithm quickly learns to assign the task of generating the vowels to one of the states, and the task of generating the consonants to the other. In addition, each of the two states assigns a higher probability to the option of shifting to the other state than to the option of staying in the same state; in short, vowels and consonants prefer to alternate, and this is easily learned. This is illustrated in Fig. 2.

Normally, one would expect there to be symbols that both states emitted with a non-zero probability. Interestingly, that is not what we find here. The data of Finnish forces the conclusion that the highest-probability model assigns, for each phoneme, a positive probability of emission from one of the states, and a probability that is negligibly far from zero for the other state: a very unambiguous categorization of the segments into two non-overlapping sets.

If a second wave of analysis takes the categories induced in the first wave and applies the same HMM learning algorithm to the stream of vowels alone we obtain the HMM in Fig. 3. Unlike the results of the first HMM learning step, the second stage does not neatly partition the vowels into two disjoint sets. Instead, the strictly front vowels $\{\ddot{a}, \ddot{o}, y\}$ are definitively associated with one state, while the strictly back vowels $\{a, o, u\}$ are associated with the other (though with o a lot less certain about it than a or u), but the neutral vowels are associated with almost equal probability to both states in the HMM. This distribution is illustrated in Table 5. The numbers that are presented there are the (base 2) logarithms, for each symbol, of the ratios of

[8]There is nonetheless an unexpected substantive point to note here. The gross generalization that consonants prefer to transition to vowels, and vice versa, could have been modeled in one of two ways, in view of the fact that there are many consonant-consonant transitions, such as *st* and a number of geminate consonants. The system might have allowed both states to generate *s* and *t*, and maintained transition probabilities of *State* 1 → *State* 2 and that of *State* 2 → *State* 1 as 1.0 (or very close to it). Indeed, the system, in learning, often stays very close to that system for quite a few learning iterations. However, it eventually decides to increase the probabilities of staying in the same states (that is, pr(*State* 1 → *State* 1) and pr(*State* 2 → *State* 2)), and dividing the segments up very strictly between the two states, so that the vowels have a zero probability of emission from the consonant state, and vice versa.

Fig. 3 A vowel-feature HMM

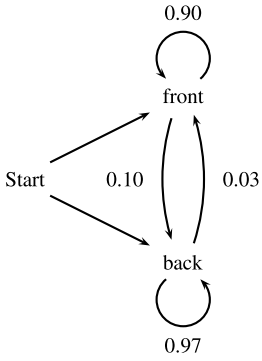

Table 5 Log ratios of emission probabilities for Finnish vowels

Vowel	Log ratio	Vowel	Log ratio	Vowel	Log ratio
ä	961	i	0.148	a	−927
ö	999	e	0.655	u	−990
y	309			o	−7.66

the probability of emission by the 'front' state to the probability of emission by the 'back' state. For example, the probability of the front vowel state emitting *ä* was 2^{961} (about 2×10^{289}) times more likely than the probability that the back vowel state emitted it. In the case of the vowel *o*, it was $2^{7.66}$, or approximately 200, times more likely that the back vowel state emitted it than that the front vowel state should emit it. On the other hand, the ratio of the probabilities of emission for the vowels i, e was not lop-sided, and was in fact quite close to 1-to-1. The ratio of the probabilities (back to front) for i is 1 : 1.1, and for e, the ratio is 1 : 1.57. The vowels are neutral, with a very small bias towards the front.

Another striking difference in the generation of the vowel-feature HMM is that the transition probabilities are quite the opposite of what was found in the prior case: the probability of staying in one state is much higher than the probability of shifting to the other. That is the nature of a harmony system. In particular, the transition probabilities are given in Fig. 3 and the log ratios of the emission probabilities for the vowels are given in Table 5. A harmony system is (essentially by definition) a two state finite state device in which the transition probability from each state to itself are greater than 0.5; the closer these transition probabilities are to 1.0, the closer it is to a perfect harmony system.

Thus the inference of categories like consonants and vowels, as well as the inference of the categories of front vowel, back vowel, and neutral vowels in Finnish, can be obtained by means of the methodological principle of maximum likelihood. The two-state HMM modeling discussed here based on the proposal in Goldsmith and Xanthos (2006) is merely one of many possible approaches to inducing these categories. To take the issue of categories to an even more concrete level, one could start from acoustic signals and create categories of segments and features using a strategy like the one proposed by Lin (2005). An initial categorization into segments was

Table 6 Mutual information on the vowel tier

V_1 \ V_2	ä	ö	y	e	i	a	o	u
ä	2.23	1.54	1.43	−0.35	0.14	−3.00	−2.64	−3.18
ö	1.29	2.54	1.53	−0.16	0.32	−1.23	−0.89	−1.29
y	1.61	3.42	2.35	0.19	0.21	−2.96	−2.71	−3.73
e	0.30	−1.85	−0.30	0.23	0.23	−0.67	−0.54	−0.20
i	0.02	0.51	−0.35	0.62	−0.15	0.11	0.06	−0.55
a	−3.54	−5.27	−3.33	−1.03	0.07	0.45	−0.36	0.02
o	−3.34	−4.33	−2.30	0.10	0.82	0.18	0.29	0.06
u	−2.93	−3.65	−2.22	0.15	−0.11	0.15	0.89	1.12

obviated in our analysis by the use of written corpora which, obviously, come pre-processed according to a (tacit) theory that segments Finnish into a set of symbols.

3.4 Building a tier-based bigram model of Finnish

The problem of inferring categories over which to build the structures used in phonological analysis (probabilistic or otherwise) is a deep and interesting one, but is mostly orthogonal to our main purpose, which is to show how probabilistic models can encode non-local dependencies and how they are to be evaluated. We say *mostly* orthogonal because the number of categories over which the models are stated does come into play in two specific ways. First, if two models A and B differ only in that the latter sub-divides categories of the former, then the MLE probability that B assigns to the corpus on which it was trained will always be better (or just as good) as that assigned by A. Second, the number of categories also plays a role once we begin to evaluate the trade-off between coverage of the data and the complexity of the model itself. We will return to the issue of model complexity in Sect. 4.4.

Given our eight-way distinction on the vowel tier, the traditional front/neutral/back categorization is represented by the shaded quadrants of Table 6. Vowel harmony is reflected by the positive MI for the harmonic front-front and back-back pairs in the upper-left and lower-right quadrants of Table 6. The dispreference for disharmonic back-front and front-back pairs is reflected by the negative MI for the disharmonic pairs in the lower-left and upper-right quadrants. Positive and negative MI values respectively encode increase and decrease in the probability of a segment in a particular environment when compared to that segment's unigram probability. For example, a front vowel increases the probability that the next vowel is front and decreases the probability that the next is back.

There are two striking features of Finnish vowel harmony that are made clear by the values in Table 6. The first is that the influences between categories are not symmetrical and the second is that the categories are not uniform. Regarding the first point, consider the MI values in Table 7 for vowel pairs in the three categories.

Keeping in mind the fact that the MI values represent deviation from the unigram probabilities, there are four immediate generalizations about $V_1\ C^+\ V_2$ sequences:

Table 7 Average MI among categories

V_1 \ V_2	front	neutral	back
front	1.99	0.06	−2.40
neutral	−0.28	0.23	−0.30
back	−3.43	0.00	0.31

(10)
i. if V_1 is back then the probability that V_2 is front is reduced
ii. if V_1 is front then the probability that V_2 is back is reduced
iii. if V_1 is front then the probability that V_2 is front is increased
iv. if V_1 is back then the probability that V_2 is back is increased

The strength of these generalizations decreases from (i) to (iv), with the strength of the last being on par with that of generalizations about the neutral vowels. On one hand, these generalizations could be seen as a reflecting the characterization in Goldsmith (1985) of Finnish vowel harmony as an instance of front harmony. On the other hand, these generalizations offer a nuanced picture of which vowels are more prevalent than expected, and which less, in each environment, generalizations that are outside the scope of autosegmental analyses such as Goldsmith (1985).[9]

We will say nothing here about *why* Finnish vowel harmony shows these specific patterns. Though the question is of indubitable linguistic importance, our current goal is to provide methods for evaluating the accuracy (and accuracy/complexity trade-off) of models of the patterns. It would certainly be interesting to generate data like that in Table 7 for a range of vowel-harmony languages (see, for instance, Baker 2009) or to generate such data for other kinds of Finnish corpora such as a running text or a morphologically decomposed lexicon. This line of research would allow one to ask whether the generalizations in (10) reflect properties of our corpus, properties of

[9]The conception of autosegmental phonology that we employ is that of Goldsmith (1976) and (1990). One of the key ideas in this model is that phonological features are strictly partitioned, and this partition involves their segregation onto separate tiers (though the partition, and hence the segregation, may be different at different levels of the grammar; that is left open as a possibility, and was employed in the work of both John Goldsmith and John McCarthy in the late 1970s and 1980s). There were two broad generalizations that supported this organization: the phenomenon of stability (referring to those cases where a featural specification remains present despite the deletion of the segment to which it was associated), and the many-to-many association patterns widely observed in tonal systems. (This conception is the most widely adopted interpretation, though it is distinct from the projection view of autosegmental representation, suggested by J.R. Vergnaud and others.)

