Chapter 21

A Bayesian evaluation of the cost of abstractness

Ewan Dunbar, Brian Dillon and William J. Idsardi

University of Maryland

'There were giants in the earth in those days...'
Genesis 6:4

21.1 Introduction

At the beginning of Bever's career it was possible to do both phonology and syntax: just three years separate *Aspects* (Chomsky, 1965) and SPE (Chomsky and Halle, 1968); four years separate Katz and Postal (1964) and Postal (1968); three years separate Bever (1967) from Bever (1970). The other contributions to this volume deal mostly with *The horse raced past the barn fell* and its fallout; we will instead try to update some of the phonological points made in Bever (1967). In arguments later cited in SPE and many other places, Bever (1967) argues for abstract, opaque phonological analyses of Menomini, specifically citing simplicity of grammatical description as a major driving force for the analysis (Bever, 1967: 18-20). These ideas have once again gained prominence with the rise of Bayesian approaches to problems of parsimony (Jaynes, 2003, Dowe, Gardner, and Oppy, 2007; on parsimony in general, see Sober, 1975, 1988, 1990, 1994), and the Bayesian approaches yield a new (probabilistic) understanding of evaluation measures and their relation to computational learning theories (Solomonoff, 1964ab, Rissanen, 1978). In addition to providing new formal tools for evaluating representational complexity, Bayesian approaches also highlight another idea championed by Bever, analysis-by-

synthesis (Bever and Poeppel, 2010), by mathematically relating posterior probability (analysis) and likelihood (synthesis). In this chapter we capitalize on the formal tools from Bayesian approaches to inference to offer a new understanding of an argument for opaque phonological analyses, which we will illustrate with a problem from Kalaallisut.

In the study of human language, as in any science, the data is noisy, is typically many layers removed from the real object of study, and for any number of other reasons, tends to underdetermine the theory. Thus, as in any science, we must constantly engage in a process analogous to what statisticians call 'model comparison', examining two or more competing theories and evaluating them to see which provides a better explanation of the data. In the study of human cognition, however, model comparison has a second significance, entirely separate from the ordinary workings of science. A productive language system develops over time in a child in response to linguistic input; the diversity of human languages and the uniformity of speakers' generalizations within a linguistic community show that the language system internalized by the learner (the internal *model*) depends on the input (the *data*). Implicitly or explicitly, then, the language learner is making comparisons between possible models of the ambient language, while the language scientist makes comparisons between possible models of the language learner.

The search for formal principles of discovery has always been of great interest within linguistics, from Harris's (1951) algorithmic recommendations for the analyst, through Chomsky and Halle's (1968) evaluation measure, to modern simulated parameter learners like those of Dresher and Kaye (1990) and Yang (2002). Yet the analyst attempting to deduce the correct mental analysis of some language still relies largely on subjective criteria; it is safe to say that, although model comparison is an integral part of linguistics, our understanding of the human language learner's principles of model comparison have

yet to reach the stage where they are useful to linguists. At the same time, however, the science of complex inference is a well developed one, with much to offer the cognitive scientist. One of the most popular modern approaches is the *Bayesian* approach, which leverages a particular kind of probabilistic reasoning. The main insight behind probabilistic approaches to model comparison is that the problems involve uncertainty, for the analyst and for the learner alike; probability theory is the simplest and most widely accepted formal theory of reasoning under uncertainty.

To demonstrate the utility of this reasoning for language scientists, we take a standard problem of abstractness in phonological grammar as an example problem. Since the publication of the Sound Pattern of English (SPE; Chomsky and Halle, 1968), phonologists have been deeply concerned with the question of what constitutes an appropriate use of abstractness in a phonological analysis (Kiparsky, 1968, 1971; Hooper, 1976). More recently, many researchers formulating grammars in Optimality Theory (OT; Prince and Smolensky, 1993) have avoided analyses which crucially rely on opaque process interactions (Sanders, 2003). This is because in its original formulation, OT captures only surface-true interactions among processes, although a number of contemporary versions of OT are specifically aimed at allowing derivational analyses to be stated (McCarthy 1999, 2007, 2010).

Here we focus on a typical case of abstractness in phonology, a simple apparent case of opacity in Kalaallisut, an Inuit language of Greenland, and argue from Bayesian reasoning that opaque or non-surface-true representations of the Kalaallisut vowel system are preferred. Although a full analysis is beyond the scope of this current paper, our goal here is to highlight the way this reasoning works. In particular, we highlight the fact that a Bayesian learner will, all other things being equal, favour simpler models; that is, if we

assume the axioms of decision making under uncertainty that underlie this approach, we immediately impute an Occam's Razor like simplicity bias to the learner. We show how a particular set of assumptions about the mechanisms of phonological grammar would compel an ideal learner to arrive at an abstract solution simply by force of these well-motivated domain-general reasoning strategies. We discuss the implications for the study of language acquisition.

21.2 Kalaallisut phonology

Kalaallisut is an Inuit language spoken in Greenland; it has been the sole official language of Greenland since 2009. The inventory of Kalaallisut, closely following Rischel (1975), is given in Table 21.1 (omitting length distinctions, which are contrastive for both vowels and consonants, but irrelevant here).[1]

The vowel inventory shown in Table 21.1 contains three phonemes, /i/, /u/, and /a/. As in many languages with uvular consonants, including the other Inuit languages, vowels are affected by following uvulars, being subject to the process described by the rule in (1) (Rischel, 1975; Dorais, 1986).

(1) $[+\text{syll}] \rightarrow [+\text{RTR}] \; / \; \underline{\quad} \begin{bmatrix} -\text{syll} \\ +\text{cons} \\ +\text{RTR} \end{bmatrix}$

The existence of the process in (1) means that the vowels of Kalaallisut each have a retracted allophone. We will notate these segments as [ɐ], [ɔ], and [ɑ] for the sake of presentational convenience, and not to make any precise claims about the phonetic values of these variants. Examples are given in (2)–(3) (examples from Bittner, undated).[2]

(2) ani + pallag + pu + q → [anipałłapp**oq**], 'went quickly'

(3) salu + qi + llu + ni + lu → [sal**oq**alunilu], ‘and he is very thin’

In addition to vowel shifts before uvular consonants, processes of regressive consonant assimilation are also common across the Inuit languages, and are most total and apply most liberally in the easternmost dialects (Dorais, 1986), including Kalaallisut. Importantly, in addition to total regressive assimilation targeting all other classes of consonants, Kalaallisut has regressive assimilation targeting uvulars, as seen in (4)–(6).

(4) aŋala + taʁ + pu + q → [aŋalasɑ**pp**oq], ‘he always travelled’

(5) uqaʁ + pu + q → [oqɑ**pp**oq], ‘he said’

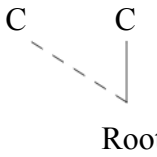
(6) sinig + niaʁ + tu + t → [sini**nn**iɑ**tt**ut], ‘he said’

As can be seen in this second set of examples, these two rules can both apply (indeed, the syllable structure of the language makes it impossible to construct an example of assimilation of a uvular in which the retraction rule would not apply), and the resulting interaction is opaque (a case of *counterbleeding* in the sense of Kiparsky, 1971).

