Simplifying subsidiary theory: statistical evidence from Arabic, Muna, Shona, and Wargamay

Colin Wilson
colin@cogsci.jhu.edu

Marieke Obdeyn
obdeyn@cogsci.jhu.edu

Johns Hopkins University

Abstract

Subsidiary features modulate the degree to which the Obligatory Contour Principle on place (OCP-Place) is violated by homorganic consonants. Statistical analysis of consonant co-occurrence patterns in four unrelated languages, combined with a method of model comparison that incorporates Occam's Razor, support a theory in which each place and subsidiary feature has a weight that contributes to gradient OCP-Place violation. Crucially, the weights of subsidiary features are not free to vary across place of articulation within a language. This **weighted feature** theory is more restrictive than the alternative proposed by Coetzee and Pater 2008b, which allows subsidiary features to vary in weight both across languages and across places within a language (an approach that has its roots in the non-quantitative theories of Yip 1989; Padgett 1991, 1995; McCarthy 1994). The weighted feature theory is less restrictive than the natural classes model of Frisch et al. 2004, but the greater descriptive freedom provided by language-particular weighting is motivated by the data even when Occam's Razor is taken into account. In addition to arguing for a novel theory of subsidiary features, this paper demonstrates that the method of evaluating analyses of co-occurrence data with respect to Observed/Expected (O/E) values — as originally proposed by Pierrehumbert 1993 and adopted by several recent papers in *NLLT* (Frisch et al. 2004, Coetzee and Pater 2008b, Anttila 2008, Coon and Gallagher 2008) — is mathematically flawed, as it confounds co-occurrence restrictions with positional probabilities. The main empirical claim of Coetzee and Pater 2008b, namely that correlations with the data of Arabic and Muna uniformly favor their weighted-constraint theory over that of Frisch et al. 2004, is based on the problematic O/E method and must be qualified. An alternative, statistically sound method of model evaluation and comparison based on probability theory is introduced and shown to support the weighted feature theory of OCP-Place over the alternatives considered.

Acknowledgments. We would like to thank Rahul Balusu, Luigi Burzio, Jenny Culbertson, Lisa Davidson, Jason Eisner, Bruce Hayes, Matt Goldrick, Kyle Rawlins, Paul Smolensky, Kie Zuraw, and the participants of the Linguistics colloquium at Yale and the the CLSP seminar at Johns Hopkins for helpful feedback. We are grateful to Andries Coetzee and Stefan Frisch for making data from Muna and Arabic available to us. Bruce Hayes deserves special acknowledgment for his role in the development of the grammatical framework that we adopt, as well as for making versions of the Shona and Wargamay data publicly accessible.

1 Introduction

The goal of research in quantitative phonology is to describe, and ultimately to explain, the knowledge that native speakers have about the categorical and gradient patterns found in linguistic sound systems (e.g., Kelly, 1991; Kelly and Martin, 1994; Pierrehumbert, 1993, 1994, 2001, 2003, 2006; Buckley, 1997; Coleman and Pierrehumbert, 1997; Kessler and Treiman, 1997; Frisch, 1996, 2000; Frisch et al., 2000; Frisch and Zawaydeh, 2001; Frisch, 2004; Anttila, 1997, 2002; Boersma, 1997; Nagy and Reynolds, 1997; Hammond, 1999, 2004; Zuraw, 2000, 2007; Albright and Hayes, 2003; Davidson, 2003, 2006; Ernestus and Baayen, 2003; Ernestus and Neijt, 2008; Goldwater and Johnson, 2003; Guion et al., 2003; Hay et al., 2003; Coetzee, 2004, 2006b; Jarosz, 2006b; McClelland and Wyk, 2006; Wyk, 2006; Goldsmith and Riggle, 2007; Goldsmith and Xanthos, 2009; Lee and Goldrick, 2008; Albright, 2009; Goldrick and Daland, 2009; Onnis and Christiansen, 2009). A central issue for this research program, like all others that are invested in explanation, is that of **model comparison** or **model/theory selection** (we use 'model' and 'theory' interchangeably throughout). Given two or more alternative models that have been assessed with respect to a given set of data, what are reasonable grounds for preferring or selecting one of them?

One common practice, adopted in many of the papers cited above, is to compare models solely in terms of statistics that measure the **fit** between the predictions of each model and the data. Commonly used statistics include various types of parametric or non-parametric correlation (e.g., Hocking 1996; Hollander and Wolfe 1999) and related values such as the proportion of variance explained. Application of such descriptive statistics is encouraged by the existence of algorithms that can fit complicated quantitative models and that are implemented in readily available software packages (such as Praat, Boersma and Weenink 2009; R, R Development Core Team 2008; BUGS, Lunn et al. 2000; and LENS, Rohde 1999). However, while such methods may be satisfactory for preliminary or exploratory investigation, or for cases in which the models under comparison have the same number and type of free parameters, they have notorious flaws when used to compare theories that differ in parametric complexity or functional form (see, for example, Cutting et al., 1992; Jacobs and Grainger, 1994; Myung and Pitt, 1997; Myung, 2000; Roberts and Pashler, 107; Duda et al., 2001; Pitt and Myung, 2002; Wagenmakers and Waldorp, 2006; Gluck et al., 2008). In essence, the problem is that more complicated models are generally more flexible or *less restrictive*, and can therefore achieve good fits or high correlations even if the data was in fact generated by a simpler model.[1]

The difficult issue of comparing models having different degrees of complexity/flexibility/ restrictiveness has been occasionally broached in quantitative phonology (e.g., Frisch et al. 2004, pp. 208-209; Coetzee and Pater 2008b, p. 329), but never addressed in a general way. Fortunately, research in other fields (primarily statistics, computer science, and psychology) has produced many methods of evaluation and comparison that balance agreement with the data against complexity of the model (e.g., Akaike 1973; Rissanen 1978, 1999; Grünwald 2000; Vitanyi and Li 2000; MacKay 1992, 2003; Hastie et al. 2001; Chen and Haykin 2002; Spall 2003; Grünwald et al. 2005; Pitt and Navarro 2005; see also Moreton 2008 for application of related ideas to the problem of acquisition). These methods can check the temptation to construct models of increasing complexity and flexibility in the pursuit of ever better fits to the data. In one form or another, they all incorporate a fundamental principle of scientific reasoning known as **Occam's razor** (e.g., Jeffreys and Berger 1992). They are also tightly connected to the Bayesian inference framework (e.g., Bernardo and Smith 1994; Gelman et al. 2004; Bishop 2006; Chater et al. 2006; Körding and Wolpert 2006; Doya et al. 2007; see especially MacKay 2003).[2]

[1]The relevant sense of 'restrictiveness' is typological: a model is restrictive to the extent that it places limits on the variation across languages (or, more generally, data sets). Restrictiveness in the sense of limiting the set of possible structures of a given language (i.e., avoiding overgeneration) is not at issue here.

[2]An alternative approach to penalizing less restrictive models is **cross-validation**, in which portions of the complete data set are held out during learning (e.g., Stone, 1974; Dietterich, 1997; Browne, 2000; Duda et al., 2001; Hastie

1.1 Model comparison and consonant co-occurrence

In this paper, we apply a well-known method of model selection, one based on the **Laplace approximation** (e.g., de Bruijn 1958; Tierney and Kadane 1986; Kass and Raftery 1995; MacKay 2003; Bishop 2006), to the theory of consonant co-occurrence restrictions. Rather than simply seeking the model that can best reproduce the quantitative patterns of consonant co-occurrence that are attested in natural languages, we aim for a theory that is compatible with known empirical patterns while being maximally restrictive. The Laplace approximation is useful for this purpose, because it combines a measure of fit to the data and a measure of restrictiveness into a single quantity. A principled approach to model comparison and selection then involves computing the quantity for each model under consideration and preferring models that achieve better values.

We focus specifically on the effects of the Obligatory Contour Principle on place (OCP-Place), and in particular on the role that **subsidiary features** play in determining the degree to which homorganic consonants violate OCP-Place. This empirical domain has been extensively studied within both non-quantitative and quantitative phonology (see §1.2 and §1.3 below). It is also the domain in which students of phonology are likely to first encounter statistical data and reasoning (e.g., McCarthy 1988; Kenstowicz 1994, pp. 163-165, 459-460; Katamba 2006; Aldrete and Frisch 2007). As we review below, a rather unrestrictive theory has developed around OCP-Place effects, with particular complexity in the treatment of subsidiary features. The main theoretical goal of this paper is to provide quantitative support for a simpler alternative, which is stated informally in (1) below (for formal development, see §2).

(1) **Weighted feature theory of OCP-Place violation**
The degree to which two homorganic consonants violate OCP-Place is determined by the similarity of the consonants, where similarity is defined as the sum of the (partly language-specific) weights of the place and subsidiary features on which the segments agree.

The evidence for our proposal comes from parallel statistical analysis of four unrelated languages. Two of the languages, Arabic and Muna (Western Austronesian), have been the subject of recent and influential quantitative studies by Frisch et al. (2004) (henceforth 'FPB') and Coetzee and Pater (2008b) (henceforth 'CP'). The main claim made by CP is that comparison of OCP-Place effects in Arabic and Muna motivates great flexibility in the treatment of subsidiary features. We will challenge this claim on both theoretical and methodological grounds, and provide additional support for our more restrictive alternative based on the previously unstudied OCP-Place patterns of Shona (Bantu) and Wargamay (Australian). We begin by reviewing the roots of CP's theory in non-quantitative phonology (§1.2), then turn to the substantially more restrictive theory of FPB (§1.3.1) and the proposal and model comparison presented in CP (§1.3.2). The organization of the rest of the paper is given in §1.4.

1.2 OCP-Place in non-quantitative phonology

Arabic verbal roots that contain two consonants of the same place of articulation are attested but relatively rare, a generalization first investigated in the contemporary era by Greenberg (1950) and McCarthy (1988, 1994). The seminal work by Greenberg and McCarthy has since been elaborated upon and extended to many other languages (on Arabic see also Yip 1989; Padgett 1991, 1995; Pierrehumbert 1993; Elmedlaoui 1995; Frisch 1996, 2000, 2004; Frisch et al. 2004; Bachra 2000; Coetzee and Pater 2008b; on related patterns in other languages see several of the references just cited as well as Mester 1986, 1988; Lamontagne 1993;

---

et al., 2001). Because more complex models are likely to 'overfit' the subset of the data to which they are exposed, they can show poorer generalization performance on the held-out set. We do not adopt this approach, preferring to let the models learn from all of the available data and imposing the penalty for lack of restrictiveness explicitly.

Berkley 1994a, 2000; Buckley 1997; MacEachern 1999; Coetzee 2005, 2008; Kawahara et al. 2006; Coon and Gallagher 2008; Ito 2007; Pozdniakov and Segerer 2007; Walter 2007; Anttila 2008; Dmitrieva and Anttila 2008; Kager et al. 2008; Graff and Jaeger 2009).[3]

McCarthy (1988, 1994) links the restriction on identical place features to a more general restriction on co-occurrence of identical elements, the **Obligatory Contour Principle** (OCP; Leben 1973; Goldsmith 1976; McCarthy 1979, 1981, 1986), therefore the constraint is generally referred to as OCP-Place. As already noted, we focus exclusively on OCP-Place effects in this paper, but anticipate that the main idea behind our proposal would readily generalized to OCP effects on other features borne by consonants (e.g., McCarthy 1988; Yip 1989; MacEachern 1999) and to identity or partial identity effects more generally (e.g., Ito 1984; McConvell 1988; Goodman 1992; Hyman 1995; Suzuki 1998; Hansson 2001; Rose and Walker 2004; see also FPB, pp. 214-215 for relevant discussion).[4]

Even the earliest studies of OCP-Place by Greenberg and McCarthy, though not set within an explicit quantitative theoretical framework, acknowledge that the evidence for the constraint is statistical (e.g., McCarthy 1994, p. 204). That is, the evidence comes not from categorical absence of all structures that violate OCP-Place, but rather from numerical measures indicating that such structures are relatively infrequent. This finding, which is replicated in many other languages (e.g. Pozdniakov and Segerer, 2007), makes the domain of OCP-Place effects an important test case for theories of quantitative phonology. What makes the case particularly interesting is the further finding — also made in the early studies by Greenberg and McCarthy on Arabic and subsequently replicated in other languages — that the OCP-Place restriction does not apply with equal strength to all sequences of homorganic consonants. Specifically, homorganic consonants are subject to stronger OCP-Place restrictions if they agree on one or more of a set of *non-place* features (e.g., McCarthy 1994; Yip 1989; Padgett 1991, 1995). Following Padgett (1991, 1995), we refer to the non-place features that influence the strength of OCP-Place as **subsidiary features**.