As its name suggests, autosegmental phonology develops a model in which the *autonomy* of separate aspects of a phonological representation is naturally represented. In general, interaction between phonological information on a given tier is restricted to operations that actually affect segments on that tier (while addition or deletion of association lines does not count as 'affecting' a tier in the relevant sense). Thus when it was noted that tones in certain African languages interact with certain voiced consonants (such consonants could add a tone to the tonal melody, or block spreading of a non-Low tone), this behavior was modeled in the framework by explicitly inserting a Low tone on the tonal tier, associated with the voiced consonant in question. Interaction with tone after that would be unproblematic, within the framework. See, for example, Kisseberth (1984), Laughren (1984), Bradshaw (1999), or more recent discussion in Downing (2008).

Table 8 Counts and frequencies for the unigrams on the timing tier

Phone	Count	Frequency	*plog*	Phone	Count	Frequency	*plog*
V	218614	0.42851	1.22	h	9018	0.01768	5.82
t	47927	0.09394	3.41	j	7048	0.01381	6.18
#	44040	0.08632	3.53	d	3734	0.00732	7.09
s	38567	0.07560	3.73	g	828	0.00162	9.27
n	35072	0.06875	3.86	b	580	0.00114	9.78
l	28060	0.05500	4.18	f	326	0.00064	10.61
k	26064	0.05109	4.29	c	312	0.00061	10.68
m	14815	0.02904	5.11	w	118	0.00023	12.08
r	13540	0.02654	5.24	x	36	0.00007	13.79
v	11487	0.02252	5.47	q	18	0.00004	14.79
p	9970	0.01954	5.68				

Finnish, or properties of vowel harmony more generally. The pursuit of these questions must, however, follow the development of tools for generating the data that will be used to answer them.

The second striking feature of the data in Table 6 is the non-uniformity of the MI values within categories. Collapsing the categories as in Table 7 shows that harmony is more robust among front vowels than back vowels but it fails to capture several nuances such as the differences between the strongest and weakest pairs in each category, the fact that $MI(a; o)$ is actually negative, and the fact that some pairs containing 'neutral' are not so neutral (e.g., contrast $MI(e; \ddot{o})$ vs. $MI(\ddot{o}; a)$, $MI(\ddot{o}; o)$, and $MI(\ddot{o}; u)$). In light of these facts, we will adopt the categories C and V in the analysis to come, but we will return to the issue in Sect. 4.4.

3.5 Applying the tier-based bigram model to Finnish

Having a model of the vowels in hand, what remains is to evaluate the rest of the language—what we shall refer to as the *timing tier*. The timing tier is essentially identical to the bigram model of Finnish as described in Sect. 3.2 save for the fact that all of the vowels have been collapsed into a single symbol 'V'. This symbol will function as a place holder on the timing tier for the vowel information recorded on the vowel tier. We speak of this collapsing as if it were 'dividing' the original corpus by the set composed of the vowel symbols: finding the 'quotient' amounts to replacing the different vowel symbols with the cover symbol V.

Collapsing vowels down to a single symbol (using the 'quotient') yields a higher probability for every bigram containing V on the timing tier, by comparison with the original bigram model. For a given word w, if we replace all of the individual vowels in it by the symbol V, we write the result as $w \div V$. For instance, in the basic bigram model the positive log probability of a and e immediately after b are 2.11 and 2.69, respectively. Once the vowels have been collapsed on the timing tier, however, the positive log probability that V follows b is 0.319—quite a significant decrease. The unigram counts and the plogs for the timing tier are given in Table 8.

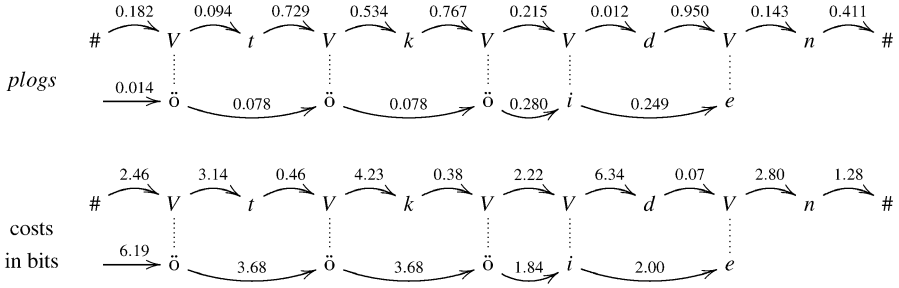

Fig. 4 A probabilistic autosegmental model of *ötököiden*

What remains, then, is to combine the two tiers to create a single model. Figure 4 illustrates the way that the tiers come together to form a probabilistic model of Finnish. The probability of a word, in this autosegmental model, is the product of the probability of the quotient string (that in which vowels have been collapsed to V) and the probability of the sequence of vowels. For a given word w, we indicate the string of vowels that it contains as $w \approx V$ (e.g., $(katab \approx V) = aa$). To compute the vowel tier probability $\text{pr}(w \approx V)$, we take the product of the unigram probability of the first vowel (because it is not preceded by a vowel) and the conditional probabilities of each subsequent vowel. This yields the expression in (11) of word probability as the product of the quotient string and vowel tier.

$$\text{pr}_{auto-V}(w) = \text{pr}_{bigram,tier1}(w \div V) \times \text{pr}_{bigram,tier2}(w \approx V) \tag{11}$$

In Fig. 4 we illustrate the probabilities and concomitant plogs that the autosegmental model assigns to the word *ötököiden*. Using plogs is especially helpful in this case because the probabilities are so small; the product of the probabilities on the timing tier is 7.62×10^{-8}, the product of the probabilities on the vowel tier is 5.79×10^{-6}, and the product of both tiers is 4.42×10^{-13}.

Representing Fig. 4 with plogs on the arcs is straightforward because the positive log of conditional probability is the joint log probability minus the log of the unigram probability, which is computable by subtracting the unigram plogs from the bigram plogs. Thus the bit cost of the first arc on the timing tier is $plog(\#V) - plog(\#) = 5.99 - 3.53 = 2.46$ which is precisely the value obtained by taking the positive log of the conditional probability: $-1 \times \log 0.182 = 2.46$. The sum of the costs on the timing tier is 23.64 bits, and the sum of the costs on the vowel tier is 17.40 bits. Taken together the cost of representing the word *ötököiden* is 41.04 bits (which is another way of saying $-1 \times \log 4.42 \times 10^{-13}$).

After collapsing the vowels, the average plog per segment on the timing tier is 2.95, and with 510,174 segments (the same number as in the bigram model), the total cost of the word-set on the timing tier (i.e. the sum of the plogs) is 1,273,648 bits. Compared with cost of the bigram model of 1,780,278 bits, we see that collapsing the vowels makes the cost of the timing tier about 28 % less than that of the bigram model. We must, however, add the cost of the vowel tier because the timing tier alone omits the vowel qualities. The average MI among bigrams on the vowel tier is 0.23 bits. This means that, among vowels, knowing the quality of the preceding vowel

reduces uncertainty about the next by about 8 %. Overall, the cost of the corpus on the vowel tier is 540,822 bits. This yields a total cost of 1,814,470 bits for the corpus in the autosegmental model. Unfortunately, this is actually higher than the cost under the bigram model.

3.6 Local C-to-V MI exceeds distal V-to-V MI

The idea behind our first tier-based model is simple—collapsing all of the vowels on the timing tier makes all of the strings on that tier more probable while exposing the non-local cases of V-to-V MI on the vowel tier makes all of the strings on that tier more probable as well.[10] If there were no neutral vowels in Finnish, and if vowels were distributed so as to respect vowel harmony but otherwise uniformly (i.e. without regard for surrounding consonants), then we would expect knowledge of vowel harmony to decrease the information present in a word with n vowels by about $n - 1$ bits because the choice of each vowel after the first would be made from a set that was only half the size of the full vowel system.

There are many words in our Finnish corpus where the results go in this direction. For instance, *ötököiden* has a cost (plog sum) of 43.64 bits in the bigram model but a cost of 41.04 in the autosegmental model—an information improvement of about 6 %. However, it turns out that about 64 % of the words in the corpus are actually assigned a lower probability under the basic autosegmental model than the bigram model. In the aggregate, these overwhelm the increase in probability for harmonic forms. In (12) we give the overall results of the three models.