Kalaallisut opacity is a somewhat nuanced, however, and helps to illustrate some of the controversy surrounding this kind of abstractness. The nature of the assimilation of uvulars in Kalaallisut has been a matter of some discussion, for two reasons. First, because, unlike all other consonant assimilations, the underlying uvular consonant rather than the surface assimilated consonant is marked in Kalaallisut orthography, so that [sininniɑttut] is written as *sininniartut,* with the assimilation marked for the [g] but not the [ʁ]. Second, it is often a detectably incomplete neutralization, even to non-native speakers. Phonetic analysis by Mase and Rischel (1971) revealed no evidence of frication in assimilated /ʁ/, but our own informal listening suggests that some trace of uvularity remains audible in a substantial number of cases.

Rischel (1974) proposes several alternate analyses of this fact. In one, the surface uvularity is cued entirely by the vowel quality. The assimilation in Kalaallisut is across-the-board total assimilation, as in (7).

(7)



This analysis claims that the interaction between the two processes is an opaque one, as shown in (8).

(8)

	/uqaʁpuq/
(1)	[oqɑʁpoq]
(7)	[oqɑppoq]
	[oqɑppoq]

Though the opaque analysis is one theoretical possibility, there is another grammatical analysis that has been preferred. Under Rischel's preferred analysis, assimilation spreads all features but [RTR] (Rischel's [± retracted]). Under this analysis, the underlying uvular consonant retains its [+RTR] feature after assimilation, and there is no opacity.

The phonetic question of whether tongue retraction is detectable on the surface 'in the consonant' or not is a crucial one, and it is characteristic of the debate that takes place in these cases. In this case, it is quite a difficult one, given the results of Alwan (1999), which suggest that, in the absence of a burst, the main cues to uvular place information are to be found in the first formant of an adjacent vowel. Nevertheless, assuming that languages can make a contrastive difference between uvular consonants and non-uvular consonants which happen to be preceded by [+RTR] vowels, the question is empirical and as yet unresolved (we recommend further MRI studies of tongue root position). It is fair to

say that much rests on empirical outcomes like this, as true cases of counterbleeding are a problem for monostratal theories of phonology (Prince and Smolensky, 1993), and substantial effort has been devoted to denying their existence, sometimes by appealing to subtle phonetic arguments.

What follows is a theoretical argument. If we assume that uvularity is obscured 'in the consonant' in at least some tokens, then it is reasonable to call the current case in some sense opaque. In the current paper we are primarily interested here in examining one sort of argument that has been made against these kinds of opaque interactions. In the absence of phonetic facts that might undermine the case for true surface opacity, researchers have given 'transparent' analyses of opacity in which the 'opaque' segments have been reanalyzed as independent phonemes.

For the well-known case of Canadian Raising, for example, where [aw] and [aj] alternate with [ʌw] and [ʌj] before voiceless stops even when they are neutralized by a subsequent flapping rule, Mielke, Anderson and Hume (2003) propose that, rather than a single pair of phonemes, /aw/ and /aj/ , subject to a raising process, there are four phonemes: /aw/ , /aj/ , /ʌw/, and /ʌj/. Storing the surface form in this manner is possible only in cases where the alternation does not occur across a morpheme boundary. For cases in which the raising does apply across a morpheme boundary, the grammar must preserve both processes. The facts are contested in the case of Canadian English (see Idsardi, 2006). Importantly for present purposes, the above examples demonstrate that both retraction and assimilation processes must apply across morpheme boundaries in Kalaallisut.

Our focus here is on this second kind of argument. Let us therefore assume that (7) is basically correct, and that the assimilation is truly total for uvulars, at least in some cases.

If the set of Kalaallisut vowels is as given in Table 21.1, then we have an *opaque analysis*; but there is clearly an alternate analysis—a *transparent analysis*—in which both rules still exist (though now perhaps as rules of allomorph selection), but the Kalaallisut vowels are as in Table 21.2.

Under such an analysis, the underlying form for a word like [oqɑppoq] would be /oqɑʁ + pu + q/, with stored retracted vowels in the first morpheme (guaranteed to be stored under the Lexicon Optimization hypothesis of Prince and Smolensky, 1993). By the process corresponding to (1), we get a retracted vowel in the second morpheme; we get assimilation of the final consonant of the first morpheme by the process corresponding to (7). Sometimes we have morphological evidence sufficient to rule out the transparent analysis. However in Kalaallisut, we do not, despite its highly agglutinative nature, as we would need a sequence /V + Q + C/, where Q is either [q] or [ʁ]. The only such morpheme we are aware of is the third person singular morpheme -/q/, but this always appears word-finally, and consequently cannot display assimilation.

This is a typical case of abstractness, an apparent case of opacity in phonology. There are several possible analyses; here we focus on two. The transparent analysis has more phonemes (possible lexical segments); the opaque anlaysis relies on interesting non-trivial properties of complex phonological systems. There is an intuition that one is somehow ‘closer’ to what is observed than the other, but the question of which analysis a human learner would select, particularly given that the crucial data appears to be obscured, is an empirical one. This is exactly where we would like some other facts about the human inference system (the language acquisition device) to come to bear. This is a case where a theory of inference under uncertainty would be informative, because there are multiple

reasonable solutions. In this case, we argue that Bayesian inference can be brought to bear directly on the question of abstractness.

21.3 Bayesian reasoning in linguistics

In the previous section we have demonstrated a typical case of model selection in linguistics. The decision between transparent and opaque models of the Kalaalisut vowel system hinges crucially on a fundamental and divisive issue in the field, that of abstractness. These two models pit storage against computation. In this, we would benefit from having an independently motivated theoretical stance on the learner.

The approach we take is to study an ideal learner. The problem of language acquisition is the problem of searching for a grammar that is in some sense an optimal model with respect to the primary linguistic data. One theoretical approach to language acquisition is to focus on the consequences of various search procedures; this is the general character of the proposals made by Dresher and Kaye (1990), Clark (1992), Niyogi and Berwick (1996), and Yang (2002), among others. Each of these proposals describes a different algorithm for exploring the set of possible grammars (in the case of Yang, 2002, a probabilistic search). On the other hand, the phonological category acquisition work of deBoer and Kuhl (2001), Vallabha *et al*., (2007), has taken a different approach. This literature applies standard statistical techniques to a learning problem—in this case, the problem of determining the location and extent of vowel categories in a language in acoustic space—in order to approximate some theoretically optimal solution. By proceeding in this way, these researchers have drawn the conclusion that the search problems under consideration are in principle solvable in a relatively straightforward search space (in the vowel-learning case, the space of possible formant values, plus several other

acoustic parameters); similarly, by adapting these same models to deal with more complicated vowel systems and more realistic data sets, Dillon, Dunbar, and Idsardi (to appear), have drawn inferences about what restrictions need to be put on the hypothesis space a priori for phonological category learning.

Here we present a study of the second kind. Rather than specifying the mechanism by which the learner reaches the adult state, we will describe the learning problem at an abstract level and attempt to find a theoretically optimal solution, in the hope that this will shed light on the question of what is learnable in phonological grammar (see Hayes and Wilson, 2008 for a proposal along similar lines emphasizing the use of a maximum entropy principle). In what follows, we state how this kind of reasoning works. We show how an Occam's Razor effect is observed as a result. We then state certain theoretical assumptions which will allow us to highlight this approach in grammatical inference; finally, we spell out some details in the current case.