In non-quantitative phonology, the theory of subsidiary features has limited content. There have been some proposals to restrict the set of universally allowable subsidiary features to those that stand in a particular feature-geometric relation with the place features (e.g., Yip, 1989; Padgett, 1991, 1995). However, such restrictions do not follow from any broader generalization about featural relations, and in any event have been challenged, sometimes implicitly, by recent quantitative work.[5] These proposals aside, there have been few attempts to restrict the deployment of subsidiary features within and across languages. In particular, it has been generally accepted since the work of Yip (1989), Padgett (1991, 1995) and McCarthy (1994) that the set of subsidiary features can be stipulated on a *language-specific* basis, and moreover on a *place-specific* basis within a given language. For example, McCarthy (1994, p. 206) specifies that in

[3]See Bachra (2000, chapter 3, pp. 25-30) and FPB (pp. 212-213) for reviews of other work not cited here.

[4]Previous research has established that perfect identity cannot be uniformly analyzed as the limiting case of similarity: a language that disallows similar but not identical consonants in a particular configuration may nevertheless allow identical consonants (e.g., McCarthy 1986; Mester 1986, 1988; Berent and Shimron 1997; MacEachern 1999; Gafos 2003; Coon and Gallagher 2008; see also FPB, p. 189, and CP, pp. 295ff., for discussion relevant to Arabic and Muna, respectively). We do not attempt to treat the representation of, or constraints on, perfect identity in this paper.

[5]Specifically, Padgett (1991, 1995) argues that only features in the **extended articulator group** — [sonorant] and all dependents of the place features (including [continuant] in Padgett's theory of the geometry of stricture features) — can be subsidiary. Padgett (1995, p. 185) cites an unpublished paper by McCarthy in which it is suggested that only stricture features can be subsidiary. See Yip (1989) for an earlier proposal that differs in some respects from those of Padgett and McCarthy. None of these proposals would allow [voice] or other laryngeal features to act as subsidiary features, contrary to the evidence from Arabic presented by Pierrehumbert (1993) and FPB (pp. 195-196, 218), from Muna by CP (p. 305), and from Yamato Japanese by Kawahara et al. (2006). Earlier analyses by Kenstowicz (1986) and (Mester, 1988, pp. 103ff.) can now be understood as claiming that [nasal] and [voice] are subsidiary features in Javanese, though their original formulations are in terms of feature-geometric dependencies. Pierrehumbert (1993), following a suggestion made briefly by Lightner (1973, pp. 58-59), proposes that all features contribute to similarity for the purposes of evaluation by OCP-Place.

Arabic the features [sonorant] and [continuant] are subsidiary (i.e., relevant for the application of OCP-Place) for coronals but not for consonants of other places of articulation. Similarly, Yip (1989) states of Arabic that "[t]he Labials . . . do not distinguish between [+continuant] /f/ and [-continuant] /b m/, but form a single identity class" (p. 361), whereas "within the Dorsal or Coronal class two similar [continuant] values may not cooccur" and furthermore "within the Coronal class two similar values of [sonorant] may not cooccur" (p. 369): that is, two Coronal (or Dorsal) consonants are allowed to co-occur unless they also agree on [continuant] or [sonorant]. Padgett (1995) encapsulates this unrestrictive approach to subsidiary feature relevance as follows: "This set of features may be stipulated for any articulator in a given language - sometimes none are stipulated" (p. 174).[6]

The level of language-particular stipulation expressed in the quote from Padgett (1995) has become a literally textbook part of the theory of OCP-Place. For example, in partially recapitulating the analysis of McCarthy (1991) (an unpublished draft of McCarthy 1994), Kenstowicz (1994, p. 164) states the version of the constraint for Arabic as follows:

(2) articulator *[x] [x]

(where x = Labial, Dorsal, Coronal, Coronal)

$$\begin{array}{cccc} & & \text{Coronal}, & \text{Coronal}) \\ & & | & | \\ & & [+\text{sonorant}] & [-\text{sonorant}, \alpha\text{continuant}] \end{array}$$

One way of formalizing the theory of subsidiary features just reviewed (though none of the works cited above have discussed it in precisely these terms) is as a system of binary (ON/OFF) parameters, *with one parameter for each place and subsidiary feature combination*. Two homorganic consonants would then violate OCP-Place iff they agree with respect to all of the subsidiary features whose parameters are set to ON for place of articulation that they share. As already noted, we know of no proposal within the analytic tradition reviewed above that would predict that the parameter for [sonorant] is set to ON for Coronal place, but OFF for Labial and Dorsal place, in Arabic. More generally, we know of no systematic attempt within non-quantitative phonology to predict which subsidiary features are relevant for a given language or a given place within a language (though see Yip 1989, p. 361 for a preliminary suggestion based on contrastivity). This aspect of OCP-Place theory thus seems an appropriate target for a methodology, like the one we set out at the beginning of this paper, that emphasizes the balance between descriptive success and explanatory restrictiveness.[7]

We will argue against the widely-held view that subsidiary features can be place-specific, proposing instead that the relevance or 'salience' or 'weight' of a feature must be the same for OCP-Place evaluation across all places within a given language. This position may seem obviously wrong in light of well-known facts, such as the greater freedom with which coronal obstruents and sonorants (e.g., [d] and [n]) combine, in contrast to other homorganic obstruents and sonorants (e.g., [b] and [m]), in Arabic verbal roots. (This and other differences among the places in Arabic can be seen, for example, in Figure 12.1 of McCarthy 1994, p. 204). However, we will demonstrate that quantitative phonology has a previously untapped capacity to

[6]McCarthy (1994, p. 206) credits Dresher (1989), Selkirk (1991), and Padgett (1991) (the unpublished version of Padgett 1995) as originators of the claim that "the scope or domain of the OCP [here, OCP-Place] can be limited by statements restricting its applicability to a particular feature in a particular context" (p. 206). We do not have access to the unpublished commentary/handouts of Dresher and Selkirk. The relevant portion of Padgett's earlier work is Padgett (1991, pp. 316-362), which contains material similar to that in Padgett (1995, pp. 173-202).

[7]Our line of argument would extend to more recent non-quantitative constraint theories of OCP effects, such as Alderete (1997, 2004)'s and Ito and Mester (2003)'s *local conjunction* approach set within Optimality Theory (OT; Prince and Smolensky 2004), if these theories were to permit subsidiary features to be freely stipulated for each place.

predict such patterns without stipulating a difference in subsidiary relevance across articulators (see §2, §4).

1.3 OCP-Place in quantitative phonology

The proposals of FPB and CP define two extremes — one very restrictive, one quite flexible — between which our own theory of OCP-Place lies. Before formalizing our proposal, we briefly review these two previous theories and discuss their relationship to the goals of explanation and description in quantitative phonology.

1.3.1 Frisch, Pierrehumbert and Broe 2004

In a departure from the analytic tradition of McCarthy, Yip, Padgett, and others, Pierrehumbert (1993), Frisch et al. (1997), and FPB develop a quantitative theory of OCP-Place in which subsidiary behavior is derived from language-specific consonant inventories (see also Broe 1993; Frisch 1996, 2000, 2004). We concentrate on FPB's proposal, which we take to supersede the earlier work. (Our own theory is somewhat closer to that of Pierrehumbert 1993.)

FPB claim that OCP-Place effects are grounded in the broader principle of *similarity avoidance*, and that the proper measure of similarity is the **natural classes metric** (FPB, pp. 196ff.). According to this metric, the similarity of two members x and y of a consonant inventory is given by the ratio of the number of natural classes that include both x and y to the number of natural classes that include x or y. Because this similarity metric is specific to OCP-Place, FPB (p. 198, note 5) limit the set of natural classes to those that are defined with at least one place feature. Letting $\mathcal{NC}$ to denote this set, the similarity of two consonants is formally given by:

(3) $$sim_{\mathcal{NC}}(x, y) = \frac{\#\{C \in \mathcal{NC} : x \in C \wedge y \in C\}}{\#\{C \in \mathcal{NC} : x \in C \vee y \in C\}}$$

where $\#S$ denotes the size (cardinality) of set S. The value of $sim_{\mathcal{NC}}(x, y)$ ranges from 0 to 1 (FPB, p. 198): identical consonants have the highest possible value of 1; consonants that share no place feature have the lowest possible value of 0, since $\mathcal{NC}$ contains no classes in which such consonants are grouped together; non-identical homorganic consonants have intermediate values.[8]

The relation between natural classes similarity and degree of OCP-Place violation is essentially one of identity, as expressed in the following equation:

(4) $$\text{OCP-Place}_{\text{FPB}}(x, y) = sim_{\mathcal{NC}}(x, y)$$

Technically, FPB take OCP-Place violation to be a more complicated transformation of (3), with the two quantities being related by a monotonically increasing step function (FPB, pp. 204ff.) or a multi-parameter logistic function (FPB, p. 209; see also Frisch 1996, 2000, 2004). However, we have found these and other functions of similarity to provide little, if any, descriptive benefit. Therefore, our model comparison in §4 will use the parameter-free formulation in (4).

The aspect of FPB's proposal that we wish to highlight is that it allows subsidiary features to vary in their influence both across and within languages, but in a very restrictive way. Variation arises from differences in consonant inventories across languages, and from differences in consonant sub-inventories

[8] For other work that adopts or evaluates the natural classes metric, see Broe 1993; Frisch 1996, 2000, 2004; Berkley 1994b,a, 2000; Bailey and Hahn 2001, 2005; Albright and Hayes 2003; ud Dowla Khan 2006.

across places of articulation within a language. The impact of such (sub-)inventory differences on natural classes similarity is discussed extensively by FPB (pp. 199ff.). Restrictiveness follows from the fact that the contribution contribution of each place and subsidiary feature to the degree of OCP-Place violation is *determined* by the contents of the inventories and sub-inventories, not by independent parameterization, weighting, or other stipulation. This approach thus stands in sharp contrast to theories in which subsidiary status is orthogonal to other aspects of the phonological system. It would, if descriptively successful, satisfy our main goal of restricting subsidiary theory.

1.3.2 Coetzee and Pater 2008

In an extended series of talks and papers (e.g., Coetzee and Pater 2005; Coetzee 2006a; Coetzee and Pater 2006), culminating in CP, Coetzee and Pater argue that the theory of FPB is *too restrictive*, and in particular not flexible enough to provide a satisfactory account of OCP-Place effects in Muna. CP's own proposal adopts the tenets of the earlier non-quantative theory reviewed in §1.2 — including the assumption that the relevance of subsidiary features can be stipulated separately for each place of articulation — and further enriches that theory with numerical weighting.

The set of universal OCP-Place constraints that CP propose is given in (5) below (repeated from CP, (20), p. 316). These constraints are derived by fully crossing the place features {Labial, Coronal, Dorsal, Pharyngeal} (denoted by { P, T, K, H }) with the subsidiary features {[sonorant], [stricture], [voice], [emphatic], [prenasalization]} (denoted by {SON, STRIC, VCE, EMPH, PRE}). In addition, there is one constraint for each place of articulation that does not mention any subsidiary feature (i.e., *P-P, *T-T, *K-K, *H-H). A constraint in (5) is violated once for every sequence of two consonants in a root that agree on all of the features that the constraint mentions. For example, *T-T-SON is violated by root consonant sequences that agree on both Coronal and [sonorant].[9]

(5)

*P-P	*P-P-SON	*P-P-STRIC	*P-P-VCE	*P-P-EMPH	*P-P-PRE
*T-T	*T-T-SON	*T-T-STRIC	*T-T-VCE	*T-T-EMPH	*T-T-PRE
*K-K	*K-K-SON	*K-K-STRIC	*K-K-VCE	*K-K-EMPH	*K-K-PRE
*H-H	*H-H-SON	*H-H-STRIC	*H-H-VCE	*H-H-EMPH	*H-H-PRE

These constraints are given independent, language-specific weights, and used along with other constraints to compute Harmony and Optimality in the way specified by Harmonic Grammar (HG; Legendre et al. 1990a,b; Smolensky and Legendre 2005; Pater 2009; see §3 of CP for a review). The Harmony of an input-output pair in HG is equal to the weighted sum of its constraint violations. For the purpose of comparing CP's theory with others, it will be useful to encapsulate the part of the sum that is due to the constraints in (5). We refer to all of these constraint together with their weights as OCP-Place$_{\text{CP}}$, and use OCP-Place$_{\text{CP}}(x, y)$ to denote the total weighted violation incurred by the co-occurrence of consonants x and y.