(12) Unigram model: 2,088,528 bits

Bigram model: 1,780,267 bits

Autosegmental: 1,814,470 bits = 1,273,648 (timing) + 540,822 (vowels)

The failure of this basic version of an autosegmental model is due mostly to the fact that it has collapsed too many distinctions on the timing tier. Many of the words that are assigned lower probability in the autosegmental model (even some that are highly harmonic) contain highly probable VC and CV pairs whose mutual information is occluded when the vowels are collapsed down to a single symbol on the timing tier. Another deficiency of the model (much less significant in its overall effect) is that the vowel tier lumps diphthongs together with vowel pairs in adjacent nuclei. In the former case, seven of the vowel pairs in the harmonic category actually have negative MI if we consider only strictly adjacent vowels and thus conflating these two cases yields a poorer account of each.

The crux of the problem with totally segregating the tiers is revealed if we rank the bigrams of the basic bigram model in term of the MI that they contribute. Table 9 lists the top twenty Finnish bigrams ranked by *weighted mutual information*, where wMI = MI × count. Using wMI provides a rough metric of the utility of each bigram in the model because it counts as most useful a bigram that has high MI *and* is

[10] This first model also provides a particularly simple way to ensure the well-formedness of the probability distributions in that the sum of the probabilities of the set of strings that map to any given template such as #kVtVb# will be the same in the bigram model and the two-tier model.

Table 9 Top 20 Finnish bigrams by weighted MI (wMI = count × MI)

Bigram	Count	MI	wMI	Bigram	Count	MI	wMI
n#	14422	2.25	32479.13	in	7521	1.13	8484.72
a#	12275	1.33	16377.40	#v	3622	1.87	6769.37
en	9019	1.81	16343.02	ko	3521	1.64	5778.19
is	10551	1.48	15609.93	ma	4204	1.36	5717.76
st	9743	1.43	13904.72	el	4651	1.18	5480.94
ll	5476	1.83	10005.10	an	6760	0.80	5422.02
ta	10345	0.97	9987.00	tu	4994	1.07	5345.63
#p	4207	2.29	9631.09	mi	3705	1.35	5001.56
#k	6234	1.47	9165.69	se	5360	0.92	4954.16
va	4632	1.87	8647.96	ää	2172	2.28	4953.32

Table 10 Performance of the three models on the test data divided into 12 sets

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
Unigram	79,441	79,407	80,018	80,025	80,492	79,530	80,521	81,109
Autoseg.	70,174	70,250	70,591	70,785	71,114	70,474	71,108	71,847
Bigram	69,111	69,129	69,548	69,468	69,853	69,174	69,764	70,490

	Test 9	Test 10	Test 11	Test 12	Sum	Mean	Variance
Unigram	80,731	79,718	79,097	78,197	958,286	79,857	642,241
Autoseg.	71,487	70,533	70,003	69,179	847,545	70,628	508,038
Bigram	70,288	69,342	68,881	67,970	833,018	69,418	442,053

prevalent in the data. Table 9 reveals that 13 out of the top 20 bigrams are either VC or CV. Thus, even though collapsing the vowels increases the probability of the timing tier, the separation of the vowels from the consonants hides the influence that the consonants have on the vowels and vice versa.

3.7 Evaluating the differences

In order to be confident that differences among the models are robust and not an artifact of over-fitting, we evaluate their performance using only half of the data (selected randomly) for training and split rest into a dozen batches for testing. To prevent accidental gaps from compromising the models' probability distributions we add half a count for each structure type (i.e., ELE smoothing) in training the models. The bit-costs of each of the twelve test-sets under the Unigram, Autosegmental, and Bigram models are given in Table 10.

For each of the batches of test data, the Bigram model fares better than the Autosegmental model, which in turn fares better than the Unigram model. Evaluating the differences in the models' performance with a paired-sample Wilcoxon signed-rank test (a non-parametric test) yields a Wilcox rank sum test statistic V of 0 with

a p-value of 0.0004883 for each pair of models indicating that the differences are highly significant in each case.

Overall, the fact that this first-pass autosegmental model does a significantly worse job predicting the data than the basic bigram model shows that, while it may be useful to incorporate mutual information between vowels across consonants, doing so by occluding the mutual information between consonants and vowels yields a net loss. In the next section we consider a more nuanced model that can utilize both sources of information.

4 Boltzmann model

4.1 An introduction to Boltzmann models

In Sect. 3, we described our probabilistic autosegmental model of Finnish phonotactics, and how the results showed that the statistical effects of vowel harmony on segment sequencing were overall slightly weaker than the aggregate of the effects of surrounding consonants on vowel quality. This result went counter to our expectations, but once we take note of the fact that our expectations were not met, it is instructive to see where our expectations came from, and it is critical to deal with a more complex linguistic reality.

Our expectations were based on the incorrect assumption that consonants do not significantly 'choose'—that is, condition—the vowel that immediately follows (or at the very least that any such effect should be weaker than the influence of vowels on one another in a vowel harmony language like Finnish). However, the fact of the matter is that linguistic structures, like most structures in the natural world, show complex partial dependencies, and we need a probabilistic model that is capable of dealing with such cases.[11]

One widely used model of this sort employs an approach which goes by a number of names, including the Boltzmann distribution. The heart of the Boltzmann model is the idea that a probability is assigned to a representation on the basis of a score, and the difference between the scores of two representations R and S is equal to the ratio of the probabilities assigned to R and S—or to put it another way, the difference in the scores of R and S is equal to the difference of the log probability of R and S.[12] This way of putting it makes it clear that if the score that is assigned has some

[11] For current views on consonant-vowel interactions and on consonant-vowel harmony systems see Padgett (2011) and Rose and Walker (2011).

[12] See Geman and Johnson (2001), for example, which is an excellent introduction for the relevance of this notion to linguistics. The notion arose in the context of statistical physics, where it is natural to define a notion of energy ϵ_i for each state that an object may be in, and then assign a probability to being in that state which is proportional to $2^{-\epsilon_i}$. In order to make these values a distribution, they must be normalized, and so one generally indicates the probability in an expression of the form $\frac{2^{-\epsilon_i}}{Z}$, where Z is the sum, for all i, of $2^{-\epsilon_i}$. In interesting cases, the model may be modified so that the influence of the differences of energy may be attenuated by introducing a notion of temperature t in a modified formula for probability of being in state i: $\frac{1}{Z}2^{-\frac{\epsilon_i}{t}}$.

In work on computational learning, the notion of a conditional random field has been explored by Lafferty et al. (2001), of which the present model is a special case; we return to this in Sect. 5.2.

linguistic meaning, the probabilistic model that the Boltzmann model creates is one in which probability is tightly linked to the score.

It is traditional to define the score in such a way that the larger the score of a representation is, the smaller is its probability. This convention is encoded in the presence of the negative sign in the exponent of 2 in (13). Thus the score in a Boltzmann model should be thought of as a measure of phonological ill-formedness.[13] For a model m and a sample space U_1 of possible phonological representations, each element $r \in U_1$ in the sample space is assigned a score $h(r)$ by m. This, in turn, yields an *exponentiated score* of $2^{-h(r)}$ that can be turned into to a probability in a well-formed probability distribution. To do this *normalization*, each $2^{-h(r)}$ is divided by Z, the sum of the exponentiated scores of all elements of the sample space.

$$\text{(13)} \qquad \text{pr}_B(r) = \frac{1}{Z} 2^{-h(r)} = \frac{2^{-h(r)}}{\sum_{s \in U_1} 2^{-h(s)}}$$

The main substance of any model thus consists of how the score function h is defined. A wide range of possibilities is available. For example, if we define the score of a representation to be the sum of the scores of the individual segments, and define the score of an individual segment as the *plog* of its unigram frequency, then the probability assigned to a representation is just its familiar unigram probability. This can be seen in (14) where the calculations show that when the score is directly based on log probabilities, it is natural that exponentiating that value should give us back probabilities, and the denominator, Z, sums to 1, as it actually sums all the probabilities of the unigram sample space.

$$\begin{aligned}
\text{(14)} \qquad \text{pr}_B(r) = \frac{1}{Z} 2^{-h(r)} &= \frac{2^{-h(r)}}{\sum_{s \in U_1} 2^{-h(s)}} \\
&= \frac{2^{-\sum_{i=1}^{|r|} plog(\text{pr}(r_{[i]}))}}{\sum_{s \in U_1} 2^{-\sum_{i=1}^{|s|} plog(\text{pr}(s_{[i]}))}} \\
&= \frac{\prod_{i=1}^{|r|} 2^{\log \text{pr}(r_{[i]})}}{\sum_{s \in U_1} \prod_{i=1}^{|s|} 2^{\log \text{pr}(s_{[i]})}} \\
&= \frac{\prod_{i=1}^{|r|} \text{pr}(r_{[i]})}{\sum_{s \in U_1} \prod_{i=1}^{|s|} \text{pr}(s_{[i]})} \\
&= \frac{\text{pr}(r)}{\sum_{s \in U_1} \text{pr}(s)} = \frac{\text{pr}(r)}{1} = \text{pr}(r)
\end{aligned}$$

[13] The intent of the notion of well-formedness described in Goldsmith (1990), Goldsmith (1991), Goldsmith (1993) was to be -1 times this quantity, and the aim of that analysis was to show that level-internal phonological processes always correspond to a decrease in ill-formedness. Those references failed to offer an explicit way to calculate the 'phonotactics'; the present paper offers the expressions that calculate the plog of a representation as the correct method for calculating such phonotactics.