21.3.1 Probability

In this section we provide a brief overview of the elements of probabilistic reasoning, using examples from phonology.

Let us begin with the phonetics-phonology mapping. Following standard assumptions, we assume that the mapping from phonetic to phonological representations is a mapping from continuous values (the outputs of lower-level audition) to discrete values (the alphabet of the phonological system). On this assumption, the learner's task is to determine exactly how this mapping is structured. This follows from the fact that identical phonetic values are mapped to different phonological categories across languages, a fact which can be seen both in the operation of phonological processes and in speech perception

(Stevens *et al*., 1969; Werker and Tees, 1984; Kazanina *et al*., 2006; Herd, 2000; Dresher, 2009).

In probabilistic modeling, the general term for a model in which each observed data point is a member of one of a finite number of categories is a *mixture model*. The intuition behind a mixture model is that, in order to generate a data point, some procedure selects a category, and, a category having been selected, some other procedure generates an instance of the selected category. In the current case, using a mixture model to describe vowels simply asserts that there is a many-to-one mapping from possible phonetic tokens to vowel categories. From a probabilistic modeling perspective, the statement of a mixture model is as in (9):

(9) $$\Pr(x) = \sum_{i=1}^{C} \Pr(c_i) \Pr(x \mid c_i)$$

Equation (9) is read as follows: the probability of some observed phonetic value x is equal to the following value, summed over all C vowel phonemes: the probability of the phoneme c_i times the within-phoneme (conditional) probability of the observed token, once we assume that x is an instance of c_i. This statement follows from the basic axioms of probability, which require that the probability of any of a finite number of mutually exclusive events (such as the occurrences of a phonetic value x conjoined with each member of the set of phonemic categories) be equal to the sum of the probability of each event (the *law of total probability*), and that conditional probabilities be related to joint probabilities (probabilities of conjunctions) by (10).

(10) $$\Pr(x \text{ and } c_i) = \Pr(x \mid c_i) \Pr(c_i)$$

Importantly, it is not the case that a mixture model treatment of this process commits the theorist to a probabilistic view of grammar, as deterministic models may be taken to be special cases of the stochastic model we formulated above. To see more clearly how a probabilistic formulation can give a deterministic model, consider the problem of recognizing speech. Given some segment with phonetic values x, the problem is to determine the phonological category c_x which generated x; that is, we must find the value of c which maximizes $\Pr(c \mid x)$. The crucial relation here is *Bayes' Rule*, given in (11):

$$(11) \qquad \Pr(c \mid x) = \frac{\Pr(x \mid c)\Pr(c)}{\Pr(x)}$$

Furthermore, we can expand the denominator using the law of total probability:

$$(12) \qquad \Pr(c \mid x) = \frac{\Pr(x \mid c)\Pr(c)}{\sum_{i=1}^{C} \Pr(x \mid c_i)\Pr(c_i)}$$

Now suppose that there is no overlap between phoneme categories, that is, that there is no acoustic value x such that the phonetics-phonology mapping would simultaneously assign $\Pr(x \mid c_1) > 0$ and $\Pr(x \mid c_2) > 0$ for $c_1 \neq c_2$; put another way, suppose there are no regions of *uncertainty*. Then, if we are given some x, there is only ever one category c_i with a non-zero value in the expansion of denominator in (12); furthermore, the probability of the correct category c_i (correct according to the model) given some data point x, is always 1:

$$(13) \qquad \Pr(c_i \mid x) = \frac{\Pr(x \mid c_i)\Pr(c_i)}{0 + \cdots + \Pr(x \mid c_i)\Pr(c_i) + \cdots + 0}$$

$= 1$

Because a mixture model is a stochastic model, it is capable of imputing detailed 'degrees of certainty' (probability) about various inputs (a probability distribution); nevertheless, probability distributions have as special cases both maximal certainty (determinism) and maximal uncertainty (uniform distributions). Because of this link, probability theory can be used as a way of formalizing reasoning in cases of high and low uncertainty alike. In the case of absolute certainty, it can be shown that it reduces to Aristotelian logic; when there is uncertainty, it can be shown to be reducible to a very small number of axioms of consistent reasoning (Cox, 1946; Jaynes, 2003). While there are other deductive systems for reasoning under uncertainty (for example, fuzzy logic, and the consequent 'possibility theory'; see Zadeh, 1978), probability theory is by far the most widely accepted.

21.3.2 Bayes Rule and model comparison

Because the calculus of probability theory gives us formal tools to evaluate inference in a flexible manner, we can cast the problem of phonological acquisition as inference about the ideal mapping between phonetic values and their associated category labels. On this formalization, the optimal solution to this problem is the model M which has maximal probability given the observed data D:

(14) $$M = \underset{m}{\operatorname{argmax}} \Pr(m \mid D)$$

Many of the theoretical approaches to learning in the literature attempt to specify the method for searching for this optimal model. For example, in Yang (2002), the learner

uses a simple reinforcement learning algorithm to incrementally update $\Pr(m \mid D)$. Our approach here is different. In what follows, we simply try to estimate what the values of this criterion would be under various possible models. We thus use Bayes' Rule, given above, to get the criterion in a more convenient form, as in (15).

$$(15) \qquad \Pr(M \mid D) = \frac{\Pr(D \mid M)\Pr(M)}{\Pr(D)}$$

This statement should be read as follows: the probability of the model after having seen some data (the *posterior;* $\Pr(M|D)$) is proportional to the probability of the data under that model (the *likelihood;* $\Pr(D|M)$), times the a priori probability of that model (the *prior;* $\Pr(M)$). When scaled down by the overall, or *marginal* probability of the data, the relation becomes one of equivalence. The *Bayesian* approach to model comparison makes use of this expansion to do inference. In particular, it accepts that having a probability distribution over possible models is reasonable; this is to be contrasted with the *frequentist* approach to statistical inference, which dominated the statistical toolbox used by scientists throughout most of the twentieth century (there has been a surge in interest in Bayesian methods in recent years: Kass and Raftery, 1995; Jaynes, 2003; Mackay, 2003; Gelman, et al., 2003; Gallistel, 2009). The frequentist approach rejects the use of $\Pr(M)$, because it interprets probability theory not as a theory of reasoning under uncertainty, but as theory of the counts of particular classes of events as the number of observations goes to infinity; in such a theory, talk of the probability of a model is incoherent, because models are not observable events. There are a number of important theoretical reasons for adopting the Bayesian approach, however, including a number of well-known paradoxes under the frequentist interpretation; more importantly, just as probability theory follows as a

straightforward generalization of Aristotelian logic, Bayesian inference is supported by a handful of very general decision-theoretic principles (see Ghosh *et al*., 2006; Robert, 2007).