In principle, the quantitative OCP-Place effects of a language could be accounted for with the OCP-Place$_{\text{CP}}$ values directly. CP argue for a different approach in which quantitative data is related to the **Acceptability** of a form, defined as the Harmony of the form minus the Harmony of the most harmonic distinct output for the same input (CP, pp. 311-314). Under the common assumption that phonotactically legal outputs are identical to their inputs (e.g., Hayes 2004; Prince and Tesar 2004; CP, p. 319), with phonotactic illegality resulting from unfaithful mapping, Acceptability values have a simple relation to constraint weights in the present context. The only other constraint employed in CP's analyses is the Faithfulness constraint IDENT-PLACE (defined as in Correspondence Theory, McCarthy and Prince 1995, 1999). Therefore,

[9]Sequences of identical consonants do not violate these constraints; for relevant discussion, see §1.2 and §4.

to determine the Acceptability of the co-occurrence of consonants x and y in the same root, simply subtract the negated weight of IDENT-PLACE from OCP-Place$_{\text{CP}}(x, y)$ (see CP, pp. 322, 325). In the case studies of §4, we consider both raw Harmony values and Acceptability values calculated in this way.

Regardless of exactly which grammatical quantities are related to the data, it is the mutually independent weights of the constraints in (5) that gives CP's theory a degree of descriptive freedom that meets and exceeds the level of earlier, parametric approaches to subsidiary features. For example, there is no grammatical pressure for the weights of *P-P and *T-T, or those of *P-P-SON and *T-T-SON, to be similar or 'tied' in a given language. Thus the subsidiary-feature difference between Labial and Coronal place in McCarthy (1994)'s analysis of Arabic verbal roots can be reconstructed by assigning a large weight to *P-P and *T-T-SON, and a small or zero weight to *T-T. It would be equally possible, in both the earlier parametric theory and CP's model, to describe a language in which (say) only [voice] is a subsidiary feature for Labial place and only [sonorant] is a subsidiary feature for Coronal place.

Where CP's theory goes beyond the earlier parametric approach is in making the precise *numerical* significance of each subsidiary feature a matter of language- and place- specific stipulation. The constraint weights that CP report for Arabic and Muna show that this greater flexibility is exercised in actual analyses: that is, there are many instances in which constraints that refer to the same subsidiary feature, such as *P-P-SON and *T-T-SON, take on quite different (non-zero) weights within a language (see CP on Arabic, (33), p. 326, and on Muna, (24), p. 320). In general, CP's theory predicts that there should be no systematic qualitative or quantitative relation among the places of a given language with respect to subsidiary effects.

CP claim that correlations with the Arabic and Muna data support their less restrictive theory over that of FPB. The r^2 values that they report (see CP, Table 14, p. 327) certainly seem convincing: for Arabic, CP's Acceptability values are reported to account for 20% more of the data variance than FPB's Similarity values when all non-identical homorganic consonant pairs are considered ($r^2 = .40$ vs. $r^2 = .20$). For Muna the corresponding difference in variance explained is reported to be 19% ($r^2 = .55$ vs. $r^2 = .36$).

These results, together with the fact that constraints referring to the same subsidiary feature have different weights in actual analysis, appears to cast doubt on the possibility of a simplified subsidiary theory. Indeed, one might pessimistically suppose that a theory that has even more descriptive flexibility than CP's is required, with (5) augmented by constraints that refer to particular values of the subsidiary features (e.g., *T-T-[+SON]) or by constraints that refer to more than one subsidiary feature simultaneously (e.g., *T-T-SON-STRIC). For analyzes that employ such constraints, see FPB (p. 194), Coetzee and Pater (2006), and Anttila (2008); CP do not take a strong position on the necessary level of specificity of individual OCP-Place constraints (CP, p. 317).[10]

1.4 Outline of the paper

The main argument of this paper begins where the model comparison reviewed in the previous subsection ends. We aim to make three points:

- Contrary to the central claim of CP, there is no compelling *correlation-* (or *variance-*) *based* evidence that favors the Harmonic Grammar model of CP over the natural classes similarity model of FPB.

[10] Anttila (2008) adopts a constraint set similar to that of CP (it is based on the earlier version Coetzee and Pater 2006) but employs partial OT rankings, together with the **Complexity Hypothesis** (Anttila 2008, p. 696), instead of weighting. While it is difficult to directly evaluate the restrictiveness of Anttila's proposal relative to that of the other theories we discuss, the constraint set used does include members that refer to a specific place of articulation together with one or more subsidiary features (e.g., Anttila's (35), p. 713). Since it is place-specific subsidiary stipulations that we are principally interested in eliminating, our arguments with respect to CP should carry over to Anttila (2008) modulo the weighting vs. ranking difference. For a critique of Anttila's proposal, see Pater (2008a).

The spurious claim of CP results from the application of a flawed method for evaluating theoretical proposals against co-occurrence data, the 'Observed/Expected' method originally due to Pierrehumbert (1993), as well as the failure to recognize the importance of complexity in theory comparison. When these methodological flaws are corrected, the correlation data from Muna and Arabic is essentially equivocal: CP's theory performs somewhat better on Arabic and — surprisingly given the empirical focus of CP's paper — FPB's theory performs somewhat better on Muna. The precise difference in fit to the data depends on the statistic computed (e.g., parametric or non-parametric correlation) and the subset of the data examined (e.g., all consonant pairs or only homorganic pairs). In light of the large difference between the theories in the number of free parameters, and hence in relative restrictiveness, we think the most reasonable conclusion to draw is the opposite of CP's. The natural classes model of FPB is supported by correlations with consonant co-occurrence in the two languages that CP examine.

- However, there is *probability-based* evidence in support of the constraint set of CP over the similarity theory of FPB, and this evidence remains intact even when Occam's Razor, as formalized by the Laplace approximation, is incorporated into the model comparison.

This evidence is available only under an explicitly *probabilistic* approach to gradient phonotactics — a kind of formal analysis that no previous approach to OCP-Place, including FPB and CP as well as Anttila (2008), has adopted. Specifically, we show that CP's proposal outperforms FPB's on correlations and other measures when it is placed within the stochastic grammar framework of Maximum Entropy (MaxEnt) and its constraints are weighted according to Bayesian principles. Because the resulting fits to the data are uniformly superior to those achieved by CP within Harmonic Grammar, with no apparent increase in theoretical or parametric complexity, these findings also support MaxEnt over HG as a framework for analyzing gradient phonotactics.

- Finally, the weighted feature theory of subsidiary features that we propose outperforms the theories of FPB and CP on the data from all four of the languages that we examine (Arabic, Muna, Shona and Wargamay).

According to the weighted feature theory, each place and subsidiary feature is associated with a (non-negative) real number. The degree to which two homorganic segments violate OCP-Place is equal to the sum of the weights of the features on which the segments agree. Like the weights of CP's constraints, these feature weights are at least partly language-specific and not derived from other aspects of the phonological system such as the inventory. But, in contrast to CP's model and the non-quantitative theories that preceded it, the contribution made by a given subsidiary feature is constant across all places of articulation in a given language (i.e., place $\times$ subsidiary 'interactions' weights disallowed). The weighted feature theory thus occupies an intermediate level of typological restrictiveness relative to previous approaches to OCP-Place.

Taken as a whole, our results indicate that probability theory is of central importance both for formalizing quantitative phonology and for determining the appropriate level of complexity or flexibility of particular phonological theories. This conclusion converges with the methods and results of many other studies in linguistics and the cognitive and brain sciences (e.g., Smolensky 1986; Oaksford and Chater 1998; Zuraw 2000; Tenenbaum and Griffiths 2001; Pierrehumbert 2001; Bod et al. 2003; Chater et al. 2006; Körding and Wolpert 2006; Ma et al. 2006; Norris 2006; Doya et al. 2007; Yang and Shadlen 2007; Clayards et al. 2008; Orbán et al. 2008).

The rest of the paper is organized as follows. We first state our weighted-feature theory of OCP-Place at the same level of explicitness with which we have described the other quantitative theories (§2). We then address the methodological issue of how different quantitative theories of consonant co-occurrence should be evaluated and compared with respect to frequency data (§3). The same section introduces the

Laplace Approximation and the MaxEnt grammar formalism, essential components of our argument that were mentioned but not defined above. The evaluation of our model and those of FPB and CP with respect to the co-occurrence data of the four languages is then reported, substantiating the claims made above (§4). We conclude by addressing some of the issues raised by our work, including the prospect of further restricting subsidiary theory by limiting the variation in feature weights across languages, and the issue of unifying the analysis of phonotactics with that of alternations in MaxEnt grammars.

2 Weighted features and similarity

Like FPB, we take degree of violation of OCP-Place to be determined by segment similarity. Our similarity metric differs from that of FPB in two ways. First, similarity is defined in terms of features rather than natural classes (cf. FPB, p. 197; 201). Second, the contribution of each feature to similarity is controlled by a (partly) language-particular weight rather than being derived from the segment inventory or other aspects of the phonological system.

The idea of defining similarity in terms of weighted features has been fruitfully applied to many non-linguistic domains (e.g., Tversky 1977; Medin and Shaeffer 1978; Nosofsky 1986; Navarro and Lee 2004; Navarro and Griffiths 2008). Within linguistics, Ernestus and Baayen (2003) employ weighted features in one of their models of Dutch past-tense formation, ud Dowla Khan (2006) argues that feature weighting provides an account of fixed segment choice in East Bengali reduplication that is superior to the natural classes metric, and Pierrehumbert (1993), FPB (p. 204), and Bachra (2000) all mention the possibility of analyzing OCP-Place restrictions with weighted features (though they stop short of proposing explicit models of this sort). It should also be noted that phoneticians have long observed that certain features carry greater perceptual weight (e.g., Miller and Nicely 1955; Ladefoged 1969; Wright 2004). Our adoption of a weighted feature theory of similarity therefore carries no great novelty. We are simply the first to develop this widespread idea into an explicit theory of OCP-Place.

To state the proposal formally, let $\mathcal{P}$ be the set of place features (e.g., the active articulator features of Sagey 1986; McCarthy 1988, 1994; Halle 1992, 1995, 2005), $\mathcal{S}$ be the set of non-place subsidiary features, and $\mathcal{F}$ be the union of $\mathcal{P}$ and $\mathcal{S}$. We do not attempt to derive $\mathcal{P}$ and $\mathcal{S}$ in this paper, but simply adopt the place and subsidiary features used by CP except for [emphatic]: thus $\mathcal{P}$ = {Labial, Coronal, Dorsal, Pharyngeal} and $\mathcal{S}$ = {[sonorant], [stricture], [voice], [prenasal]}.[11]

The similarity of two segments x and y is now defined as the sum of the **weights** of the features in $\mathcal{F}$ on which x and y agree:

(6) $$sim_{\mathbf{w}}(x, y) = \sum_{F \in \mathcal{F}} w_F \cdot \delta_F(x, y)$$

[11] CP's [emphatic] subsidiary feature serves the same purpose as the [acute] feature of FPB (pp. 200-201): namely, to distinguish the non-emphatic coronal obstruents [t d s z] from the emphatic coronals [tˤ dˤ sˤ zˤ] in Arabic. But this distinction is most likely one of simple vs. complex place of articulation, not one of non-place subsidiary features (for discussion, see McCarthy (1994); Kenstowicz (1994, p. 456-460); Bachra (2000); and FPB, p. 195). Rather than adopting a cover feature of uncertain status, we leave the issue of the substantive specification of emphatics to further research on the phonetics and phonology of these sounds (e.g., Goldstein 1994; Jongman et al. 2007; Zeroual et al. 2007). See the end of §4.1 for further discussion of emphatics.