But the beauty of a probabilistic model such as this is that it allows a wider range of freedom than simply to use the log probability of an element as its score. We could, for example, set up a list of regular expressions c_i, each associated with a weight, and then assign a score to a representation which was equal to the sum of the weights associated with each expression c_i. If each c_i modeled some characteristic that the system tries to avoid, so to speak, then such a Boltzmann model assigns a probability based on these terms, weighted by the 'strength' of each particular expression.[14]

Returning to the general point, the score assigned to any given representation is the exponential of (-1 times) the weighted sum of the values associated with each phonological 'feature' associated with (or simply found in) a given representation. By 'feature,' again, we do not mean simply phonological features in the usual sense (although these could be features in the present sense), but a set of features selected from any property at all that can be measured.

4.2 A Boltzmann model for two tiers

For the case of Finnish, we propose to use three sources for features in the Boltzmann scoring: the unigram positive log probabilities of the segments (a quantity often referred to as *self-information*), the mutual information between pairs of consecutive segments, and the mutual information between non-adjacent vowels. (We return below to the question of whether this decision is made on a language-particular basis or more generally.)

$$U(w) = \sum_{i=1}^{|w|} \log \mathrm{pr}(w_{[i]}) \tag{15}$$

$$M_1(w) = \sum_{i=1}^{|w|} MI\left(w_{[i-1]}^{tier1}; w_{[i]}^{tier1}\right)$$

$$M_2(w) = \sum_{i=1}^{|w|} MI\left(w_{[i-1]}^{tier2}; w_{[i]}^{tier2}\right)$$

$$Score(w) = U(w) - M_1(w) - M_2(w)$$

The reader will recall that the first term in (15), alone, expresses the unigram model probabilities, and that the first two terms together express the bigram model. It is thus the presence of the third term that incorporates the analysis of vowel harmony into the model. What we propose to do is sketched in Fig. 5.

The solid arrows in Fig. 5 give the plogs of the conditional probabilities in the basic bigram model, which are equal to the plog of the unigram probabilities minus the mutual information of the bigrams. Given this base value, we then subtract a second line of mutual information for non-adjacent vowels to create our Boltzmann model.

[14] Proposals along these lines have been made by Goldwater and Johnson (2003) and Hayes and Wilson (2008) where expectation maximization is used to find optimal weights over each c_i, and by Wilson (2006) who puts the various c_i together in a conditional random field to create a model whose structure is, in many ways, similar to what we propose in Sect. 4.2.

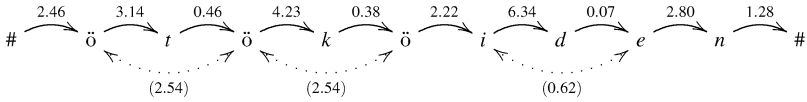

Fig. 5 Plogs of conditional probabilities +V–V mutual information

We use the second source of MI only for non-adjacent vowels because interactions among adjacent vowels are already captured by the bigram model.

In both models, we compute the statistical connections between items and their neighbors by means of computing mutual information. Consonants thus have only one neighbor (to their left or right), but some vowels have two neighbors (on each side): a 'local' neighbor and a more distant neighbor, a notion that is in effect modeled by the autosegmental representation.

4.3 Computing Z

In light of the discussion in the previous section, we calculated a set of scores for the words in the Finnish corpus described above. Each word w's score, $h(w)$, is equal to the sum of the plogs of its phones, less the mutual information between successive phones and the mutual information between successive non-adjacent vowels. Each word was assigned an exponentiated score $2^{-h(w_i)}$, which is transformed into a probability by division by Z, the partition function.

Computation of Z is often the hardest part of developing a Boltzmann model. Happily, the models that we are working with obey a simple structural restriction that makes it possible to compute Z relatively easily. Even though the harmony component of our model is recursive (i.e. can operate over arbitrary distances), it can nonetheless be encoded as a simple weighted finite state automaton (wFSA).

We represent the harmony component of the model as a wFSA, H, whose arcs are labeled with mutual information scores for vowel pairs. We represent the bigram component of our model as a wFSA, B, whose arcs are weighted with positive logs of conditional probabilities. Using these representations, it is possible to construct a new wFSA, $B \times H$, by intersecting the structures of B and H (cf. Hopcroft and Ullman 1979) and assigning new weights to the arcs by subtracting the V–V MI from the plog of the conditional probability for any vowel that has a vowel antecedent. This new machine represents both segment-to-segment interactions, and distal vowel-to-vowel interactions within a single weighting function. The only wrinkle is that, after the intersection, the weights on the arcs of the new wFSA no longer represent a well-formed probability distribution. To recover a probability distribution from the new model we must sum the weight assigned to every possible representation $r \in U_1$; that is, we must compute Z.

It is not possible to compute Z by incrementally summing weights because there are infinitely many possible phonological representations in U_1. This is where the use of a finite-state representation for our linguistic forces is most useful. Because the combined wFSA $B \times H$ is a finite encoding of the weighting of the infinite range of representations, we can 'solve' the weight that the dynamic system assigns to U_1 by a recursive computation on $B \times H$ in the manner described by Eisner (2002). Following this procedure, we find that $Z \approx 1.0177$ for the MLE model with probabilities

Table 11 Performance of the Bigram and Boltzmann models on the 12 sets of test data

	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6	Test 7	Test 8
Bigram	69,111	69,129	69,548	69,468	69,853	69,174	69,764	70,490
Boltzmann	68,380	68,429	68,850	68,800	69,155	68,491	69,060	69,821

	Test 9	Test 10	Test 11	Test 12	Sum	Mean	Variance
Bigram	70,288	69,342	68,881	67,970	833,018	69,418	442,053
Boltzmann	69,575	68,653	68,207	67,283	824,704	68,725	441,487

generated from the whole corpus.[15] The fact that Z is quite close to 1.0 means that the exponentiated scores need only be reduced slightly to yield a well-formed probability distribution, and hence the probability of the corpus (which is what we are most interested in calculating) is enhanced by this model.

At a more abstract level, what this means is that the exponentiated score that we have calculated, if summed over the entire space of representations, would be slightly more than 1—alternatively put, by and large, and on the average, adding mutual information between vowels to our scoring function does not help to improve words generated randomly (Finnish words are not random). Hence, the fact that including mutual information in the scoring of the real Finnish data did improve the exponentiated score is highly non-trivial. Put more simply, the relationship between the vowels in the Finnish corpus increases the probability of the data, and thus should be captured in a statement of Finnish phonology. In (16) we give the bit-costs for the models.

(16) Unigram model: 2,088,528 bits

Bigram model: 1,780,267 bits

Boltzmann: 1,760,523 bits

The improvement offered by the Boltzmann model is relatively small when compared to the difference between the Unigram model and the Bigram model. Nonetheless, the differences between the models are significant. A paired-sample Wilcoxon signed-rank test yields a Wilcox rank sum test statistic V of 0 with a p-value of 0.002516. The bit-costs of the data in the test-sets under the Bigram and Boltzmann models are given in Table 11.

In evaluating what appears to be a relatively small improvement over the bigram model, one needs to keep in mind the fact that the bigram model encompasses *all* segment-to-segment effects (including those among vowels) while the harmony tier encodes only the non-local interactions among vowels. To put this in perspective, the harmony tier has $8 \times 8 = 64$ free parameters to track non-local V–V pairs while the bigram tier has $28 \times 28 = 784$ free parameters to track all pairs. The bigram model replaces the 28 free parameters of the unigram model with 784 free parameters and in so doing decreases the cost of the corpus by about 14.8 %. By contrast, the Boltzmann

[15] For the ELE model, if probabilities are generated from the whole corpus then $Z \approx 1.0175$ and if probabilities are generated from the training data used in our comparisons then $Z \approx 1.0177$.

model adds 64 free parameters on top of the bigram model and in so doing decreases the cost of the corpus by about 2.2 %. In the former case the number of parameters differ by a factor of 28 and in the latter by a factor of ∼1.08.