Bayesian reasoning gives us the decision rule in (16), the *Bayes decision rule*.

$$\text{(16)} \qquad \frac{\Pr(D \mid M_1)\Pr(M_1)}{\Pr(D \mid M_2)\Pr(M_2)} > 1 : M_1, \quad \text{otherwise} : M_2$$

The left-hand side in (16) is the ratio of $\Pr(M_1 \mid D)$ and $\Pr(M_2 \mid D)$. The rule is read as follows: if the left-hand side (the *Bayes factor*) is greater than one, decide in favour of model M_1; if the Bayes factor is less than one, decide in favour of model M_2; the larger the Bayes factor, the better the evidence for M_1. This can be interpreted as an 'odds,' in the gambler's sense. (Comparisons are usually done in log, so that, for example, a difference of two orders of magnitude is considered strong evidence; see Goodman, 1998).

The important thing to note here is that the likelihood and the prior are in a trading relation. We can maximize $\Pr(D \mid M)$ by maximizing the likelihood if the prior is uninformative, or by maximizing the prior if the likelihood does not help in the model comparison. An immediate consequence of this is, all other things being equal, we should pick the a priori more probable model.

As has often been pointed out (Mackay, 2003; Jaynes, 2003), a Bayes factor analysis gives an automatic model complexity penalty, because models with more free parameters yield smaller probabilities. To see this intuitively, consider the simple case in which two models are under comparison, one of which has a single binary valued parameter, and the other of which has two binary-valued parameters. Suppose that under either model, there is a single parameter value ($\hat{\theta}_1$, $\hat{\theta}_2$ respectively) that gives a reasonably

good fit—that is, gives a reasonable likelihood—and the others (or the single other) give near-zero likelihood. We expand out $\Pr(D \mid M)$ (a *marginal likelihood*, because it averages over all parameter values under model M) to get the crucial decision ratio in (17).

$$(17) \qquad \frac{\Pr(D \mid \hat{\theta}_1, M_1)\Pr(\hat{\theta}_1 \mid M_1)\Pr(M_1)}{\Pr(D \mid \hat{\theta}_2, M_2)\Pr(\hat{\theta}_2 \mid M_2)\Pr(M_2)}$$

Suppose both parameter values are equally likely under Model 1, and all four parameter values are equally likely under Model 2. If the two models are equally likely, and they assign equal probability to the data under the single good parameter value for each, we get the decision rule in (18).

$$(18) \qquad \frac{\Pr(\hat{\theta}_1 \mid M_1)}{\Pr(\hat{\theta}_2 \mid M_2)} > 1 : M_1,$$
$$\text{otherwise} : M_2$$

Since there are *four* possible parameter values under M_2, and under M_1 only *two*, if they are all equally likely a priori, the Bayes factor is one quarter divided by one eighth—Model 1 is twice as probable.

Importantly, this means the following: Bayesian reasoning not only tells us that, all other things being equal, we should pick the most probable model (or, of course, conversely, the priors being equal, we should pick the model that assigns the higher probability to the data); it also tells us that we should in general pick the model *with fewer free parameters*. In essence, we derive Occam's Razor.

A word of warning is in order. Fully Bayesian inference will compare models by averaging over all possible parameter values (thus, by using the marginal likelihood). In our example, we assumed that there was only one parameter value worth looking at, because

the rest assigned negligible probability to the observed data; thus averaging would be pointless, because we would multiply in likelihood values close to zero for the other parameter values. We will continue to use this oversimplified reasoning to illustrate how the Bayesian approach can bring this important complexity penalty to linguistics. In reality, as we increase the number of free parameters, a number of things change about the performance of the model. First, we can eventually find parameter values that give greater likelihood to the observed data (imagine a model with as many parameters as data points); second, we can find *more* high-likelihood models (there are more ways to get the same data). Thus, averaging, we might find that all things are not equal, not only because the best parameter value may be better under the more complex model, but also because there might be more 'best' parameter values to choose from. There will be some tradeoff against model complexity, as we have shown, of course; the question is simply how quickly the likelihoods and the number of good fits grow, as compared to how quickly the conditional priors on the parameters shrink. This can only be determined given the particular model and data set we are working with.

Abstracting away from this, however, the logic is clear: all other things being equal, Bayesian reasoning tells us to prefer simpler models. This is the essence of the reasoning we use in this paper: simpler models are preferred. In the current case, models with fewer phonemes are preferred. What follows is simply filling in the details.

21.3.3 Theoretical assumptions

In order to illustrate our point, we will need to make some assumptions about the shape of the phonological model.

Recall from the preceding discussion that to assume discrete phonemes is to assume

a mixture model, in which there is a choice between some finite number of categories, and each category has some distribution.

$$\Pr(x) = \sum_{i=1}^{c} \Pr(x \mid c_i) \Pr(c_i) \quad (19)$$

This probability has two parts for each component: a *class-conditional* probability $\Pr(x \mid c_i)$, and a *mixing probability* $\Pr(c_i)$. For example, following deBoer and Kuhl, 2001, Vallabha et al., 2007, and Feldman *et al*., 2009, we might assume that $\Pr(x \mid c_i)$ (yielding the probability distribution for acoustic tokens under each phoneme, or *component of the mixture*) follows a multivariate Gaussian distribution; we might consider assuming other distributions, including uniform distributions, though the speech perception literature seems to us to suggest that a uniform distribution is an inappropriate model for vowels, since identification rates vary in proportion to distance from the category centre (see for example, Pisoni, 1975; Kuhl, 1991; Savel,a 2009). For current purposes, $\Pr(c_i)$, the mixing probability, is immaterial; it is most often modeled as a multinomial distribution (Vallabha et al., 2007), but Feldman et al. (2009) construct a more complicated model which, seen as a mixture, essentially uses a draw of a word from a simulated lexicon to get these probabilities.

We will further assume a model of the phonetics-phonology mapping in which the computation of allophony is a subsymbolic process, in particular, the model argued for by Dillon, et al. (to appear). In this model, phonetic categories are fit simultaneously with a set of subsymbolic shifts in phonetic space corresponding to allophonic rules. In this model, there crucially are no phonetic categories (‘phones’), in the sense of phonemes with all postlexical processes applied to them. This model can be seen as taking extremely seriously

Liberman and Pierrehumbert's (1984) hypothesis that post-lexical rules are actually phonetic rules, so that the surface inventory and allophonic 'categories' are epiphenomenal. This model has many consequences discussed elsewhere, but, here, crucially, it is not the case that, in order to get a model with three phoneme categories, the learner must first find six phonetic categories; rather, the learner will find three phonetic categories corresponding in this case to the lexical vowels of Kalaallisut. It is also not the case that the three phonetic categories discovered will each need to cover the entire phonetic space covered by *both* (retracted and non-retracted) allophonic variants; the retracted variants will be shifted to fall into the phonetic region covered by the unretracted ones.

The final assumption we make is a theory of possible underlying forms. Under the *Richness of the Base* theory, 'which holds that *all inputs are possible in all languages*, distributional and inventory regularities follow from the way the universal input set is mapped onto an output set by the grammar,' (Prince and Smolensky, 1993; emphasis added). One way to interpret this is to say that, a priori, no sequence of length N is more probable than any other. This has the consequence that, for some underlying sequence /ABC/ , $\Pr(/\mathrm{ABC}/) = \Pr(/\mathrm{A}/)\Pr(/\mathrm{B}/)\Pr(/\mathrm{C}/)$.