The other subsidiary features mentioned in the text are fairly standard. Note that [stricture], which replaces [continuant] (CP, p. 299), has three values: narrow (e.g., stops), intermediate (e.g., fricatives), and wide (e.g., liquids and glides). See Padgett (2002b, 2008) for related work on the representation of constriction degree in phonology. CP make use of contrastive underspecification (e.g., Steriade 1995), and we follow them in this regard for the purposes of theory comparison; see, for example, the Muna feature chart given in CP, Appendix A, p. 332.

where **w** is the vector of feature weights, w_F stands for the weight of feature F, and $\delta_F(x, y)$ is equal to 1 if x and y agree on F and 0 otherwise. Feature weights are required to be non-negative in order to ensure that, universally, similarity is a non-decreasing function of feature overlap. A negatively weighted feature, if such were allowed, would make segments that agree on it *less* similar than segments that disagree. We do not think such dissimilarity-by-agreement effects are found in natural phonological systems.[12]

In our theory, as in the most straightforward application of FPB's proposal, the OCP-Place violation incurred by the co-occurrence of two (non-identical) homorganic segments x and y in the same root is equal to the similarity of x and y. Non-homorganic segments of course do not violate OCP-Place. The constraint is defined formally in (7) below.[13]

(7) $$\text{OCP-Place}_{\mathbf{w}}(x, y) = \begin{cases} sim_{\mathbf{w}}(x, y) & \text{if } \exists P \in \mathcal{P} \text{ such that } \delta_P(x, y) = 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that the set $\mathcal{P}$ of place features acts as a precondition or **gating** factor in the constraint. If two segments agree on one or more place features, then all of the features in $\mathcal{F}$ (place and subsidiary alike) contribute to the degree of OCP-Place violation. But if two segments have no place feature in common, all of their other specifications are moot — the degree of OCP-Place violation is identically 0. The structure of our constraint thus recalls the 'two-step' process of OCP-Place evaluation proposed by Yip (1989, p. 361, 369-370) and Padgett (1995, p. 181) (see also FPB, note 5, p. 198). Segments x and y are first checked for homorganicity; iff this test is passed, further calculation is performed on their features.

We take the gating function of place to be a significant empirical generalization about phonological systems, one that is implicit in the segment classes of Greenberg (1950) and that has been formalized in the work of McCarthy (1988, 1994); Yip (1989); Padgett (1991, 1995) and others (see also Padgett 2002a, pp. 99-102 for related discussion). For careful demonstration that subsidiary feature effects are parasitic on place agreement — rather than being reducible to independent OCP constraints on the subsidiary features themselves (e.g., hypothetical OCP-Sonorant or OCP-Nasal) — see Yip (1989, p. 369-370), Pierrehumbert (1993, ms. p. 3-4), Padgett (1995, p. 173, 178, 196-200), and CP (p. 301).[14]

---

[12]The only subtlety to (6) lies in applying the notion of feature **agreement** to underspecified segments (on the possibility of underspecification, see the previous footnote). Although not explicitly stated in CP, it is evident from their tableaux (e.g., their (26), p. 320, and (27), p. 321) that two segments specified $[-F]$ and $[0F]$, or $[0F]$ and $[0F]$, should be taken to agree on feature F in the sense that is relevant for OCP-Place evaluation. For previous discussion of the role of underspecification in co-occurrence restrictions, see especially Yip (1989).

Note that it would be possible to amend (6) so that *disagreement* on a feature F actively lowers similarity values (as is possible in the **contrast model** of Tversky 1977; see also Pierrehumbert 1993). Simply redefine δ_F so that it returns -1 rather than 0 given two F-disagreeing segments. This would sometimes result in negative similarities, which could be eliminated by redefining $sim_{\mathbf{w}}(x, y)$ as $\max(0, \sum_{F \in \mathcal{F}} w_F \cdot \delta_F(x, y))$. We leave exploration of these and other variants of the theory for future research.

[13]An alternative implementation of our main idea would use a set of separately weighted OCP-Place[F] constraints, one for each feature F $\in \mathcal{F}$, where OCP-Place[F] is violated by sequences of homorganic segments that agree on feature [F]. We do not have an argument against this alternative, but adopt (6) and (7) in the hope that future research will discover evidence for the language-specific feature weights that is independent of OCP-Place effects (and therefore not attributable to OCP-Place[F] constraints).

[14]It should be noted, however, that Elmedlaoui (1995) and Bachra (2000) argue for non-parasitic dissimilatory effects of subsidiary features in the root co-occurrence pattern of Arabic. Furthermore, Zuraw and Lu (2009) employ constraints similar to OCP-Sonorant and OCP-Nasal in their analysis of non-static phonology (alternations and affix placement). Therefore, it seems likely that such constraints are needed independently of (7), in which case the question arises of whether a single set of feature weights suffices to account for both place-parasitic and non-place-parasitic OCP effects in a given language.

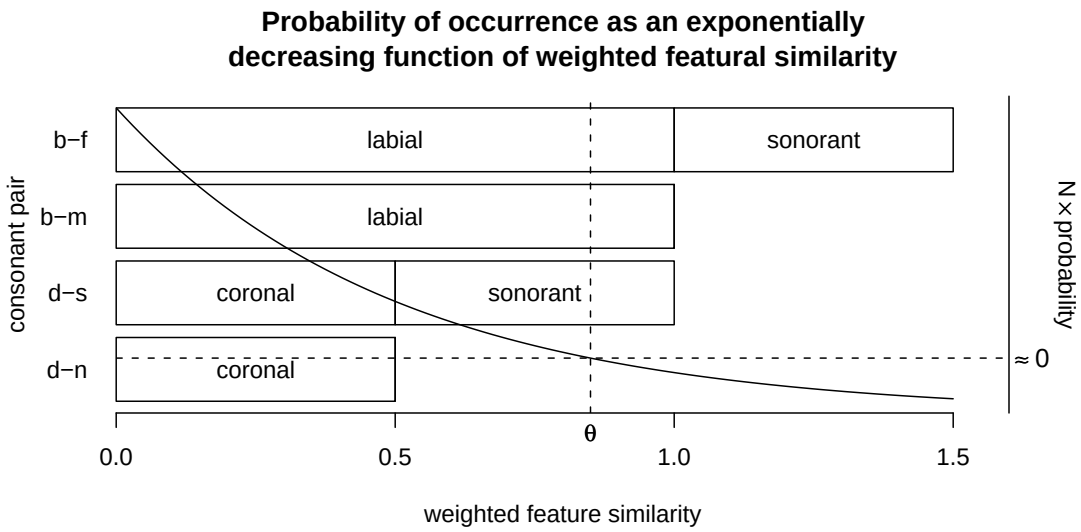
* * *

Having set out our proposal, we now compare its properties with those of the theories reviewed in §1. In one respect, our proposal is richer than the previous, non-quantitative theories of OCP-Place that parameterize the relevance of a subsidiary feature in a binary fashion. Parametric *irrelevance* corresponds to a weight of zero in our theory. But in place of parametric *relevance* we have a (potentially) infinite range of feature weights. The empirical motivation for this descriptive flexibility is the same as that given by CP for their weight-based theory (see CP, pp. 293-306; see also the case studies in §4).

In another respect, our proposal is less flexible than the non-quantitative theories and the quantitative theory of CP, because we do not allow the relevance of a subsidiary feature to differ across places of articulation in the same language. The same feature weights **w**, and hence the same similarity function $sim_{\mathbf{w}}$, determine the degree of OCP-Place violation in (7) regardless of which place of articulation is shared by homorganic segments. *The individual place features can have different weights* — a quantitative manifestation of the more general idea that the place features are independent entities, each on its own 'tier' (e.g., McCarthy 1988) — *but the subsidiary feature weights cannot vary across places*. The concomitant increase in explanatory potential is great, especially relative to that of CP's factorial constraint set (5), since the number of independent stipulations is limited to the sum of the number of place features and the number of subsidiary features, $|\mathcal{P}| + |\mathcal{S}|$ ($= 8$ given the feature set adopted here), instead of their product, $|\mathcal{P}| \times |\mathcal{S}|$ ($= 16$). The difference in the number of independent weights grows by a factor of $|\mathcal{P}|$ for every subsidiary feature added to the theory. (Recall the suggestions of Lightner 1973; Pierrehumbert 1993 that *all* features can act as subsidiary for OCP-Place evaluation).

Under the present hypothesis, apparent place-specific differences in subsidiary behavior must be due to the weights of the place features themselves, not to true place $\times$ subsidiary interactions. For example, suppose that the weight of Labial place is much greater than that of Coronal place in a given language, $w_{\text{Labial}} \gg w_{\text{Coronal}}$. Suppose further that w_{Labial} is so large that the similarity of *any* two labial consonants, and hence their degree of OCP-Place violation, prohibits them from co-occurring. It will thus appear that all subsidiary features are irrelevant for OCP-Place effects on labials; that is, the weights of the subsidiary features will be effectively 'masked' by the weight of the place feature itself. In contrast, those weights will be 'unmasked' among the coronals, and the subsidiary features will thus appear relevant for this place, if w_{Coronal} is sufficiently small. This state of affairs, which we will argue to be part of the proper characterization of the difference between Labial and Coronal OCP-Place effects in Arabic (see §4.1), is schematized in the figure below.

Graff and Jaeger (2009) propose a theory of consonant co-occurrence restrictions that is in some ways similar to ours but which seems to dispense with gating by place of articulation altogether. This strikes us as incorrect or at least insufficiently argued, given the large body of evidence in favor of a gradiently violated OCP-*Place* constraint. Preliminary attempts to compare our theory with that of Graff and Jaeger (2009) favoring the former, but further research along these lines is necessary before making a definitive judgement. (Thanks to Kie Zuraw, p.c., for bringing Graff and Jaeger's work to our attention.)



The horizontal axis of the figure indicates the similarity (in arbitrary units) of the consonant pairs displayed on the vertical axis. The contributions that the place features and [sonorant] make to similarity are indicated by horizontally stacked bars. For purposes of concrete illustration, $w_{\text{Labial}} = 1.0$ and $w_{\text{Coronal}} = w_{\text{sonorant}} = .5$. By hypothesis, [sonorant] has a constant weight across the two places. However, the quantitative consequences of agreeing on [sonorant] is different depending on whether there is also agreement on Labial or Coronal. The reasoning goes as follows.

Assume, as indicated by the curved solid line, that the probability of co-occurrence of segments x and y (within a root or some other domain) is an exponentially decreasing function of their weighted feature similarity, $sim_{\mathbf{w}}(x, y)$. This assumption will hold, modulo differences in the baseline probabilities of the segments themselves, in the stochastic model we develop later in the paper (§3.2.3, §4). At some critical or threshold similarity value, indicated in the figure by θ, the probability of co-occurrence will become so small that not even one instance of the pair xy would be expected in a root lexicon of total size N. For concreteness, we have set θ equal to $.75$.

Because two consonants that agree on Labial cross the threshold *regardless of whether or not they also agree on [sonorant]*, we would not expect to find a single instance of either [b-m] or [b-f] in the lexicon. However, because agreement on Coronal does not suffice to cross the threshold, some number of lexical items are expected to contain [d-n]. Only Coronal-Coronal pairs that also agree on [sonorant], such as [d-s], are so similar that they lie on the same side of the threshold as all Labial-Labial pairs. It would thus appear from the expected lexical distribution that [sonorant] behaves as a subsidiary feature for Coronal only, when in fact it and all other subsidiary features make the same contribution to similarity regardless of which place is shared.

We claim that the same type of relationship among feature weights, similarity, and expected probability explains away all cases of apparently place-specific subsidiary effects.[15]

[15] A large body of experimental work has established that native speakers can make well-formedness or acceptability distinctions among unattested structures (e.g., Greenberg and Jenkins 1964; Scholes 1966; Ohala and Ohala 1986; Berent and Shimron 1997; Berent et al. 2007; Bailey and Hahn 2001; Hay et al. 2003; Myers and Tsay 2005; Davidson 2006; Kirby and Yu 2007; Coetzee 2008). In light of this general finding, the most basic prediction of our theory is that (all other things being equal) [b-f] would be judged to be at least as unacceptable as [b-m] given the feature weights discussed in the text. Whether native speakers would show a dispreference for [b-f] relative to [b-m] would depend on how accurately acceptability judgments (or other experimental measures) track grammatical differences. (See CP, pp. 306-307, 309, 317, for related discussion.)