4.4 Model complexity

One of the most relevant properties of the model that we have proposed is that it is a synthesis of second order models rather than a third (or higher) order model. One could conceivably try to capture Finnish vowel harmony by casting an ever wider net using 3-grams, 4-grams, 5-grams, and so on with exponentially larger numbers of probabilistic parameters. However, the alternative that we advocate here is simply the addition of another relatively compact model alongside the bigram model that allows vowels to act on each other at arbitrary distances. By that same token, instead of trying to capture harmony with a trigram-like model in which each consonant is split into eight categories according to its preceding vowel, we obtain a smaller model by treating the harmony tier and bigram tier as independent.

In developing an approach of the sort described here, it is critical to be able to express quantitatively the complexity of the formal model used, and there needs to be some explicit price 'paid' for increasing the complexity of a model. The reason for this is somewhat complicated, but, for our current purposes it suffices to note that the notion of complexity measure in Chomsky (1956/1975) is closely related to this trade-off between model complexity and fidelity of description, as is the notion of algorithmic complexity.[16]

A crucial insight of the theory of algorithmic complexity (see Li and Vitányi 1997) is that the units that are used to measure a grammar's length are the same units used to measure the logarithm of the probability of an object (or rather, the plog): in both cases, we use bits as our units. In light of this, we recognize that the choice for the language-learning system is not necessarily, "should I include a bigram model of my phones in the phonological account?" but rather, "for how many, and for which, pairs of segments, or categories of segments, should I keep track of the relevant mutual information statistics?"

The way to answer these questions that we believe is the most linguistically illuminating is embodied by the Minimum Description Length Principle. Using MDL we select the $h \in H$ that minimizes:

$$h_{MDL} = \arg\min_{h \in H} L_{C_1}(h) + L_{C_2}(D|h) \tag{17}$$

where $L_C(x)$ is the description length of x under encoding C. Because it is possible to encode the data D using n bits where n is the positive log of the probability assigned to D by h, this is an expression of Bayes rule $\arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)}$ under the assumption that the prior probability of hypothesis h is the number of bits that it takes to encode h. In an idealized sense, this approach embodies a *Kolmogorov-prior*

[16] Various approaches have been suggested to the task of assigning a prior probability over models. One approach, incorporated into Minimum Description Length models (Rissanen 1989), assigns a probability of $2^{-|m|}$, where $|m|$ is the length of the grammar in some appropriately compact formulation.

Table 12 Total costs of models and data

Model	# Param.	Model cost	*Cost*(data|model)	Total cost
Unigram	28	448	2,088,528	2,088,976
Bigram	784	12,544	1,780,267	1,792,811
Boltzmann	848	13,568	1,760,523	1,774,091
VC-combo	5,040	80,640	1,710,499	1,791,139
Trigram	21,952	351,232	1,553,644	1,904,876

wherein we assume that the prior probability of a hypothesis is the reciprocal of 2 raised to its complexity (length) in bits. Unfortunately, Kolmogorov complexity is not generally computable, so we are left to propose encoding schemes for classes of hypotheses and then work within those.

One of the most basic properties of models is their number of free parameters. We can get a base-line cost for representing the models under consideration here by calculating the cost of representing the parameters. (These are really families of models with the same structure but different parameter values.) Assuming that 16 bits will allow sufficient precision in representing the probabilities in the models, we will need 16 bits for each parameter in the model.[17] This gives us a rough estimate of the cost of each model. In Table 12, we add the costs of various models to the (MLE) cost that each model assigns to the corpus. For comparison we also include a trigram model and a model *VC-combo* in which the bigram and harmony tier are not treated as independent.

What Table 12 shows is that even though moving to a trigram model would do a better job compressing the corpus, this improvement would be overshadowed by the cost of encoding the model itself. Because the Boltzmann model is a linear combination of the bigram model and the harmony tier, which has 64 free parameters, we need only add $64 \times 16 = 1{,}024$ bits to the cost of the bigram model to reduce the bit-cost of the corpus by about 20,000 bits.

4.5 Model complexity revisited

Baker (2009) applies the approach discussed in this paper to several languages including Finnish, and he proposes a range of modifications and innovations to the proposals here, including a strategy for evaluating each parameter of a model in terms of the degree to which it compresses the corpus. Baker found that the neutral vowels on the harmony tier were not particularly useful in this regard, and suggested that they be effectively removed. This revision to the Boltzmann model actually improves the compression of the corpus to 1,755,148 bits—a 1.1 % savings—and does so with 28 fewer parameters (i.e. the harmony tier tracks only $6 \times 6 = 36$ pairs). Because the neutral vowels of Finnish are transparent to harmony (see Ringen 1975/1988), removing them allows the model to capture interactions among non-neutral vowels at greater distances. This provides a perfect example of the fact that probabilistic models and traditional linguistic structures are not at all antithetical.

[17] This assumption is very generous to large models because most of the probabilities we observe require more than the 4-decimal-place fidelity that can be recorded with 16 bits.

Table 13 Front/back counts and front/back mutual information

V_1 \ V_2	Front	Back	*Total*
Front	9,753	2,575	12,328
Back	2,045	54,963	57,008
Total	11,798	57,538	69,336

V_1 \ V_2	Front	Back
Front	2.22	−1.99
Back	−2.25	0.22
wMI	1.46	0.12

Given the fact that removing the category of neutral vowels from our harmony tier reduces the number of parameters in the model without reducing the probability assigned to the corpus, it is natural to ask whether collapsing the front or back vowels down to a single category would offer similar benefit. Considering only the strictly-front and strictly-back vowels, $\{\ddot{a}, \ddot{o}, y, a, o, u\}$, our Finnish data contains 69,336 pairs of vowels, (V_1, V_2), separated by at least one consonant. In 17 % of the instances V_2 is front and in the other 83 % of the instances V_2 is back. When the cases are separated into groups based on the color of V_1, however, quite a reversal is revealed; though back vowels are about five times more common in general, they are actually four times *less* common than front vowels when V_1 is front and about 25 times more common than front vowels when V_1 is back.

Table 13 offers insight into why the effects of adding vowel harmony to the model are rather small despite the fact that the generalizations are fairly robust. Most of the vowels that could be subject to harmony are back and the vast majority of these occur, as expected, following back vowels. In the environment $V_1C^+V_2$, the probability that V_2 is back is 0.83, and when V_1 is back, this jumps to 0.96, but the difference between *plog*(0.83) and *plog*(0.96) is only about 0.22 bits. Overall, saving 0.22 bits in describing the fifty-five thousand back vowels that occur after back vowels while spending an extra 2 bits to describe the front vowels that occur after back vowels yields an average savings of 0.12 bits per vowel in this environment. The average savings are much greater in describing vowels that follow front vowels but there are far fewer of these.[18]

A similar set of generalizations can be made about the interactions of the eight vowels on the harmony tier introduced in Sect. 3.4, but they are much easier to see once the vowels are collapsed into just two groups. The critical question from the perspective that we advocate here is whether using 4 rather than 64 parameters to describe the vowel interactions is superior in terms of the cost of the model plus the cost of the data given the model. We contrast a Boltzmann model with only two classes of vowels, Boltz-2, with our original model, Boltz-8, in Table 14.

Assuming, as we did above, that each free parameter costs 16 bits, we find that, for this data and this metric of model cost, the reduction in the cost of the model is outweighed by the loss in predictive power.

One might ask how things would have to be different in order for the Boltz-2 model to be the best. This would be especially germane in a case where we had

[18] It is also important to keep in mind that these numbers are for the harmony tier all by itself. Once the harmony tier is included with the bigram tier in the Boltzmann model these values will be normalized to produce a well-formed probability distribution.

Table 14 Evaluating a Boltzmann model with fewer parameters

Model	# Param.	Model cost	*Cost*(data—model)	Total cost
Bigram	784	12,544	1,780,267	1,792,811
Boltz-2	788	12,608	1,765,451	1,778,059
Boltz-8	848	13,568	1,760,523	1,774,091

some independent evidence that the smaller model was somehow the 'right' one. This could happen if, for instance, one were attempting to model experimental data for phonological generalizations made by humans in which there was evidence that the front vowels were treated as a unit for the purposes of harmony. Though we are expressly not trying to model humans' phonological generalizations here (for that we would need very different kind of data), the evaluation metric that we have proposed can be straightforwardly applied to such a task.

One natural way to tip the balance in favor of the Boltz-2 model would be to increase the cost for encoding each parameter by an order of magnitude; our assumption of 16 bits was quite low to begin with. This would work for the data at hand but would likely break down for larger sets of data because the importance of the model's cost diminishes as the data grows.[19] A more interesting approach would be to assume an upper bound on the amount of data that can be taken into consideration when choosing between the models. If, for instance, the decision was made based on a window of the 20,000 most common (or recent) words, then the relative importance of minimizing the size of the model would increase.