We believe that most of these assumptions are well justified. More importantly, we take up these assumptions in part because they allow us to highlight the Occam's Razor effect of Bayesian reasoning. While there are many benefits to be reaped from taking the theory of reasoning under uncertainty seriously, we believe that this particular point will be of deep interest to linguists.

21.3.4 The need for fewer categories: a bias in the prior

In this section we show how the simplicity preference inherent in Bayesian inference manifests itself in the prior by showing how a plausible set of assumptions about what it means to learn categories and grammars would force the abstract solution.

Following the reasoning given above, we compare two different models for the Kalaallisut vowel space: m_o, an opaque model that incorporates three phoneme categories and a system for deriving surface pronunciations, and m_t, a transparent model that contains six vowel phonemes and no opaque interactions, using a decision rule as in (20).

$$\text{(20)} \qquad \begin{array}{l} \dfrac{\Pr(D \mid m_o)\Pr(m_o)}{\Pr(D \mid m_t)\Pr(m_t)} > 1 : m_o, \\ \text{otherwise} : m_t \end{array}$$

Recall from the previous section that, ordinarily, in model comparison, the hypotheses under comparison each consist of a range of possible parameter values, and in order to compare the two models, we integrate over all parameter values. In the present case, this type of comparison would require far more involved mathematical analysis than is appropriate here. To get at the intuition behind the approach, we will thus attempt a simpler comparison, between two particular sets of parameter values under the two models, but taken in the abstract.

Recall also the fact that, if the likelihoods are equal under two models, model comparison will be driven by the priors. To illustrate the logic, we will assume this to be true in this section. This is of course not a reasonable assumption in general (otherwise the data would never have any effect on the outcome of learning), but it is at least plausible for the optimal solutions under either number of categories. In any case, it is a formal way of stating the bind we take ourselves to be in: the theory is truly underdetermined by the data,

to the point that neither model is a better explanation of the observation. In such a situation, in the model comparison rule in (20), $\Pr(D \mid m_o)$ is always equal to $\Pr(D \mid m_t)$, and we always get (21).

$$(21) \qquad \begin{array}{l} \dfrac{\Pr(m_o)}{\Pr(m_t)} > 1 : m_o, \\ \text{otherwise} : m_t \end{array}$$

A model of the phonetic/phonological grammar has several parts. First, we must know the number of categories, K. For m_o, we have $K = 3$; for m_t, $K = 6$. Second, there will be some grammar, G_o for m_o, G_t for m_t. Finally, we have some set of parameter values for each category in each model; for m_o, call these $\theta_{/\text{i}/,o}$, $\theta_{/\text{a}/,o}$, $\theta_{/\text{u}/,o}$, and call the whole collection C_o; for m_t, call them $\theta_{/\text{i}/,t}$, $\theta_{/\text{e}/,t}$, $\theta_{/\text{a}/,t}$, $\theta_{/\text{A}/,t}$, $\theta_{/\text{u}/,t}$, $\theta_{/\text{o}/,t}$, and call the whole collection C_t. (These parameter values, might, for example, be the means and covariance matrices of multivariate Gaussians.) We thus state the models as in (22).

$$(22) \qquad \begin{array}{l} m_o := \langle K = 3, G_o, C_o \rangle \\ m_t := \langle K = 6, G_t, C_t \rangle \end{array}$$

We can write out the function in (21) in terms of this parameterization and expand it using the chain rule of probability to obtain (23).

$$(23) \qquad \frac{\Pr(m_o)}{\Pr(m_t)} = \frac{\Pr(G_o \mid C_o, K = 3)\Pr(C_o \mid K = 3)\Pr(K = 3)}{\Pr(G_t \mid C_t, K = 6)\Pr(C_t \mid K = 6)\Pr(K = 6)}$$

This can be seen as three separate ratios. The leftmost ratio compares the two grammars. The ratio $\dfrac{\Pr(G_o \mid C_o, K = 3)}{\Pr(G_t \mid C_t, K = 6)}$ will be different from one to the extent that there is

an inherent cost to crucially derivational grammars (assuming that, apart from the ordering, the two grammars are the same); this cost might be different depending on the rest of the model, but, again, this bias, if any, would be an a priori one. For example, if there were a coherent rule-based analysis in which the two rules were in some sense 'unordered,' this would have twice the probability of either ordered rule analysis if the two orders were equally probable. In an ideal learner model, this is in fact a very reasonable way to spell out the intuition that the opaque system is 'hard to learn,' or that the learner would 'wait for certain data points'—like the crucial case of both rules applying across morpheme boundaries—to posit the opaque analysis. The intuition behind these statements is that, even though both the opaque and the transparent model can give the same strings, the transparent model is inherently preferred unless there is some data that it would not generate—that is, that has lower probability (perhaps not zero, though, since the learner can always treat it as noise).

The rightmost ratio, $\frac{\Pr(K=3)}{\Pr(K=6)}$, asks whether there is an inherent preference for some number of categories. We can think of this as being a bias inherent to Universal Grammar—are languages with three vowel categories treated as inherently more probable by learners than languages with six vowel categories? This is different from a bias driven by properties *of the deductive system*, as we will see.

Finding the values of these two ratios means solving two rather difficult empirical questions—indeed, this is so even if the null hypothesis is for the learner to be in some sense unbiased, because the structure of the model we assume will induce biases even if the precise details are all totally unknown. Without any knowledge about what these two biases are, let us leave their combined effect as a constant J. If J is less than one, the decision

will be biased in favour of the transparent analysis; if it is more than one, the decision will be biased in favour of the opaque analysis.

The interesting ratio here is $\frac{\Pr(C_o \mid K=3)}{\Pr(C_t \mid K=6)}$. Let us expand this factor in the decision rule.

$$\begin{aligned} & \frac{\Pr(G_o \mid C_o, K=3)\Pr(C_o \mid K=3)\Pr(K=3)}{\Pr(G_t \mid C_t, K=6)\Pr(C_t \mid K=6)\Pr(K=6)} \\ (24) \quad = & \left[\frac{\Pr(G_o \mid C_o, K=3)\Pr(K=3)}{\Pr(G_t \mid C_t, K=6)\Pr(K=6)}\right] \cdot \frac{\Pr(\theta_{/\mathrm{i}/,o}, \theta_{/\mathrm{a}/,o}, \theta_{/\mathrm{u}/,o} \mid K=3)}{\Pr(\theta_{/\mathrm{e}/,t}, \theta_{/\mathrm{A}/,t}, \theta_{/\mathrm{o}/,t}, \theta_{/\mathrm{i}/,t}, \theta_{/\mathrm{a}/,t}, \theta_{/\mathrm{u}/,t} \mid K=6)} \\ = & J \cdot \frac{\Pr(\theta_{/\mathrm{i}/,o}, \theta_{/\mathrm{a}/,o}, \theta_{/\mathrm{u}/,o} \mid K=3)}{\Pr(\theta_{/\mathrm{e}/,t}, \theta_{/\mathrm{A}/,t}, \theta_{/\mathrm{o}/,t}, \theta_{/\mathrm{i}/,t}, \theta_{/\mathrm{a}/,t}, \theta_{/\mathrm{u}/,t} \mid K=6)} \end{aligned}$$

The decision ratio in (24) compares (in addition to the fixed cost ratio for the rule ordering and the number of categories) the probability of the particular categories (parameter values of some phonetic probability distributions) recovered under each solution. Intuitively, the ratio will be smaller than one, because the set of three-category solutions is a proper subset of the set of six category solutions; each time we must estimate a new category, we add further uncertainty to the solution.