3 Evaluating and comparing quantitative models

Given competing models of OCP-Place such as those presented in §1 and §2, and a lexicon or set of consonant co-occurrence frequencies, how should we evaluate and compare the models with respect to the data? FPB and CP follow the method proposed by Pierrehumbert (1993), in which the quantitative predictions of a model are assessed by their *correlation* with 'Observed over Expected' (O/E) values computed from co-occurrence frequencies (see also Frisch 1996, 2000; Kawahara et al. 2006; Aldrete and Frisch 2007; Ito 2007; Anttila 2008; Coon and Gallagher 2008). We do not adopt this method, and moreover argue that it should never be used for this purpose. In this section, we explain why we take this position and present our alternative, which is based on probability theory. We also introduce the probabilistic MaxEnt grammar formalism that is used in the model fitting and comparison of §4.

3.1 Against O/E

The O/E method of Pierrehumbert (1993) works as follows (see also Pierrehumbert 1993, ms. p. 2; FPB, pp. 185-187; CP, p. 290). Let Σ be the set of segments under consideration; in the present context, Σ is the consonant inventory of a given language. For each consonant pair $xy \in \Sigma \times \Sigma$, the O/E value of xy is gotten by dividing the frequency of the pair (its 'O' value), O_{xy}, by the following product: the total frequency of all consonant pairs, $N = \sum_{x'} \sum_{y'} O_{x'y'}$, multiplied by the proportion of pairs that have x as their first member, $(\sum_{y'} O_{xy'})/N = O_{x+}/N$, and the proportion of pairs that have y as their second member, $(\sum_{x'} O_{x'y})/N = O_{+y}/N$. The following equation summarizes this calculation:

(8) $$O_{xy}/E_{xy} = O_{xy}/(N \cdot \frac{O_{x+}}{N} \cdot \frac{O_{+y}}{N})$$

The significance of the denominator $E_{xy} = N(O_{x+}/N)(O_{+y}/N)$ is that it gives an estimate of the number of times the pair xy would occur by random combination of consonants, in a sample of size N, *if there were no phonological restrictions on* ***any*** *of the consonant pairs* (i.e., if the choice of the first and second member of each pair were independent). It is this sense, and this sense only, in which E_{xy} is an 'expected' value of the frequency O_{xy}. In general, E_{xy} is *not* the expected value of O_{xy} if one or more of the consonant combinations is subject to grammatical restriction or constraint. Of course, the existence of at least one constraint on combinations is presupposed by Pierrehumbert (1993), FPB, CP, and all other work on the proper formulation of OCP-Place.[16]

Pierrehumbert (1993)'s method assesses the fit of a model to the data by evaluating the correlation between the empirical O/E values of consonant pairs and model-defined quantities such as natural classes similarity values. The general assumption is that smaller O/E values correspond to stronger restrictions or greater degrees of OCP-Place violation. Pierrehumbert (1993) and FPB report results from non-linear regressions of O/E values against (transformed) similarity values (FPB, pp. 207-209), whereas CP evaluate models with linear regression (CP, pp. 327-328). Regardless of the specific type of regression performed, this method is valid only if the O/E value of a consonant pair provides a statistically sound estimate of the degree to which the pair violates OCP-Place (or, in CP's application, of the pair's Acceptability value, which is a linear function of its OCP-Place Harmony). Neither Pierrehumbert (1993) nor subsequent research has attempted to prove that this necessary statistical relationship holds. The preceding discussion of E_{xy} leads

[16] In statistics, the complete absence of restrictions on combination is a common null hypothesis, with the presence of one or more restrictions being the alternative that is tested against the null. For definition of the null hypothesis, and its connection to E_{xy} values, see Agresti (2002, pp. 22, 25, 36-39, 78) and Wickens (1989, pp. 17-27).

us to suspect that the relationship does not systematically hold when OCP-Place or other constraints on consonant combination are in force. The following simple example confirms this suspicion.

* * *

Consider an abstract language in which the set of consonants is $\Sigma = \{P, T, K\}$ and the probability of each possible combination $xy \in \Sigma \times \Sigma$ is determined, up to a normalizing constant, by the product of three terms: $\psi_{1(x)}$ — a penalty for having x in the first position of the pair (i.e., in position '1'), $\psi_{2(y)}$ — a penalty for having y in the second position (i.e., in position '2'), and $\psi_{12(xy)}$ — a penalty for combining x and y in that order (i.e., the joint effect of having x in position 1 and y in position 2). The $\psi_{1(x)}$ and $\psi_{2(y)}$ terms are proxies for position-specific constraints on individual segments (e.g., the 'positional preferences' of Bachra 2000, pp. 62ff. or the positional Markedness constraints of Zoll 2004). The $\psi_{12(xy)}$ terms instantiate co-occurrence restrictions such as the strict OCP or OCP-Place. Hypothetical values for all of these terms are given in table (a) below, with $\psi_{1(x)}$ values along the rows, $\psi_{2(y)}$ values along the columns, and $\psi_{12(xy)}$ values in the individual cells.

	$P\ (\frac{1}{3})$	$T\ (\frac{1}{2})$	$K\ (\frac{1}{6})$
$P\ (\frac{1}{3})$	$\mathbf{\frac{1}{2}}$	1	1
$T\ (\frac{1}{2})$	1	$\mathbf{\frac{1}{2}}$	1
$K\ (\frac{1}{6})$	1	1	$\mathbf{\frac{1}{2}}$

(a) Terms for margins and cells

	P	T	K
P	$\frac{1}{3\cdot3\cdot2}$	$\frac{1}{3\cdot2\cdot1}$	$\frac{1}{3\cdot6\cdot1}$
T	$\frac{1}{2\cdot3\cdot1}$	$\frac{1}{2\cdot2\cdot2}$	$\frac{1}{2\cdot6\cdot1}$
K	$\frac{1}{6\cdot3\cdot1}$	$\frac{1}{6\cdot2\cdot1}$	$\frac{1}{6\cdot6\cdot2}$

(b) Products of terms

	P	T	K
P	0.069	0.207	0.069
T	0.207	0.155	0.103
K	0.069	0.103	0.017

(c) Normalized products

	P	T	K
P	345	1034	345
T	1034	776	517
K	345	517	86

(d) Frequencies in a sample

	P	T	K
P	**0.58**	1.29	1.06
T	1.29	**0.72**	1.17
K	1.06	1.17	**0.48**

(e) O/E values for the sample

Note that the terms $\{\psi_{1(x)}\}$ and $\{\psi_{2(y)}\}$ define probability distributions over the rows and columns of table (a), respectively. For simplicity, the distribution is the same along both dimensions: $\Pr(P) = 1/3$, $\Pr(T) = 1/2$, and $\Pr(K) = 1/6$. Note also that the combination term $\psi_{12(xy)}$ is equal to 1 — which corresponds to the absence of a constraint on the combination xy — except along the diagonal, where it imposes a *uniform* penalty of $1/2$ on all identical combinations. Thus the collection of $\psi_{12(xy)}$ terms acts as a soft OCP constraint: a constraint that is violated only by the three identical combinations PP, TT, and KK. For what follows it is crucial to keep in mind that the degree of violation is the same for all three identical combinations.

Multiplication of each cell-specific term by the corresponding row and column terms yields table (b). For example, the value in the KK $\psi_{1(K)} \cdot \psi_{2(K)} \cdot \psi_{12(KK)} = \frac{1}{6\cdot6\cdot2}$. When all such products are normalized (divided by their sum), the result is the probability distribution over consonant pairs shown in table (c). (In the interest of legibility the probabilities have been rounded to three decimal places and therefore do not sum to exactly 1.) In a random sample of consonant pairs of total size N, each combination would be expected to have a frequency equal to about N times its probability. Table (d) shows those frequencies for $N = 5000$. (The frequencies have been rounded to integers, with the result that the total sample size is slightly different from 5000.) Finally, the O/E values corresponding to the frequencies in table (d) are given in table (e).

Despite the fact that the co-occurrence frequencies in table (d) were derived from a grammar in which the OCP applies with the same strength to all identical combinations (recall table (a)), the O/E values on the diagonal of table (e) are not equal. In particular, the O/E values make it appear that the OCP constraint applies more strongly to the combination KK than to PP and TT and more strongly to PP than to TT (because $O_{KK}/E_{KK} < O_{PP}/E_{PP} < O_{TT}/E_{TT}$). This cannot be due to true differences in the strength of OCP violation, since there are none. Instead, it must reflect differences in positional probability (the $\psi_{1(x)}$ and $\psi_{2(y)}$ terms), which are independent of restrictions on combination.

Consider now how two hypothetical theories would fare on this data if Pierrehumbert's method were applied. According to one theory, which is true by hypothesis, the OCP is violated to the same degree by all identical consonant pairs. According to the other theory, which is false, the OCP applies more strongly to KK than to PP (i.e., more strongly to 'Dorsal-Dorsal' pairs than to 'Labial-Labial' pairs), and more strongly to PP than to TT (i.e., more strongly to 'Labial-Labial' pairs than to 'Coronal-Coronal' pairs). Correlations between the O/E values and the predictions of these theories would support the false conclusion that the strength of the OCP constraint varies across identical pairs in the way that the second theory prescribes.

What this hypothetical case illustrates is that O/E values can provide a distorted index of the strength with which constraints like OCP or OCP-Place apply to combinations, making it appear that there are differences in strength when in fact there are none or exaggerating differences that do exist. We have no reason to believe that this problem is limited to simple artificial examples such as the one just described. Note also that the largest difference in O/E values here, .72 - $0.48 = .24$, is well within the range of values that are taken to be significant in FPB's analysis of Arabic; see their table IV, p. 203.[17]

* * *

It might be objected at this point, by proponents or practitioners of the O/E method, that multiplicative combination of terms as in table (b) above does not correspond to the grammatical theory for which the method is intended or appropriate. Indeed, the Harmonic Grammar model of OCP-Place effects proposed by CP involves summing weighted constraint violations (CP, p. 308), maximizing over the set of Harmony values (CP, p. 308), and subtracting Harmony values to derive Acceptability values (CP, p. 313). Multiplication is only used in the weighting operation (CP, p. 308). The Optimality Theory model of Anttila (2008) involves no multiplication whatsoever.

---

[17]This is not the place for an in-depth mathematical analysis of O/E (see [author reference, in prep.]), but a brief formal discussion may provide an indication of the generality of the result in the text. By construction, each cell in table (d) has a value that can be written (ignoring the effects of rounding) as $N\psi_0\psi_{1(x)}\psi_{2(y)}\psi_{12(xy)}$, where N is the sample size and ψ_0 is a multiplicative constant that is common to all cells. This value is O_{xy} for the cell xy in the ideal case that cell frequencies are identical to cell probabilities multiplied by the sample size. In order for O_{xy}/E_{xy} to provide a relative measure of $\psi_{12(xy)}$ — that is, to measure the strength of the restriction on the combination xy relative to all other combinations — it must be the case that E_{xy} is equal to $\alpha\psi_{1(x)}\psi_{2(y)}$, where α is a constant that is the same for all cells. (Note that dividing $N\psi_0\psi_{1(x)}\psi_{2(y)}\psi_{12(xy)}$ by $\alpha\psi_{1(x)}\psi_{2(y)}$ yields $[(N/\alpha)/\psi_0]\psi_{12(xy)}$, in which all of the terms except $\psi_{12(xy)}$ are constant across all cells.)

It is easily shown, however, that there is no such α for the hypothetical case considered in the text. The values of $\psi_1(x)\psi_2(y)$ along the diagonal in table (a) are $(\frac{1}{3\cdot 3}, \frac{1}{2\cdot 2}, \frac{1}{6\cdot 6}) \approx (0.111, 0.25, 0.028)$, and the corresponding E_{xy} values along the diagonal for table (d) are $N \cdot (0.119, 0.217, 0.035)$. Pointwise division of the latter by the former yields $(0.935, 1.154, 0.773)$. In order to transform this vector into one that is constant (i.e., has the same value in all three positions), α would have to be both greater than and less than 1 — an impossibility.