This example illustrates but one of hugely many possible groupings of Finnish segments into categories; there is an extensive literature on strategies for doing this kind of grouping (for an introduction, see Kaufman and Rousseeuw 2005). In this work we adopted the categories C and V following Goldsmith and Xanthos (2006) in order to implement our tier-based harmony model and to show that such a model is 'simple' in the right way (i.e. tiers can capture non-local interactions while adding relatively little complexity). The problem of searching the space of models is highly relevant but is beyond the scope of this current paper. All that we will say about it here is that a basic strategy for searching the model-space can be obtained by combining the evaluation metric in (17) with any clustering algorithm and the premise that things in the same category can interact non-locally on a tier.[20]

5 Discussion

In this section, we will comment on some general issues that are raised by the kind of approach that we have envisioned in this paper. The first is the relationship of

[19] This is exactly as it should be; even small deviations from the (unigram) expectations are significant and worth encoding if they hold over large enough sets of observations.

[20] This premise can be seen as an implementation of the idea that harmony (and respectively disharmony) operates over elements that are sufficiently similar (see, for instance, Cole 1987, 2009; Walker 2000, 2005; Hansson 2001; Rose and Walker 2004). The approach in Cole (2009), though couched in an exemplar-based model, is conceptually quite close to what we advocate here.

probabilistic models to generative models and the second involves where this line of research may be taking us.

5.1 Information theory and generative phonology

Our goal in the work described here is to develop a framework of phonological analysis which is explicit enough to algorithmically determine which of a finite set of candidate analyses is the best, given a set of data from a language. In a sense, our approach is entirely within the original framework of generative grammar, though to our knowledge, relatively little work along these lines has actually been carried out since Chomsky (1956/1975) and Chomsky (1957: 52f). Our goal is not to discover that Finnish has vowel harmony—that was known well before there was such a thing as generative phonology—but rather to develop a device that quantitatively and algorithmically substantiates that kind of analysis of a Finnish corpus.

It may seem odd to hear probabilistic models being touted as exemplars of generative grammar, but we have tried to emphasize that probabilistic models are always as formal as any other type of grammar: a probabilistic model is subject to the constraints that each element in U_1 be assigned a probability and the summation of these values (which is typically infinite) must yield the value 1.0. We suspect that the reason some linguists find this odd is that early work on probabilistic models typically used very simple unigram and bigram models, even to handle syntactic phenomena, and probabilistic models were, quite erroneously, identified with the simplest of finite state models. However, there is no reason to limit the application of probabilistic tools to models that are linguistically simple or naive.

It is fair to say that probabilistic models are not widely employed at the center of phonological work in the United States today. Bod et al. (2003) write: "One of the foundations of modern linguistics is the maxim of categoricity: language is categorical. Numbers play no role, or, where they do, they are artifacts of non-linguistic performance factors." As a remark about the foundations of modern linguistics, we disagree; Trubetzkoy (1939/1968), surely one of the founders of modern phonology, wrote, "All the particular properties that give a language its unique phonological character can be expressed in numbers." Along a similar line, Henry Kučera (1982: 167) wrote, "[The] correlation between structure and language statistics is important since it supports the basic notion that the markedness relation, in reflecting an informational economy in language coding, has the expected statistical effects."

Still, probabilistic models do have a strong presence in fields closely related to phonology: in all work today on speech recognition (see Huang and Acero 2001 for a recent overview); in much of the work in computational linguistics (see, e.g., Manning and Schütze 2000 for a recent overview);[21] in some European traditions, notably the work of Gabriel Altmann (1980, e.g.) and others publishing in the *Journal of Quantitative Linguistics*; and a wide range of phonological work by authors writing

[21] Also, Pereira (2000) makes an argument regarding probabilistic models of syntax much in the spirit of what we argue here. He points out that there are many widely held misconceptions about weaknesses of probabilistic models that stem from overgeneralizing the weaknesses of straw-man models in well-known papers. Like Pereira, we aim to correct the misconception that probabilistic models are antithetical to linguistic structure.

in the 1950s and 1960s, including Morris Halle, Roman Jakobson (see, e.g., Cherry et al. 1953, Henry Kučera 1982; G. Herdan 1956, 1960, 1964; see also Halle 1958 for a review of the first), and many others.

Since the 1990s, another wave of scholarship on probabilistic phonological models has emerged that, from a historical point of view, forms an organic whole with the earlier work cited above. Indeed, one could argue that, if a historical fact is in need of explanation, it is the temporary absence of work in this area in the period from the late 1960s until the early 1990s. A partial list of this more recent wave of probabilistic work includes Pierrehumbert (1994, 2001, 2003), Coleman and Pierrehumbert (1997), Seidenberg (1997), Frisch et al. (2000), Albro (2000), Bailey and Hahn (2001)), Dainora (2001), Goldsmith (2002), Billerey-Mosier (2003), Jäger (2004), Goldwater and Johnson (2004), Hayes and Wilson (2008), and Cole (2009) to name just a few. The latter three of these, especially Hayes and Wilson (2008), have many similarities to our proposal here.

We find it very encouraging that our work is among several lines of research that are independently converging on what seems to be a common information theoretic framework for modeling phonological phenomena. There are, however, a few significant differences in our approach that bear mentioning.

First, we believe the distributional properties of the data are sufficient for discovering the categories that are relevant for local and non-local interactions in a given language. Thus, unlike Hayes and Wilson (2008), we do not assume that natural classes or specific tiers for non-local interaction are antecedently given. Hayes and Wilson make this assumption, in part, to restrict the model search-space, which is a more central focus of their work. Thus, we acknowledge that our assumptions may turn out to be problematic when we move on to the next stage of this work, which will involve more model building. Nonetheless, we are encouraged by the fact that the evaluation metric we propose can be used to adjudicate among the clusterings of segments (i.e. classes) produced by any clustering algorithm.

A second central difference, is the way that our evaluation metric incorporates both the complexity of a model and its fit of the data. Hayes and Wilson employ heuristics in evaluating models that also favor low complexity but this complexity is assessed in terms of the natural classes they take as a starting assumption. Though our model-complexity metric of 16 bits per parameter is admittedly crude, we think that it is ultimately more illuminating to demonstrate that 'classical' linguistic structures like natural classes can emerge from distributional patterns in the data. Cole (2009) makes much the same point and goes further to assess theories about why the patterns are there in the data.

5.2 Further directions: towards a phonological field theory

It is natural to ask what the origin is of the surprising aspect of the model in (15) and illustrated in Fig. 5, which is to say: why should there be vowel harmony in a language? We have focused thus far on the use of mathematical tools, particularly those associated with information theory, and we have thought of the generation of a word in Finnish (or any other language) as a stochastic process in which the probability of a given phoneme is conditioned by preceding phones. In this final section, we will

briefly describe a somewhat different view that we think is lurking behind the structure of vowel harmony, and one which attempts to exploit mathematical methods used in statistical mechanics. In this view, the probability of a given string is related to its 'potential,' and this potential is related in turn to the degree to which the segment at a particular position in an utterance is in agreement with the statistical expectation established by the segments surrounding it.

According to this picture, assimilatory and dissimilatory effects, both local and distant, are the effects of a field of 'forces' present in varying degrees at every point in a representation; each segment contributes to the field in its neighborhood, and each segment is in turn influenced by the orientation of the force-field at its location. The image we have in mind is much like that of a magnetic field: each atom contributes to the magnetic field in which it exists, and each atom in turn is affected by the potential which increases as the magnetic orientation of the atom fails to match up with the preferred direction of the field at that point.

We imagine that a phonological representation can be viewed as a sequence of units each of which can assume different 'states,' much like a die can be flipped so as to show different faces pointing upward. Let us call this sequence of positions $\{s_i\}_i$. Each unit s_i is a skeletal position, and the 'state' that it is in defines which segment (selected from the inventory of the language) it represents. To clarify the metaphor a bit, we may say that there are $\mathcal{P}$ different phonemes in the inventory of the language, and that each position s_i is in one of these different states.[22] If one is willing to go beyond the dice metaphor, we can say that each skeletal position s_i is associated with a vector in a space of dimensionality $R^{\mathcal{P}}$ which we call its characteristic vector. The direction of that vector at each position is what we think of as the surface realization of that position. In the terms of conditional random fields, mentioned above (see footnote 12), the underlying representation specifies the 'input' random variables, and the surface form is specified by the 'output' values. In many applications of conditional random fields, the output values are interpreted as (otherwise hidden) labels of observed data; this is not the case here, however.[23]

The direction of the characteristic vector is determined by three things: the underlying specification of the position, the influence of the phonological force field at the position in which the position occurs, and the 'stiffness' associated with a twisting of the direction of the vector from one direction to another—in effect, the cost associated with shifting from one segment-type to another (here, twisting it from its underlying 'direction' to its surface direction). The direction of the characteristic vector directly determines what the surface form of the position is.