To make this true in our case, we need some assumptions. As discussed above, under the theory of Dillon *et al*. (to appear), the continuous input space in which the phonetic categories are fit has already had the effects of allophonic processes removed (of course, the categories must be learned simultaneously with the grammar). This means that, ideally, if we can find the true categories in the data, we should have $\Pr(\theta_{/\mathrm{i}/,t}, \theta_{/\mathrm{a}/,t}, \theta_{/\mathrm{u}/,t} \mid K=6)$ exactly equal to $\Pr(\theta_{/\mathrm{i}/,o}, \theta_{/\mathrm{a}/,o}, \theta_{/\mathrm{u}/,o} \mid K=3)$, because the recovered categories will be the same. Of course, as discussed in greater detail elsewhere, it might be the case that, under one or the other hypothesis, it is more difficult to find the true

categories (indeed, this is almost certainly the case); but, so long as there is no strong prior on the phonetic location and extent of categories, the two should be roughly equal.

This means that we can productively expand the decision rule in (21) and (24) using the chain rule. If, as we assume, some of the categories are shared between the two solutions and the probabilities cancel, then we have (25).

$$(25) \qquad J \cdot \frac{\Pr(\theta_{/i/,o}, \theta_{/a/,o}, \theta_{/u/,o} \mid K = 3)}{\Pr(\theta_{/e/,t}, \theta_{/A/,t}, \theta_{/o/,t}, \theta_{/i/,t}, \theta_{/a/,t}, \theta_{/u/,t} \mid K = 6)}$$
$$\approx J \cdot (\Pr(\theta_{/e/,t}, \theta_{/A/,t}, \theta_{/o/,t} \mid \theta_{/i/,t}, \theta_{/a/,t}, \theta_{/u/,t}, K = 6)^{-1}$$

Clearly, the second factor must be greater than one, because the probability inside the reciprocal can by definition be no more than one. We thus have a direct comparison: whatever the inherent cost of process ordering, and whatever inherent bias learners might have for more categories (if this is plausible), their combined value (some $J < 1$) must overcome the inherent cost of *estimating* three new categories in order for a transparent solution to get off the ground.

In order for this to be the case it would need to be that, at least given the correct estimates for the three categories /i/, /a/, /u/, the remaining three sets of parameter values were extremely probable. Assuming each to be equiprobable, they would each need to have (conditional) probability $\sqrt[3]{J}$. Even for apparently quite strong biases like $J = 10^{-3}$, we get that each set of parameter values would need to have probability 0.10, which indicates substantial bias toward certain phonetic categories.

21.3.5 An analysis of Kalaallisut underlying representations: a bias in the likelihood

In this section we build on the analysis of the previous section, applying the same reasoning

to a slightly different part of the problem. In particular, while in the previous section we assumed that the likelihoods were comparable under the two hypotheses, we will weaken that assumption here. We show how the same type of reasoning applies: when there are more things to estimate under a particular model, the probability of any individual solution under that model drops, so that, to the extent that the solutions under that model are roughly as good and as probable as under the simpler model, we should prefer the simpler model.

In particular, recall that the Bayes factor for model comparison is a ratio of two model probabilities, where each is as in (26).

$$\text{(26)} \qquad \Pr(M \mid D) = \frac{\Pr(D \mid M)\Pr(M)}{\Pr(D)}$$

In this section we focus on the fact that by hypothesis the underlying phonetic/phonological model M provides information about phonetic values only by way of phonological categories. If the model is relatively uninformative with respect to n-gram probabilities *of potential phonological strings*, then a model with more phonemes will assign lower probability to an individual string. This affects the likelihood, $\Pr(D \mid M)$, which we previously assumed to be roughly equal under the two hypotheses. In the extreme case, if the probability of a phonemic string—say /puq/—is simply the product of the probabilities of the individual phonemes, then the fact that having phoneme categories means greater uncertainty will mean smaller string probabilities. As discussed above, the assumption that all phonemic strings are equiprobable is roughly the Richness of the Base hypothesis of Prince and Smolensky, 1993. In this section we specify more precisely how such an assumption would interact with the kind of model comparison under discussion.

Given data D equal to some phonetic input x, the learner must compare models

using a Bayes factor constructed from (26). For x a single one-segment data point we have (27), where each c_i is one of the K phoneme categories.

$$\text{(27)} \qquad \Pr(M \mid x) = \frac{\left[\sum_{i=1}^{K} \Pr(x \mid c_i, M) \Pr(c_i \mid M)\right] \Pr(M)}{\Pr(x)}$$

The expansion in (27) says that each token might have been generated by any of the K phoneme categories, and that the learner (and the listener) must decide which; equality follows from the law of total probability. Similarly, if we consider $\bar{x}$ corresponding to a sequence of phonemes, we have (28), where w ranges over all possible underlying category sequences.

$$\text{(28)} \qquad \Pr(M \mid \bar{x}) = \frac{\left[\sum_{w} \Pr(\bar{x} \mid w, M) \Pr(w \mid M)\right] \Pr(M)}{\Pr(\bar{x})}$$

Making the assumption that the data consists of a sequence of independently drawn sequences of phonetic values (that is, that the probability assigned by the model to one phonetic string does not depend on the identity of the previous ones), we get that the learner will do model comparison using the Bayes factor in (29), where $\bar{x}$ ranges over all phonetic sequences in the data, and w ranges over all possible phonemic strings.

$$\text{(29)} \qquad \frac{\Pr(D \mid m_o)}{\Pr(D \mid m_t)} = \prod_{\bar{x}} \frac{\sum_{w} \left[\Pr(\bar{x} \mid w, m_o) \Pr(w \mid m_o)\right] \Pr(m_o)}{\sum_{w} \left[\Pr(\bar{x} \mid w, m_t) \Pr(w \mid m_t)\right] \Pr(m_t)}$$

As we know from the previous section, the rightmost factors (the priors) will tend to favour m_o by some amount. It is not clear a priori which of the two likelihood terms should

dominate. Note, however, that under certain assumptions the contribution of $\Pr(\bar{x} \mid w, m_o)$ as versus $\Pr(\bar{x} \mid w, m_t)$ will be nil. In particular, under the model of the phonetics–phonology interface discussed above, the interface categories are estimated using phonetic values corrected for the effects of allophonic processes. The consequence of this is that, in a three-category system, the one-to-many mapping from categories to phonetic values does not result in three large categories.