The fundamental error in the O/E method is the assumption that the **marginal effects**, represented by our $\psi_{1(x)}$ and $\psi_{2(y)}$ terms, can be estimated by the corresponding components of E_{xy}, namely the **marginal proportions** O_{x+}/N and O_{+y}/N, even when the $\psi_{12(xy)}$ terms are not all equal to 1. If this assumption were correct, then dividing O_{xy} by E_{xy} would of course always yield the combination term $\psi_{12(xy)}$ (up to a normalizing constant). But the assumption is false: terms that apply to combinations will make the empirical marginal *proportions* differ from the underlying marginal *effects*, and hence taking O/E ratios will not in general recover the combination effects. This is unsurprising, since the frequencies O_{x+} and O_{+y} reflect all of the terms, not just the marginal ones.

However, it is precisely these aspects of the theories that make the adoption of the O/E method by CP, Anttila (2008), and others so incongruous. The O/E statistic is, at the most basic level, a *ratio* of two terms: one term that is specific to a combination itself (O_{xy}), and one that is determined by the two elements of the combination (E_{xy}). If the theory of grammar does not specify that (weighted) constraint violations are combined by multiplication, why should the method of isolating the degree to which one type of constraint is violated — or derivative quantities such as Acceptability values — involve division? Why is the appropriate statistic not O *minus* E, as might seem more appropriate for CP's additive theory of constraint combination, or some other function of the co-occurrence frequencies? Such questions have not even been raised, let alone answered, in previous work.

In conclusion, we find there to be no justification, in any theoretical framework of which we are aware, for estimating degrees of constraint violation or related quantities with O/E values, the practice of Pierrehumbert (1993) and many others notwithstanding. If constraints interact multiplicatively, then the O/E ratio does not generally deconfound constraints on combinations (related to the $\psi_{12(xy)}$ terms in our notation) from position-specific constraints on individual elements (related to the $\psi_{1(x)}$ and $\psi_{2(y)}$ terms). If constraints interact in some non-multiplicative way, there is no apparent motivation for working with a ratio statistic such as O/E at all. We would reject as incomplete any response to the preceding discussion that does not prove that O/E in fact measures the grammatical quantities that it is supposed to measure. Such a demonstration would require a commitment to a theory of constraint interaction and a formal link between constraint evaluation and co-occurrence frequencies, topics that we turn to next.

3.2 A probability-based alternative

The O/E method can be understood as a (flawed) attempt to pre-theoretically distinguish the effects of constraints on combinations from those of constraints on individual segments, and thereby to avoid postulating an explicit theory of how such constraints collectively account for gradient data. FPB characterize constraint interaction, particularly insofar as it concerns gradient constraints like OCP-Place, as an open problem (pp. 221-222). However, general mathematical models of constraint interaction that encompass both gradient and, in the limit, categorical constraints were developed throughout the two decades prior to the publication of FPB (e.g., Hopfield 1984; Ackely et al. 1985; Hinton and Sejnowski 1986; Smolensky 1986; Legendre et al. 1990a,b; Berger et al. 1996; Rosenfeld 1996; Della Pietra et al. 1997). As discussed by Hayes and Wilson (2008) (see also Albright 2009, ms. pp. 7-9; p. 24, note 6), two properties of a model of constraint interaction are of particular importance for phonological analysis: (i) the ability to evaluate the **probability of the data** —- equivalently, the **likelihood of the model** — given values of the free parameters (e.g., feature or constraint weights) and (ii) the ability to form coherent grammars over phonological representations in which each element has multiple descriptions (e.g., multiple features and multiple structural roles).

We know of only a few types of model that combine these properties: those based on stochastic (or 'noisy') versions of OT (e.g., Boersma 1997; Zuraw 2000; Boersma and Hayes 2001; Jarosz 2006b) and HG (e.g., Boersma and Pater 2007b; Jesney 2007), and models based on the principle of **Maximum Entropy** ('MaxEnt'; e.g., Berger et al. 1996; Rosenfeld 1996; Della Pietra et al. 1997; Jelinek 1999; Goldwater and Johnson 2003; Klein and Manning 2003; Clark and Curran 2007; Hayes and Wilson 2008; see also Ackely et al. 1985; Hinton and Sejnowski 1986; Smolensky 1996 for closely related connectionist formalisms). Among these, MaxEnt models have the further advantages that the probability of the data is easily evaluated (up to a normalizing constant; Hayes and Wilson 2008, pp. 384-386), and that parameter learning can be performed by descending a gradient, or vector of partial derivatives, that is straightforward to calculate or approximate (Hayes and Wilson 2008, pp. 385-389). These properties are not presently available for stochastic OT models, in which calculating the probability of even a single output requires either multidimensional integration for which there is no analytic solution (e.g., Zuraw 2000) and/or summation over all

possible inputs (e.g., Jarosz 2006b), and for which there is no published learning algorithm that has been proved correct and does not involve explicitly representing all possible constraint rankings (see Pater 2008b in particular on the learning algorithm of Boersma 1997). The same considerations favor MaxEnt over stochastic versions of HG.[18]

For these reasons, and in light of the results obtained previously by Hayes and Wilson (2008), we adopt the MaxEnt grammar formalism, and fit and compare many different formulations of OCP-Place within that framework (see §4). The remainder of this section is organized as follows. We first introduce the notions of maximum likelihood and maximum a posteriori fitting (§3.2.1). We then factor Occam's razor into the method of model evaluation, returning to themes raised in the introduction to the paper (§3.2.2). Finally, we review the MaxEnt formalism itself (§3.2.3)

3.2.1 Maximum likelihood and maximum a posteriori estimation

The general problem we address here is how to assess the 'fit' of a model to a given body of gradient data. There are numerous methods, ranging from various types of correlation to measures based on integrating over the members of a large class of hypotheses. Many of the methods presuppose the ability to calculate the probability of the data given the model. If the model has one of more free parameters, as all models considered in this paper do, it is conventional to assess the fit of the model in terms of the *maximum* data probability that can be achieved by adjusting the parameters. This is the ubiquitous method of **maximum likelihood** (e.g., Duda et al. 2001; MacKay 2003; Bishop 2006; for a tutorial review, see Myung 2003).

Formally, let $\mathbf{G}$ be a model (e.g., a set of constraints and a grammar formalism regulating their interaction) and θ be the associated vector of free parameters (e.g., the weights of the constraints). Further let D be the data to which the parameters of the model are to be fit (or, if one prefers, from which the parameter values are to be learned). In the analysis of gradient consonant co-occurrence restrictions, D standardly takes the form of a list of consonant pairs *tokens*. This is equivalent to a set of consonant pair *types* together with their token frequencies (i.e., the number of times each pair occurs in the lexicon or set of roots). Following the notation of §3.1, we use xy to denote a generic consonant pair and O_{xy} to denote the token frequency of xy in the data.[19]

If $\mathbf{G}$ is a proper stochastic model, then it is possible to calculate the probability of D for any value of θ. We write the data probability as $\Pr_{\mathbf{G},\theta}(D)$. Under the common, simplifying assumption that the elements of D are drawn independently from the distribution $\Pr_{\mathbf{G},\theta}(\cdot)$, the probability of D factors into the product of the probabilities of the forms:

(9) $$\Pr_{\mathbf{G},\theta}(D) = \prod_{xy \in D} \Pr_{\mathbf{G},\theta}(xy)^{O_{xy}}$$

[18]Jesney (2007) claims that there is a learner for stochastic HG, but there is no proof that such an algorithm exists (see 3.2.1 for related discussion). The main point of Jesney (2007) is that stochastic HG is more restrictive than MaxEnt because (i) only the former prevents harmonically-bounded forms from being produced and (ii) the latter 'predicts that [variable] processes will always pattern independently' (p. 4). The first point reflects a confusion about MaxEnt grammars. Such grammars define a probability distribution over a set Ω of forms (or candidates), but do not determine the contents of Ω. If there is evidence that probability should be distributed among only the forms that are not harmonically bounded (i.e., the *contenders*, Riggle 2004, 2009), this limitation could be imposed on Ω. Jesney provides no more formal statement of the second point, and offers no empirical example in which the claimed lesser restrictiveness of MaxEnt relative to HG is problematic.

[19]The method described here can be generalized to any empirical domain by taking D to be the relevant set of phonological types together with their token frequencies. For example, D could represent the observed frequencies of a set of (input,output) pairs, as is common in applications of stochastic OT (e.g., Boersma and Hayes 2001).

Note that the contribution to (9) of any unseen (zero-frequency) pair xy is identically 1, because $\Pr_{\mathbf{G},\theta}(xy)^0 \doteq 1$. This is computationally useful, because the product can be restricted to only those pairs that actually occur in the data (as opposed to all possible pairs). It also follows that maximum likelihood learning requires only positive evidence from the environment, making it a plausible high-level characterization of human language acquisition. (This applies also to MAP learning, a variant of maximum likelihood learning that is introduced below.)

Because the right-hand side of (9) is a product of probabilities, each of which is necessarily less than or equal to 1, $\Pr_{\mathbf{G},\theta}(D)$ is typically a very small value and numerical difficulties can arise in calculating it exactly. Therefore, it is standard to instead work with the log of the data probability, $\log \Pr_{\mathbf{G},\theta}(D)$. This is by definition equal to $\ell(\theta; \mathbf{G}, D)$, the **log-likelihood** of the parameters:

$$\ell(\theta; \mathbf{G}, D) \doteq \log \Pr_{\mathbf{G},\theta}(D) = \log \prod_{xy \in D} \Pr_{\mathbf{G},\theta}(xy)^{O_{xy}} = \sum_{xy \in D} O_{xy} \log \Pr_{\mathbf{G},\theta}(xy) \tag{10}$$

Maximizing $\ell(\theta; \mathbf{G}, D)$ per se is typically an ill-posed problem (e.g., Szeliski 1986): the parameter values that maximize the probability of the data may not exist, and when they do exist they are often not unique. Furthermore, aspects of a language-specific data pattern can pull the parameters to extreme values, thereby exaggerating the extent of variation across languages. In the present application, this undesirable artifact of maximum likelihood estimation would arise, for example, if homorganic consonants that agree on a particular place or subsidiary feature never co-occur in a particular language.

A general response to the problem of ill-posed likelihood functions is provided by the Bayesian approach to parameter estimation. This involves imposing a data-independent, or universal, prior probability distribution $\Pr_{\mathbf{G}}(\theta)$ on the parameters (e.g., Lehmann and Casella 1998; MacKay 2003; Gelman et al. 2004; Bishop 2006). The function to be maximized is thus not the log-likelihood (data probability) itself, but rather the 'regularized' log-likelihood:

$$\ell'(\theta; \mathbf{G}, D) \doteq \log[\Pr_{\mathbf{G},\theta}(D) \cdot \Pr_{\mathbf{G}}(\theta)] = \log \Pr_{\mathbf{G},\theta}(D) + \log \Pr_{\mathbf{G}}(\theta) \tag{11}$$

We assume that $\Pr_{\mathbf{G},\theta}(D)$ is as in (9). For $\Pr_{\mathbf{G}}(\theta)$, we adopt a commonly-used prior according to which each parameter θ_i is drawn independently from a Gaussian (normal) distribution with mean μ_i and variance σ_i^2 (Chen and Rosenfeld 1999; Johnson et al. 1999; see also Rumelhart et al. 1996; Goldwater and Johnson 2003; Hayes and Wilson 2008, and Rumelhart et al. 1996; Chen and Rosenfeld 2000; Goodman 2004 for alternatives). In this paper, we assume a relatively broad prior distribution for the parameters, setting $\mu_i = 0$ and $\sigma_i^2 = 10$ for all θ_i. Our intent was to impose sharp prior penalties for parameter values greater than about 20. Within the theory of constraint interaction we adopt (§3.2.3), this allows parameter fitting to be sensitive to differences among languages but prevents cross-linguistic variation from being exaggerated with unduly extreme weights.