As we have tried to show in this paper, the critical factor describing quantitatively the effect of a segment in one position on another position is the mutual information; we assume here that for any two segments, there are two such parameters, *local mu-*

[22] For a segment to shift from one phoneme to another (a *d* might become a *t*, for example), the unit takes on a different state, which metaphorically could be viewed as a die whose *d*-labelled face is up flipping to a state where the face labelled *t* is up.

[23] There is a growing body of literature on the automatic learning of the features used in conditional random fields (see McCallum 2003 and more recent papers). These methods could be used for the automatic learning of the present model.

tual information $MI_L(p;q)$ and *distal mutual information* $MI_D(p;q)$, just as we did in the previous two sections.

We indicate the force field at position i as $\mathcal{F}(i)$; this is a vector in $R^{\mathcal{P}}$. For a given string S, the strength of $\mathcal{F}(i)$ in a given direction p is determined by the unigram plog of segment p, minus the local and distal mutual informations; this is approximated in (18):

$$\begin{aligned}(18)\qquad \mathcal{F}_p(i) = plog(p) - MI_D\big(S[i-2],p\big) - MI_L\big(S[i-1],p\big)\\ - MI_L\big(p,S[i+1]\big) - MI_D\big(p,S[i+2]\big)\end{aligned}$$

A natural definition for the potential of string S at position i in the presence of field $\mathcal{F}$ is given in (19) where $\langle .,. \rangle$ represents the vector inner product.

$$(19)\qquad \mathcal{H}(S,\mathcal{F},i) = \big\langle \overrightarrow{S[i]}, \mathcal{F}(i)\big\rangle,$$

A natural way to deal with alternations in this framework is to include a set of $\mathcal{P}^2$ parameters $\tau_{j,k}$ describing the potential associated with twisting an underlying phoneme-state j to a surface phoneme-state k. For the first time, we need to distinguish the underlying specification of a segment from its surface form, and so we extend our notation slightly: when there is a discrepancy between the two, we take $S/i/$ to represent the ith position of the underlying string, and $S[i]$ to represent the ith position of the surface string. The potential described in (19) would then include a term expressing the phonological probabilistic cost of mismatch between deep and surface form. (This is similar to models that compare strings in bioinformatics; see, for example, Durbin et al. 1999, Chap. 2.) Let us assume that we have established an alignment between underlying and surface forms; this alignment need not be one-to-one, and it will certainly be probabilistic, and we use it to establish a probability distribution over pairs of segments $\mathrm{pr}_d(/x/,[y])$ as in (20).

$$(20)\qquad \sum_{x,y\in\Sigma} \mathrm{pr}_d\big(/x/,[y]\big) = 1.$$

In this model, it would be natural to interpret the potential function as a linear combination of the expression in (19) and sum of the plogs of the deep-surface segment pairings $\{S/i/, S[i]\}_i$. We are pursuing these developments in work in progress as natural extensions of the models described here.

6 Conclusion

This paper has been an exploration of the usefulness of information theoretic tools for understanding phonological phenomena. It began, from our point of view, with the conjecture presented in Sect. 3.6: the effects of vowel harmony in a language like Finnish should result in a decrease in entropy if we condition the probability of a vowel on the vowel that precedes, at whatever distance, rather than by the immediately preceding segment. To our surprise, this hypothesis was soundly defeated by the data. As we continued to explore why the hypothesis was wrong, we came to be more and more impressed by the usefulness of a methodology that allows structure

in the phonological data to speak so forthrightly to the researcher. We were never in a situation where our preference as to styles of phonological analysis (autosegmental, rule-based, constraint-based, innate features, etc.) played a significant role in the testing of any hypothesis; at best, or at worst, such preferences may have limited the range of hypotheses our creativity was limited to.

We are only too aware of additional steps that should be taken even in the analysis of Finnish vowel harmony, not to mention similar problems in every other language in the world. Perhaps the most important next step is to compare the model developed in this paper based on whole words to a model based on words divided into morphemes, so that vowel harmony strictly inside of morphemes can be studied separately and compared to cross-morpheme vowel harmony.

In the final analysis, we believe that information theory is a critical tool for analyzing and understanding linguistic data—and linguistic data is what linguists build their theories on. Adoption of these mathematical tools is in no sense a repudiation of the essential and integral role of formal modeling or abstract structure in linguistic theories. Indeed, the main goal of this paper has been to understand how a particular kind of abstract structure—the vowel tier—can arise as a necessary consequence of the use of probabilistic models in which we seek to maximize coverage while simultaneously minimizing model complexity. As E.T. Jaynes (2003) aptly puts it, we view probability distributions as *carriers of information*.

The most important next step is to better explore the relationship between increasing the size of our model and increasing the probability of the data. We need to better understand and to explicitly model the trade-off between expanding the information contained in a grammatical description of a language, on the one hand, and the improvement in the log probability of the data that follows from that expansion of the grammar. In addition, we plan to look to see whether the economy that information theory provides will help us better understand the trade-off between phonology and morphology. There is already a fair literature on the information theoretic complexity of morphological analysis (e.g. Goldsmith 2001), and it should be possible to explicitly compare the relative complexity of analyzing various phenomena in morphological terms with that of analyzing them phonologically. If we can achieve that goal, we will have arrived at a significantly deeper explanation of the relationship of two significant components of linguistic grammar, and a new understanding of how the phonological analysis of a natural language is justified.

Acknowledgements For helpful and insightful discussion, suggestions, and comments we would like to thank: Max Bane, Ryan Bennett, Pierre Collet, Antonio Galves, Sharon Goldwater, Yu Hu, Junko Itô, Mark Johnson, Theano Starvinos, John Sylak, Colin Wilson, and Alan Yu.

References

Albro, Daniel. 2000. A probabilistic ranking learner for phonotactics. Ms., UCLA.

Altmann, Gabriel, and Werner Lehfeldt. 1980. *Einführung in die quantitative phonologie*. Vol. 7 of *Quantitative linguistics*. Bochum: Studienverlag Dr. N. Brockmeyer.

Bailey, Todd M., and Ulrike Hahn. 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? *Journal of Memory and Language* 44: 568–591.

Baker, Adam C. 2009. Two Bayesian approaches to finding vowel harmony. Technical report, University of Chicago. http://home.uchicago.edu/~adamc/papers/vharmony.pdf.