This is important, because, ordinarily, when fitting a mixture model, the choice between one category or two categories results in a roughly equal tradeoff between having greater or smaller mixing probabilities and requiring narrower or wider coverage. Figure 21.1 illustrates this. In Figure 21.1, a single Gaussian is overlaid with a pair of Gaussians having roughly the same coverage. Above each is shown a mixing probability, the probability of selecting that category. If we treat the two Gaussians as an alternate solution to the single Gaussian, then, clearly, the mixing probability in the single category solution will of necessity be greater than either of the individual mixing probabilities in the two category solution, because probabilities must sum to one. This will be traded off, however, against the fact that the single Gaussian will need greater coverage, and thus any individual value will be smaller, again because probabilities must sum to one. Thus, comparing the probability density at an individual point will come out roughly equal, and comparing individual segment likelihoods will be uninformative to the extent that the best fit under the two solutions has basically the same coverage.

On the other hand, under the model we assume, the single category phoneme model needs only to have the extent of one of the allophonic variants, not both. Thus, although the mixing probability is greater, the individual densities are not smaller, and comparing points

will favour the single-category solution.

In particular, if the likelihood values for individual points in phonetic space are roughly the same, we can say something about the comparison between $\Pr(\bar{x} \mid w, m_o)\Pr(w \mid m_o)$ and $\Pr(\bar{x} \mid w, m_t)\Pr(w \mid m_t)$, by comparing the probabilities of various underlying forms. In particular, we will get a model comparison ratio in which the important terms (the ones that differ between numerator and denominator) will be probabilities of underlying forms containing retracted vowels under m_t, but non-retracted vowels under m_o, as in (30).

(30)
$$\begin{aligned}
&\frac{\Pr(D \mid m_o)}{\Pr(D \mid m_t)} \\
&= \prod_{x \text{ with } [\ldots \text{eZ} \ldots]} \frac{\cdots + \Pr(x \mid /\ldots \text{iZ}_\text{o} \ldots /, m_o)\Pr(/\ldots \text{iZ}_\text{o} \ldots / \mid m_o)\Pr(m_o) + \cdots}{\cdots + \Pr(x \mid /\ldots \text{eZ}_\text{t} \ldots /, m_t)\Pr(/\ldots \text{eZ}_\text{t} \ldots / \mid m_t)\Pr(m_t) + \cdots} \\
&= \prod_{x \text{ with } [\ldots \text{eZ} \ldots]} \frac{\cdots + \Pr(/\ldots \text{iZ}_\text{o} \ldots / \mid m_o)\Pr(m_o) + \cdots}{\cdots + \Pr(/\ldots \text{eZ}_\text{t} \ldots / \mid m_t)\Pr(m_t) + \cdots}
\end{aligned}$$

The summation is over possible alternate underlying forms for x; by removing the likelihood term we make the simplifying assumption that we can basically ignore the 'incorrect' underlying forms potentially posited by the learner/hearer, and that the remaining likelihoods are roughly equal across all possible underlying forms in each model, and roughly equal across the two models. This is a stronger version of the assumption just discussed—that the probability of individual phonetic segments does not change under the two hypotheses; this is the crucial premise to our version of Occam's Razor, but now operating 'inside' the likelihood function.

The reasoning is now similar to the reasoning from the previous section. By the chain rule of probability, we obtain (31) from (30).

(31)

$$\prod_{x \text{ with } [\ldots eZ \ldots]} \frac{\cdots + \Pr(x \mid /\ldots iZ_o \ldots/, m_o)\Pr(/\ldots\_Z_o\ldots/ \mid /i/, m_o)\Pr(/i/ \mid m_o)\Pr(m_o) + \cdots}{\cdots + \Pr(x \mid /\ldots eZ_t \ldots/, m_t)\Pr(/\ldots\_Z_t\ldots/ \mid /e/, m_t)\Pr(/e/ \mid m_t)\Pr(m_t) + \cdots}$$

The assumption of Richness of the Base given above then crucially tells us the following:

(32) $$\frac{\Pr([eqa] \mid /iqa/, m_o)\Pr(/i/ \mid m_o)\Pr(/q/ \mid m_o)\Pr(/a/ \mid m_o)}{\Pr([eqa] \mid /eqa/, m_t)\Pr(/e/ \mid m_t)\Pr(/q/ \mid m_t)\Pr(/a/ \mid m_t)}$$

Given that the probability of the surface string is roughly the same under both grammars, this reduces to (33).

(33) $$\frac{\Pr(/i/ \mid m_o)}{\Pr(/e/ \mid m_t)}$$

Assuming a uniform distribution of segments, the fact that Kalaallisut has fifteen consonant phonemes gives us a ratio of $\frac{15+6}{15+3} \approx 1.17$, preferring the opaque solution. Clearly, the same will hold for any other sequence. Here, as above, then, we see that putting linguistic assumptions into a formal framework for decision making under uncertainty can often be informative; in this case, we see how simple principles of reasoning under uncertainty can take hold under the right circumstances to give interesting results that inform our understanding of general issues like abstractness in learning.

Note, however, that we are not finished: this was just one form. The model comparison ratio is a *product*, taken over the entire data set; this means that each data point will contribute by *multiplying in* its probability, which, being less than one, will shrink overlall probabilities exponentially. In the analysis of scientific data, Bayes factors are

usually compared as logarithms; comparisons of 3 or more in favour of a model are generally considered very strong evidence (Goodman, 1998). The log score in favour of an abstract model for a here is $N \log 1.17 + \log \frac{\Pr m_o}{\Pr m_t} \approx 0.154 + \log \frac{\Pr m_o}{\Pr m_t}$, where N is the number of data points. Clearly it will take very little time for this number to reach 3, regardless of how the model priors compare. The more times the learner must use its grammar to encode speech, the less probable that particular data set is.

As we have seen, this type of result falls out under the Bayesian approach to reasoning under uncertainty, because the Bayesian approach is to assume probability distributions over parameter vectors and models; as shown here, however, under certain models, this type of effect can even be obtained within the likelihood term, because the structure of certain models (like a model in which phonetic values are generated by discrete phonemes) implies a kind of 'hidden prior', in this case so that if the observed phonetic values are roughly equally probable under either model, we fall back on the probabilities of underlying phoneme sequences. The correct interpretation of this quantity is up for debate, but it is plausibly *not* informatively modeled under either hypothesis, leading us to conclude that, in the case of the number of phonemes in the model, the tendency to minimize the objects in the model is very strong.

One possible objection here is to our interpretation of the Richness of the Base. According to the Richness of the Base, the choice of phonological model does not affect the set of possible lexical encodings. Thus, one might conclude that we have a choice between /i/ and /e/ under *either* model. The consequence of this, however, depends on what it means to 'learn the discrete category /i/.' If the category /i/ is really just a point in a finite-valued feature space, and learning that there are only three categories in m_o simply

means learning that some feature is truly irrelevant (except at the phonetics–phonology interface, where its effect will be restored), then it is reasonable to suppose that an encoding of /i/ is still an encoding of /i/, regardless of the value for that feature. Thus when we talk about 'representations containing /i/,' we are referring to representations with either feature value, and are thus summing over both of the representations possible in m_t; the conclusion clearly does not change.

In any case, there will be tradeoffs to be made under any set of assumptions. If there is a substantially better fit to the phonetic data under one theory than another, then the improved fit will accrue in the same way, multiplying through for each data point; and, if there are some surface forms that are ambiguous under one theory but not another, then those points would be more probable under that theory, because they would have more possible sources. We would be satisfied, however, regardless of the correct answer, simply to have the debate about learnability take place at this level rather than in the realm of speculation.