Because the $\log$ function is monotonic, the value of θ that maximizes (11) is the same as the value that maximizes $\Pr_{\mathbf{G},\theta}(D) \cdot \Pr_{\mathbf{G}}(\theta)$. This **maximum a posteriori** (MAP) value of the parameter, denoted by θ_{MAP}, is important for two reasons. First, finding θ_{MAP} is a reasonable goal for the language learner, one that embodies and generalizes notions of 'frequency matching' (e.g., Boersma and Hayes 2001, p. 53ff.) and of finding the 'most restrictive' grammar compatible with language-specific evidence (e.g., Hayes 2004; Prince and Tesar 2004; Jarosz 2006b). Furthermore, as already noted, this type of learning can be performed from positive environmental evidence only. From this perspective, $\ell'(\theta; \mathbf{G}, D)$ is a plausible objective function

for language acquisition.[20]

Second, and most importantly for this paper, the value of the function $\ell'(\theta; \mathbf{G}, D)$ evaluated at the point θ_{MAP} provides a measure of how well the model can possibly account for the data. In essence, this is our proposed solution to the problem of assessing models based on gradient data. The solution does not require computation of theoretically-unmotivated ancillary statistics, such as O/E values, and it imposes no restrictions on the models other than those required by probability theory. Instead of attempting to isolate the effect of one type of constraint, such as OCP-Place, the solution quantifies how well *all* of the constraints in a model, interacting together, explain the data. Different models can then be compared on an equal probability-based footing, each assessed with its own MAP parameters. Immediately below, we augment this model-comparison method by incorporating a term that penalizes more complex (less restrictive) models. But even in its augmented form the method will make crucial use of $\ell'(\theta; \mathbf{G}, D)$.

3.2.2 Incorporating Occam's Razor

All other things being equal, a model with more free parameters will achieve a value of $\ell(\theta; \mathbf{G}, D)$ or $\ell'(\theta; \mathbf{G}, D)$ that is at least as high as — and typically higher than — that of a model with fewer free parameters. Even when all other things are not equal (i.e., when neither model is 'nested' within the other), the model with additional parameters will typically attain a higher value on more possible data sets. It will therefore be *less restrictive* than the model with fewer parameters: its ability to 'fit' many possible patterns entails a failure to predict that only some of those patterns will be found. In addition to the number of parameters, the functional form of a model also affects its restrictiveness (e.g., Myung and Pitt 1997). For example, the generalized linear regression model (McCullagh and Nelder 1989), which includes logistic regression functions of the type considered by FPB (pp. 182-183, 209) and Frisch (1996, 2000), is less restrictive than ordinary linear regression. Similarly, a connectionist network with a layer of hidden units is less restrictive than a network with only input and output units (e.g., Minsky and Papert 1969; Rumelhart et al. 1986; Hornik et al. 1989).

Given two or more models that differ in number of parameters or functional form, it may sometimes be appropriate to compare them purely on the basis of likelihood (10), MAP value (11), or other measures of fit to the data such as correlation. Strictly exploratory or descriptive investigation, particularly in new empirical domains, could profitably proceed in this way. But in areas that have been the subject of intense theoretical investigation for several decades — as OCP-Place effects have been — and when explanation is part of the research goal — so that finding the right balance of descriptive flexibility and restrictiveness

[20]Note that maximizing $\ell'(\theta; \mathbf{G}, D)$ is a learning **objective**, not a learning **algorithm** (see also Tesar and Smolensky 1998, 2000; Goldwater and Johnson 2003; Prince and Tesar 2004; Jarosz 2006b,a on the objective/algorithm distinction). Because the objective for the MaxEnt models that we adopt is convex (e.g., Della Pietra et al., 1997), there are many algorithms that could find its maximum value (e.g., (improved) iterative scaling, steepest ascent, conjugate gradient ascent, simulated annealing; see Malouf 2002 for a review and comparison). Each of these algorithms could be modified to apply in 'batch' or 'on-line' modes. The choice of algorithm and mode would be relevant for modeling the detailed pattern of language acquisition, but not necessarily for understanding the adult state that is ultimately obtained. Goldwater and Johnson (2003) make similar points in relation to the Gradual Learning Algorithm (GLA) of Boersma (1997).

We emphasize the distinction between objectives and algorithms because the discussion in CP is unclear on this point. For example, CP reject the regression analysis of their §2.5 — which has an objective function that could be optimized in many ways — partly on the grounds that they do not regard it "to be a particularly realistic model of human language acquisition" (p. 318). The learning algorithm that they endorse does not appear to have a (known) objective that relates to predicted probability or Acceptability. Of the two references that they give for their algorithm, Fischer (2005) and Boersma and Pater (2008), the first develops an on-line learning algorithm for the MaxEnt framework that we adopt — not the Harmonic Grammar model proposed by CP — and the second presents only a proof about learning input-output mappings, not about learning relative degrees of probability or well-formedness from gradient data.

is important — comparisons based only on data fits are inadequate and potentially misleading (e.g., Pitt and Myung 2002). What is needed instead is a method comparison that penalizes less restrictive models and thereby offsets their almost inevitably higher log-likelihoods or correlation values. In short, model evaluation and comparison must incorporate Occam's razor.

The particular formalization of Occam's razor that we adopt is known as the **Laplace approximation** (or **saddle-point approximation**; e.g., de Bruijn 1958; Tierney and Kadane 1986; Kass and Raftery 1995; MacKay 2003, pp. 341, 350, 501-503; Bishop 2006, pp. 213-217). This is an approximation to the typically high-dimensional integral $\int_\theta \ell'(\theta; \mathbf{G}, D) d\theta$, which is in turn proportional to the Bayesian **posterior probability** of model $\mathbf{G}$ under the unbiased assumption that all models have equal prior probability. It is derived by Taylor-expanding $\ell'(\theta; \mathbf{G}, D)$ around the maximizing point or 'mode' θ_{MAP} and noticing that the exponential of the result has the form of an unnormalized Gaussian distribution centered at the mode (e.g., Bishop 2006, pp. 213-217). The approximation to the posterior probability of $\mathbf{G}$ is then $\ell'(\theta_{\text{MAP}}; \mathbf{G}, D)$ multiplied by the normalizing constant of that Gaussian distribution. Omitting the technical derivation (for which see the references cited), this works out to be:

$$\begin{aligned} \log \Pr_{\mathbf{G}}(D) &\approx \log[\, \Pr_{\mathbf{G},\theta_{\text{MAP}}}(D) \cdot \Pr_{\mathbf{G}}(\theta_{\text{MAP}}) \cdot (2\pi)^{M/2}/|A|^{1/2}\,] \\ &= \log \Pr_{\mathbf{G},\theta_{\text{MAP}}}(D) + \log \Pr_{\mathbf{G}}(\theta_{\text{MAP}}) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{A}| \end{aligned} \tag{12}$$

where M is the number of parameters in the model (i.e., the length of the vector θ), $\mathbf{A} = -\nabla\nabla \log \Pr(\theta|D, \mathbf{G})|_{\theta_{\text{MAP}}}$ is the **Hessian** (matrix of second partial derivatives) evaluated at the point θ_{MAP}, and $|\mathbf{A}|$ is the **determinant** of $\mathbf{A}$ (sometimes denoted by $\det \mathbf{A}$). (Note that the sum $\log \Pr_{\mathbf{G},\theta_{\text{MAP}}}(D) + \log \Pr_{\mathbf{G}}(\theta_{\text{MAP}})$ is the regularized log-likelihood, as in (11), evaluated at θ_{MAP}.)

The first term on the right-hand side of (12) rewards the model $\mathbf{G}$ in proportion to how well it accounts for the data with its own MAP parameters θ_{MAP}. The second term, which is the other component of $\ell'(\theta; \mathbf{G}, D)$, penalizes $\mathbf{G}$ in proportion to how far the MAP parameters deviate from their prior values. This term does not uniformly favor less complex models: indeed, a model with a smaller *number* of parameters may have to resort to more extreme parameter *values* than a model with a richer parameter set. Rather, model complexity is penalized by the term $-\frac{1}{2}\log|\mathbf{A}|$. In general, models that are compatible with more possible data patterns will have larger values of $|\mathbf{A}|$ and hence smaller values of this term. (The remaining term, $\frac{M}{2}\log(2\pi)$, arises in the formal derivation of the Laplace approximation and does not have a direct interpretation in terms of rewarding data fit or penalizing complexity.)[21]

[21] A more intuitive understanding of the Laplace approximation can be gained by considering the special case of a one-parameter model. θ_{MAP} is then a single value (scalar) that locally maximizes $\ell'(\theta; \mathbf{G}, D)$, and the log of the Laplace approximation to the posterior of model $\mathbf{G}$ is given by:

$$\log \Pr_{\mathbf{G}}(D) \approx \log \Pr_{\mathbf{G},\theta_{\text{MAP}}}(D) + \log \Pr_{\mathbf{G}}(\theta_{\text{MAP}}) + \log \sqrt{(2\pi)/c}$$

where c is the **curvature** of $\ell'(\theta; \mathbf{G}, D)$ in the vicinity of θ_{MAP}. Large curvature indicates that the parameter must be close to θ_{MAP} in order to attain a high value of the objective function. In other words, large curvature indicates that the model fails to properly account for the data except in a small region of the parameter space; values outside of that region correspond to other possible data patterns. Since complex models are able to fit more data sets, and therefore have less parameter space to devote to any given pattern, curvature is a quantitative index of complexity. The term $\log \sqrt{(2\pi)/c}$ translates larger curvatures into smaller values of $\log \Pr_{\mathbf{G}}(D)$, thereby penalizing more complex models. This line of reasoning extends to multidimensional parameter spaces, with $|\mathbf{A}|$ (the determinant of the Hessian) playing the role of the curvature.

Note that we assume, here and in the text, that θ_{MAP} is a maximizing point of $\ell'(\theta; \mathbf{G}, D)$ (rather than a minimizing point or saddlepoint), and that $\ell'(\theta; \mathbf{G}, D)$ has a single global maximum. Both assumptions hold in the MaxEnt

In adopting the Laplace approximation and its penalty for complexity, we depart from the position expressed in the following quote from CP (p. 329):

> It is perhaps not surprising that the Harmony-based account outperforms the similarity metric [i.e., natural classes similarity], since our learning simulation essentially fits the constraint values to the observed data. This is in line with one of the central points of our paper: that the cross-linguistic differences found in place co-occurrence patterns go beyond what is predicted by the similarity metric. Given such differences, an account of them must provide learners with a means of constructing language-specific grammars.

We do not dispute the claim that some cross-linguistic differences must be stipulated in the form of parameter settings, constraint rankings or weightings, etc. Nevertheless, we thank that the view expressed in this quote oversimplifies the model comparison problem, as it could be used to justify essentially any degree of theoretical complexity or lack of restrictiveness. For example, suppose we were interested in comparing two HG theories of co-occurrence restrictions: the constraint set of CP, and an alternative in which there is a separate constraint for every universally-possible consonant pair. Both theories have a finite set of constraints whose weights could be learned from the data in the way proposed by CP. The more restrictive theory (in this case CP's) could not possibly fit the lexical data better than the less restrictive alternative. And, because the sample sizes involved are relatively small (i.e., the number of roots available for any given language is limited), it is to be expected on purely statistical grounds that consonant pairs with the same grammatical status will often have different frequencies. Therefore, it is almost certain that CP's theory would fit the data *worse* than the less restrictive theory. We would not want to automatically take this as evidence that CP's theory is incorrect, or that the 'theory' of consonant co-occurrence should amount to little more than a list of observed frequencies in each language. Rather, we need a basis on which to determined *how much worse* the more restrictive theory must fare, relative to the data, in order for the less restrictive theory (or some theory of intermediate complexity) to be supported.

We have adopted one well-motivated quantitative approach to this problem. Many alternatives are available and could, if properly applied, yield important results that converge or conflict with our own.[22] What we do not find appropriate is an approach that ignores the problem and just pursues better fits to the quantitative details of individual languages (or to other types of phonological data, such as typological generalizations and experimental findings).