Belevitch, Vitold. 1956. *Langage des machines et langage humain*. Bruxelles: Office de Publicité.
Billerey-Mosier, Roger. 2003. Exemplar-based phonotactic learning. Paper given at SWOT 8, Tuscon, AZ.
Bod, Rens, Jennifer Hay, and Stefanie Jannedy. 2003. *Probabilistic linguistics*. Cambridge: MIT Press.
Bradshaw, Mary. 1999. A crosslinguistic study of consonant-tone interaction. PhD diss, Ohio State University.
Cherry, Colin, Morris Halle, and Roman Jakobson. 1953. Toward the logical description of languages in their phonemic aspect. *Language* 29: 34–46.
Chomsky, Noam. 1956/1975. *The logical structure of linguistic theory*. New York: Plenum.
Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton.
Cole, Jennifer. 1987. Planar phonology and morphology. PhD diss., MIT.
Cole, Jennifer. 2009. Emergent feature structures: Harmony systems in exemplar models of phonology. *Language Sciences* 31 (2–3): 144–160. doi:10.1016/j.langsci.2008.12.004. Data and Theory: Papers in Phonology in Celebration of Charles W. Kisseberth.
Coleman, John, and Janet B. Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. In *Computational phonology. Third meeting of the acl special interest group in computational phonology*, 49–56. Somerset: Association for Computational Linguistics.
Courant, Richard, and Herbert Robbins. 1941. *What is mathematics?* New York: Oxford University Press.
Cover, Thomas M., and Joy A. Thomas. 1991. *Elements of information theory*. New York: Wiley.
Dainora, Audra. 2001. An empirically based probabilistic model of intonation in American English. PhD diss., University of Chicago.
Downing, Laura. 2008. Where does depression come from in Nguni Bantu languages? Paper presented at the Old World Conference in Phonology 5, Toulouse, France.
Durbin, Richard, Sean R. Eddy, Anders Krogh, and Graeme Mitchison. 1999. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge: Cambridge University Press.
Eisner, Jason. 2002. Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th annual meeting of the association for computational linguistics (acl)*, Philadelphia, 1–8.
Frisch, Stefan, N. R. Large, and David B. Pisoni. 2000. Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language* 42: 481–496.
Geman, Stuart, and Mark Johnson. 2001. Probability and statistics in computational linguistics, a brief review. Ms: http://www.cog.brown.edu/~mj/papers/Review.pdf.
Goldsmith, John. 1985. Vowel harmony in khalkha mongolian, yaka, finnish and hungarian. *Phonology* 2: 253–275.
Goldsmith, John A. 1990. *Autosegmental and metrical phonology*. Oxford: Blackwell.
Goldsmith, John A. 1991. Phonology as an intelligent system. In *Bridges between psychology and linguistics: A swarthmore festschift for Lila Gleitman*, eds. Donna Jo Napoli and Judy Kegl. Mahwah: Lawrence Erlbaum.
Goldsmith, John A. 1993. Harmonic phonology. In *The last phonological rule*, ed. John Goldsmith, 221–269. Chicago: University of Chicago Press.
Goldsmith, John A. 2001. The unsupervised learning of natural language morphology. *Computational Linguistics* 27 (2): 153–198.
Goldsmith, John A. 2002. Probabilistic models of grammar: Phonology as information minimization. *Phonological Studies* 5: 21–46.
Goldsmith, John A. 2007a. Analogy in morphology: Only a beginning. In *Proceedings from a conference on analogy*, eds. James Blevins and Juliette Blevins.
Goldsmith, John A. 2007b. Towards a new empiricism. In *Recherches linguistiques à vincennes*, ed. Joaquim Brandao de Carvalho, 9–36.
Goldsmith, John A., and Aris Xanthos. 2006. Discovering phonological categories.
Goldwater, Sharon, and Mark Johnson. 2003. Learning OT constraint rankings using a maximum entropy model. In *Proceedings of the workshop on variation within optimality theory*, eds. Jennifer Spenader, Anders Eriksson, and Östen Dahl, 111–120. Stockholm: Stockholm University.
Goldwater, Sharon, and Mark Johnson. 2004. Priors in Bayesian learning of phonological rules. In *Proceedings of the seventh workshop of the ACL special interest group in computational phonology*, 35–42.
Good, Irving John. 1980. Some history of the hierarchical Bayesian methodology. *Trabajos de Estadística y de Investigación Operativa* 31: 489–519.
Halle, Morris. 1958. Review of Herdan, language as choice and chance. *Kratylos* 3: 20–28.
Hansson, Gunnar. 2001. Theoretical and typological issues in consonant harmony. PhD diss., University of California, Berkeley.

Hayes, Bruce, and Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39 (3): 379–440.
Herdan, Gustav. 1956. *Language as choice and chance*. Groningen: P. Noordhoff.
Herdan, Gustav. 1960. *Type-token mathematics*. 's-Gravenhage: Mouton.
Herdan, Gustav. 1964. *Quantitative linguistics*. London: Butterworths.
Hopcroft, John E., and Jeffrey D. Ullman. 1979. *Introduction to automata theory, languages, and computation*. Reading: Addison-Wesley.
Huang, Xuedon, and Alex Acero. 2001. *Spoken language processing: A guide to theory, algorithm and system development*. New York: Prentice Hall.
Jaynes, Edwin T. 2003. *Probability theory: The logic of science*. Cambridge: Cambridge University Press.
Jäger, Gerhard. 2004. Maximum entropy models and stochastic optimality theory. Rutgers Optimality Archive: ROA 625.
Kaufman, Leonard, and Peter J. Rousseeuw. 2005. *Finding groups in data: An introduction to cluster analysis*. New York: Wiley-Interscience.
Kiparsky, Paul. 1973. Phonological representations. In *Three dimensions of linguistic theory*, ed. Osamu Fujimura, 1–136. Tokyo: TEC.
Kisseberth, Charles. 1984. Digo tonology. In *Autosegmental studies in bantu tone*, eds. G.N. Clements and John Goldsmith, 105–182. Dordrecht: Foris.
Kucera, Henry. 1982. Markedness and frequency: A computational analysis. In *Coling 1982*, ed. J. Hoercky, 167–173. Amsterdam: North Holland.
Lafferty, John D., Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data, 282–289.
Laughren, Mary. 1984. Tone in zulu nouns. In *Autosegmental studies in bantu tone*, eds. G.N. Clements and John Goldsmith, 183–234. Dordrecht: Foris.
Lewin, Jonathan. 2003. *An interactive introduction to mathematical analysis*. Cambridge: Cambridge University Press.
Li, Ming, and Paul M. B. Vitányi. 1997. *An introduction to Kolmogorov complexity and its applications*, 2nd ed. New York: Springer.
Lin, Ying. 2005. Learning features and segments from waveforms: A statistical model of early phonological acquisition. PhD diss., University of California Los Angeles.
Manning, Christopher D., and Hinrich Schütze. 2000. *Foundations of natural language processing*. Cambridge: MIT Press.
McCallum, Andrew. 2003. Efficiently inducing features of conditional random fields. In *Nineteenth conference on uncertainty in artificial intelligence (uai03)*, 403–410.
Padgett, Jaye. 2011. Consonant-vowel place feature interactions. In *The Blackwell companion to phonology*, eds. Elizabeth Hume, Marc van Oostendorp, Colin J. Ewen, and Keren Rice. Oxford: Blackwell.
Pater Joe, Christopher Potts, and Rajesh Bahtt. 2007. Harmonic grammar with linear programming. ROA: 984-0708.
Pereira, Fernando. 2000. Formal grammar and information theory: Together again? *Philosophical Transactions of the Royal Society* 358: 1239–1253.
Pierrehumbert, Janet B. 1994. Syllable structure and word structure: A study of triconsonantal clusters in English. In *Phonological structure and phonetic form: Papers in laboratory phonology III*, ed. Patricia A. Keating, 168–188. Cambridge: Cambridge University Press.
Pierrehumbert, Janet B. 2001. Stochastic phonology. *GLOT* 5 (6): 195–207.
Pierrehumbert, Janet B. 2003. Probabilistic phonology: Discrimination and robustness. In *Probability theory in linguistics*, eds. Rens Bod, Jennifer Hay, and Stefanie Jannedy. Cambridge: MIT Press.
Prince, Alan, and Paul Smolensky. 1993/2004. *Optimality theory: Constraint interaction in generative grammar*. Cambridge: MIT Press.
Ringen, Catherine O. 1975/1988. *Vowel harmony: Theoretical implications*. Garland Press, NY. Indiana University PhD dissertation, 1975. Published in 1988 by Garland Press, NY.
Ringen, Catherine O., and Orvokki Heinämäki. 1997. Variation in finnish vowel harmony: An OT account. *Natural Language & Linguistic Theory* 17: 303–337.
Rissanen, Jorma. 1989. *Stochastic complexity in statistical inquiry*. Vol. 15 of *Series in computer science*. Singapore: World Scientific.
Rose, Sharon, and Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80: 457–531.
Rose, Sharon, and Rachel Walker. 2011. Harmony systems. In *The handbook of phonological theory*, 2nd ed., eds. John Goldsmith, Jason Riggle, and Alan Yu. New York: Wiley-Blackwell.

Seidenberg, Mark. 1997. Language acquisition and use: Learning and applying probabilistic constraints. *Science* 275: 1599–1604.

Shannon, Claude. 1951. Prediction and entropy of printed English. *Bell Systems Technical Journal* 30: 50–64.

Shannon, Claude, and Warren Weaver. 1949. *The mathematical theory of communication*. Urbana: University of Illinois Press.

Solomonoff, Ray. 1959a. A new method for discovering the grammars of phrase structure languages. In *Proceedings of the international conference on information processing*, 256–290. Paris: UNESCO.

Solomonoff, Ray. 1959b. A progress report on a machine to learn to translate languages and retrieve information. In *Advances in documentation and library sciences*, 3, 941–953. New York: Interscience.

Solomonoff, Ray. 1964. A formal theory of inductive inference. *Information and Control* 7: 224–254.

Solomonoff, Ray. 1997. The discovery of algorithmic probability. *Journal of Computer and System Sciences* 55 (1): 73–88.

Trubetzkoy, Nicolas Sergueevitch. 1939/1968. *Grundzüge der phonologie* (translated in French by Jean Cantineau: Principes de phonologie). Paris: Klincksieck.

Walker, Rachel. 2000. Long-distance consonantal identity effects. In *Proceedings of the west coast conference on formal linguistics 19*, 532–545.

Walker, Rachel. 2005. Weak triggers in vowel harmony. *Natural Language & Linguistic Theory* 23 (4): 917–989. doi:10.1007/s11049-004-4562-z.

Wilson, Colin. 2006. Learning phonology with substantive bias: An experimental and computational study of velar palatalization. *Cognitive Science* 30: 945–982.