21.4 Discussion

In this chapter we have shown how Bayesian reasoning applies to problems of inference in linguistics, which arise both in the context of normal scientific reasoning, and because inference is part of the object of study.

We selected a simple problem of phonological abstractness, in which more abstract solutions are pitted against solutions with more phonemes, to demonstrate an important feature of reasoning under probability theory, and, more specifically, Bayesian reasoning: more complex solutions are dispreferred, all other things being equal. In the case of

Kalaallisut, the opaque 3-vowel solution is preferred to the transparent 6-vowel solution because of biases in both the prior and the likelihood terms involved in model selection. From the point of view of the prior, models with fewer free parameters will always be preferred because of their relative representational simplicity. In addition to this bias in the prior, we argued that there is a bias in the likelihood as well. The core of this argument was that having a smaller phonemic vocabulary maximized the probability of the output. This means that the learner is in general more confident of any parse under a smaller, opaque vowel system. Furthermore, relatively low-confidence parses that result from a transparent vowel system are compounded every time the phonological grammar is used to parse speech, leading to a substantial increase in the bias towards opaque systems as the number of data points grows. Taken together, these results provide an argument for the opaque analysis of the Kalaallisut vowel system on the grounds of representational simplicity.

It is important to underline that our conclusions were based on a number of simplifying assumptions about the nature of phonology and phonetics as cognitive systems. Under other assumptions, or under a more realistic model comparison scheme, it is of course possible that we would have obtained different results.

More crucially, model selection is almost inevitably strongly dependent on the parameterization of the space of possible hypotheses. Even under extensionally equivalent theories of grammar with the same general architecture, we might conclude that some grammar is far less likely in one theory than another, perhaps because it requires substantially more machinery to state; changing the distributional assumptions for our phonetic categories (even changing how those distributions are parameterized) will, of course, also change the solution in general. We believe that this simply indicates that the current state of the art in phonology is inadequate for proper, complete model comparison.

If the various current models of interacting phonological processes could be reduced to their bare theoretical essentials and stated in a common metalanguage (for example, an automata-theoretic formulation along the lines of Heinz, 2007), then we would have a much clearer basis for comparison; arguments about the correct intensional statement of grammars would then to some degree be arguments about the priors.

Our study has been of an ideal solution to an inference problem, thus a study at the computational level in the sense of Marr (1982), in that it specified the learning problem precisely without giving an algorithmic account of how the learner would arrive at the ideal solution. This is a more abstract approach than has been taken in some other theoretical language acquisition literature. It is in the spirit of the *evaluation measure* theory of Chomsky and Halle (1968), in the sense that it attempts to specify a cost function for grammar induction without specifying a search algorithm. In this case, the difference does not appear to be important, since the grammatical part of the solution—the two processes in (1) and (7)—is the same across both models. In other cases, the search function might need to be exposed to certain crucial data points in order to 'discover' certain rules that would allow it to escape from local maxima in the cost function. Nevertheless, we believe that specifying the cost function first is a fruitful approach in any case.

Although we reach a similar conclusion to Chomsky and Halle—namely, that the evaluation measure includes an Occam's Razor like principle—it should be reiterated that the goal of the present work was to point out that such a principle follows from general principles of reasoning under uncertainty. Indeed, the 'Occam factor' obtained by Bayesian model comparison can be restated as (the limiting case of) a principle of Minimum Description Length (Rissanen, 1978), consonant with the counting-symbols cost function of Chomsky and Halle.

Furthermore, our goal has not been to show definitively that an abstract solution for this particular problem is correct, but simply that a tendency towards abstract solutions falls out from simple, domain-general assumptions about rational decision making: you will pay for every extra phoneme with every word, but you will only pay for the grammar once. We believe that a future approach to linguistic theory that attempts to find optimal statistical solutions to the problems of inference we face will be highly informative, since it touches on the fundamental issues of simplicity and abstractness. Perhaps contrary to expectations, abstractness is not inherently more costly or difficult for the learner; indeed, it may be optimal.

Table 21.1: The phonemic inventory of Kalaallisut, roughly following Rischel 1975. Length is also contrastive for both vowels and consonants (omitted here). The uvular nasal [ɴ] is marginal.

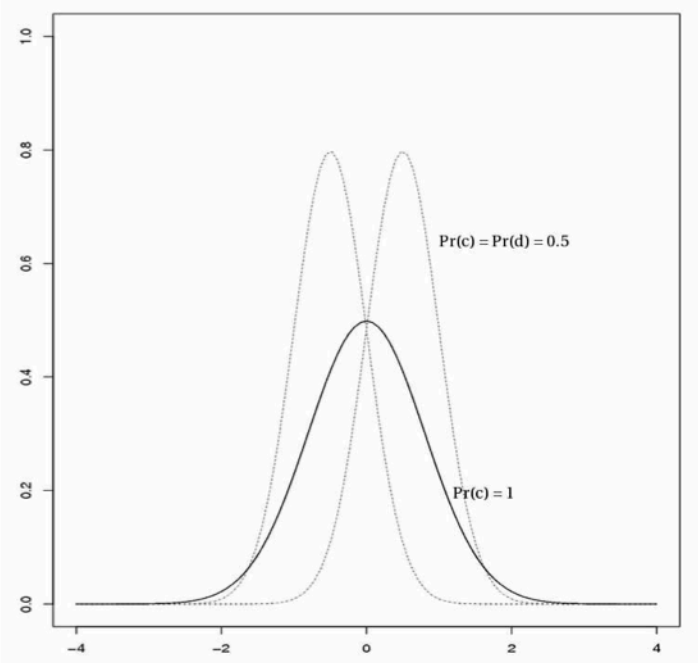
Bilabial	Coronal	Velar	Uvular
p	t	k	q
v	j l s ʃ	ɣ	ʁ
m	n	ŋ	ɴ

Vowels	
i	u
a	

Table 21.2: The phonemic inventory of Kalaallisut under a transparent analysis (length omitted as above). Position in the chart is not intended to suggest any particular featural analysis.

i		u
e		o
a	ɑ	

Figure 21.1: A two-component Gaussian mixture distribution as versus a single Gaussian of similar shape to the combination of the two smaller ones. The pair of smaller distributions will each individually give greater likelihood values than the single Gaussian (as shown by the height of the peaks), but this must be traded off against the mixing probabilities (probabilities of the categories) by which each data point must be multiplied in model comparison. If the single distribution only needed to be the width of one of the two components, however, the greater mixing probability for a single category would favour the single category solution because of the increased likelihood.



[1] Table 21.1 deviates from the inventory adduced by Rischel in that it omits an underlying voiceless fricative series. The question is irrelevant for current purposes, and the argument in favour of such an analysis would seem dated by modern standards, as it turns only on the

maintenance of the taxonomic phonemic level; see the original.

[2] The non-low retracted variants are notated in the standard Kalaallisut orthography as *e* and *o*; the two variants of the low vowel are collapsed in the orthography as *a*. Rischel (1975) describes the variants as being lowered and pharyngealized.