* * *

To summarize, the method of model comparison that we adopt is as follows. Given a data set D and competing models $\mathbf{G}^{(1)}, \mathbf{G}^{(2)}, \ldots$, each with its own parameter set $\theta^{(1)}, \theta^{(2)}, \ldots$, compute the Laplace approximation of $\log \Pr_{\mathbf{G}^{(i)}}(D)$ for each model. This single quantity, which estimates the log posterior probability of the model, encompasses both degree of fit to the data and complexity of the model. As anticipated in §3.2.1, calculating $\log \Pr_{\mathbf{G}^{(i)}}(D)$ involves finding $\theta^{(i)}_{\text{MAP}}$. It also involves calculating the Hessian $\mathbf{A}^{(i)} = -\nabla\nabla \log \Pr(\theta|D, \mathbf{G}^{(i)})|_{\theta^{(i)}_{\text{MAP}}}$, which is a function of the MAP parameters; many implementations of function optimization can return the Hessian (or a numerical approximation of it) from a MAP fit. Once

framework that we adopt below (e.g., Della Pietra et al. 1997). This framework thus appears well-suited to the Laplace approximation, which is known to be poor when there are multiple local minima of the objective function (e.g., Bishop 2006, p. 216).

[22] The alternatives include Minimum Description Length (MDL; e.g., Rissanen 1978, 1996, 1999, 2001; Grünwald 1998, 2000; Grünwald et al. 2005; Pitt et al. 2002); landscaping (e.g., Kim et al. 2004; Wagenmakers et al. 2004); and Parameter Space Partitioning (PSP; Pitt et al. 2006). The alternative that is closest to our own involves approximating posterior probabilities with simulation rather than the Laplace method (see, e.g., Myung and Pitt 1997). For additional discussion of MDL, see §5.2.

the Laplace approximations have been calculated, preference is given to those models that achieve higher values, which correspond to better explanations of the data.

3.2.3 Maximum Entropy grammars

The preceding discussion establishes the roles that the quantities $\ell(\theta; \mathbf{G}, D)$, $\ell'(\theta; \mathbf{G}, D)$, and $\log \Pr_{\mathbf{G}}(D)$ play in model selection, but says nothing about the form of the model or grammar $\mathbf{G}$ and its parameters θ. For reasons already discussed (see the introduction to §3), we adopt the MaxEnt grammar formalism, which is briefly reviewed here (see also Goldwater and Johnson 2003; Hayes and Wilson 2008 for other applications of MaxEnt to phonology).

A MaxEnt grammar consists of a set of constraints $\mathbf{G} = \{C_1, C_2, \ldots, C_n\}$ and a corresponding set $\theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$ of constraint weights. (The sets $\mathbf{G}$ and θ can also be treated as vectors by imposing a fixed but arbitrary ordering on the constraint indices.) The **Harmony** of a representation xy according to the grammar is defined as the negative of its weighted sum of constraint violations, exactly as in Harmonic Grammar (Smolensky 1986; Legendre et al. 1990a,b; Smolensky and Legendre 2005; Pater 2009):

$$H_{\mathbf{G},\theta}(xy) = -\sum_{i=1}^{M} \theta_i \cdot C_i(xy) \tag{13}$$

Note that we require all constraint violations and strengths to be greater than or equal to 0. This justifies the negative sign in (13), since greater violation should entail lower Harmony.

Harmony values are transformed to probabilities by taking the exponential and normalizing:

$$\Pr_{\mathbf{G},\theta}(xy) = \exp[H_{\mathbf{G},\theta}(xy)]/Z_{\mathbf{G},\theta}; \qquad Z_{\mathbf{G},\theta} = \sum_{x'y' \in \Omega} \exp[H_{\mathbf{G},\theta}(x'y')] \tag{14}$$

where Ω stands for the set of all phonological representations under consideration. In this paper, $\Omega = \Sigma \times \Sigma$, where Σ is the consonant inventory. That is, we follow a long line of research on OCP-Place in analyzing co-occurring consonant *pairs* together with their frequencies.

The term Z_θ is a **normalizing constant** (also called the **partition function**) that ensures that the total probability of all representations in Ω sums to 1. In general, calculation of $Z_{\mathbf{G},\theta}$ is non-trivial because Ω is infinite: the calculation therefore cannot be done explicitly (i.e., by applying (13) to each member of a list of representations), instead requiring so-called dynamic programming techniques (e.g., Mohri 1998, 2002; Eisner 2001, 2002). In the present context, however, the set of possible representations is finite by hypothesis. It is the set of consonant pairs drawn from the language-particular segment inventory Σ. Therefore, $Z_{\mathbf{G},\theta}$ is easily calculated by directly summing the exponentials of the harmonies of all consonant pairs.

A major advantage of explicitly stochastic grammar formalisms like MaxEnt — in comparison to the non-probabilistic methods that have been employed in previous quantitative research on OCP-Place — is that the relation between the probability of a consonant pair and the expected attestedness of the pair follows mathematically rather than being stipulated. Assume that the size of the corpus of consonant pairs, which is determined by the number and types of roots in the language, is fixed. Specifically, let N be the total number of consonant pair tokens in the data. If the observed consonant pairs are assumed to result from a generative process that makes N independent draws from the probability distribution defined by (14), then the frequencies of the consonant pair types will have a **multinomial distribution** (e.g., Agresti, 2002; Simonoff, 2003; Wasserman, 2004) with parameters N and $\{\Pr_{\mathbf{G},\theta}(xy)\}_{xy \in \Sigma \times \Sigma}$.

From general properties of the multinomial distribution it follows that the **expected** or **predicted value** of the frequency of a consonant pair type xy, O_{xy}, is $\mathbb{E}_{\mathbf{G},\theta}[O_{xy}] = N \cdot \Pr_{\mathbf{G},\theta}(xy)$. (Note that this is not generally identical to the E_{xy} value of the O/E statistic.) Furthermore, because the marginal distribution of each consonant pair type is binomial (e.g., Simonoff, 2003, p. 68), pointwise error bars around the predicted frequencies could be formed by using the binomial variance equation $\mathbb{V}_\theta[O_{xy}] = N \cdot \Pr_{\mathbf{G},\theta}(xy) \cdot (1 - \Pr_{\mathbf{G},\theta}(xy))$. (The error bars are 'pointwise' or 'cellwise' because they do not take into account the negative correlations among the frequencies of different consonant pair types; see, for example, Simonoff 2003, p. 68). All of this follows from definition (14) without further assumption.[23]

There are a number of ways to find the parameters θ_{MAP} that optimize $\ell'(\theta; \mathbf{G}, D)$, and thereby calculate expected co-occurrence frequencies and other quantities, given the constraints of a MaxEnt grammar. In the case studies reported below, we used the L-BFGS-B algorithm (Liu and Nocedal 1989) as implemented in the `optim` function of R (R Development Core Team 2008). The same function returns a numerical approximation of the Hessian evaluated at θ_{MAP}; we used this in calculating the Laplace approximation of $\log \Pr_{\mathbf{G}}(D)$.

4 Case studies

The preceding section established a method for evaluating and comparing quantitative models against consonant co-occurrence data. In this section, we compare the three quantitative models described in §1.3 and §2 on data from Arabic, Muna, Shona, and Wargamay. The first case study, on Arabic, provides additional details about the types of grammars that we tested.

4.1 Arabic

The data for this subsection is the list of *all ordered pairs of adjacent consonants* in the electronic dictionary of Arabic verbal roots analyzed by FPB and CP.[24] The consonants represented in the dictionary are [b t d tˤ dˤ k g q ʔ f θ ð s z sˤ zˤ ʃ χ ʁ ħ ʕ h m n l r w j]. Consonant pairs containing the glides [w j] were excluded from the analysis, as these segments may exhibit special behavior (McCarthy 1994, p. 204; cf. FPB, p. 196). The data set includes both homorganic and non-homorganic consonant pairs, as well as pairs containing identical consonants. We have found no reason to collapse across consonant order (cf. FPB, p. 186; CP, p. 295, note 6). The total frequency count of the ordered consonant pairs is $N = 4,119$.

The features we assume for Arabic consonants are similar to those of FPB and CP, except that we do not include [acute] or [emphatic] (see §2) and (following CP) we use a trivalued [stricture] feature in place of [continuant].[25] Emphatic consonants were specified Coronal but not Pharyngeal (see McCarthy 1994; Kenstowicz 1994 and discussion at the end of this subsection).

Five MaxEnt grammars, differing only in the form of the OCP-Place constraint, were tested in each of two conditions. The model that we refer to as **FPB Similarity** has a single OCP-Place constraint that assigns violations equal to natural classes similarity values (see (4) of §1.3.1). The models **CP Acceptability** and **CP Harmony** also has a single OCP-Place constraint for the purposes of model fitting. Violations

[23] In contrast, CP assume that Acceptability values should be correlated with O/E values, but do not derive or justify this relation based on any principle of Harmonic Grammar. Similarly, Anttila (2008) stipulates but does not derive the statement that OT-based **Complexity** (Anttila 2008, p. 702) should be correlated with O/E values. The discussion surrounding the definition of Complexity in Anttila (2008) might lead us to expect that it should be negatively correlated with *frequency*, but even this is not formally shown. Why there should be a relationship between Complexity and O/E values is unclear.

[24] We are grateful to Stefan Frisch for providing us with this dictionary, which is based on Cowan (1979).

[25] The feature charts and R code used for fitting the models are available from the first author upon request.

of the constraint were set equal to the Harmony and Acceptability values, respectively, of CP's analysis of Arabic (see §1.3.2 for discussion of how Harmony and Acceptability values were calculated).[26] The model **CP MaxEnt** employs the full set of OCP-Place proposed by CP (see (5) of §1.3.2), except that the constraints mentioning [emphatic] were excluded. Finally, our own model, referred to as **Weighted Features**, has a single OCP-Place constraint that assigns violations according to equations (6) and (7) of §2. The weighted feature similarity function of this model has one parameter for each place feature (Labial, Coronal, Dorsal, Pharyngeal) and one for each subsidiary feature ([sonorant], [stricture], [voice]; since Arabic lacks prenasalized segments, the [prenasalization] feature is irrelevant to this case study).

In addition to its OCP-Place constraint or constraints, each model includes a strict OCP constraint that is violated by identical consonants. To avoid redundant violations, strictly identical consonants were exempted from violating OCP-Place (as also in CP and FPB).

Each model was tested in two conditions, which differ only in the treatment of constraints on the individual members of a consonant pair. As observed by FPB (p. 208; see also Pierrehumbert 1994; Frisch 1996), position-specific segment distributions by themselves account for a substantial proportion of the variation in consonant co-occurrence frequencies. Our two conditions correspond to two rather different formalizations of this observation. In the first condition, there is one constraint ${}^*r(x)$ for each consonant x in the inventory and each position r (first or second member of an ordered consonant pair). The weights of these constraints were learned along with those of the OCP-Place and strict OCP constraints by maximizing the objective in (11). Given the optimizing weights θ^* ($= \theta_{\text{MAP}}$), the probability of a consonant pair xy according to the MaxEnt grammar is (as in equation (14)):

$$\text{(15)} \qquad \Pr_{\mathbf{G},\theta^*}(xy) \propto \exp(-\theta^*_{\text{OCP-Place}}\text{OCP-Place}(x,y)) \cdot \exp(-\theta^*_{\text{OCP}}\text{OCP}(x,y)) \cdot \exp(-\theta^*_{1(x)}) \cdot \exp(-\theta^*_{2(y)})$$

where the constant of proportionality is determined by summing over all pairs of consonants drawn from the inventory. (In the CP MaxEnt grammar, the term $-\theta^*_{\text{OCP-Place}}\text{OCP-Place}(x,y)$ is replaced by a sum of terms, one for each constraint in (5).) Note that the terms $\exp(-\theta^*_{1(x)})$ and $\exp(-\theta^*_{2(y)})$, which are the penalties for having consonants x and y in the first and second position of a consonant pair, respectively, correspond to $\psi_{1(x)}$ and $\psi_{2(y)}$ in the critique of O/E in §3.1.

The first condition contains one constraint for every possible consonant/position combination, and so is quite unrestrictive with respect to positional or marginal effects. While it is a strength of the MaxEnt framework that it allows probabilistically sound grammars to be constructed from many interacting constraints, and even though we granted all of the competing formulations of OCP-Place the same degree of descriptive freedom with respect to consonant/position association, we also wanted to consider an alternative with many fewer few parameters. Therefore, in the second condition, the weights of the ${}^*r(x)$ constraints were not optimized but rather fixed by the equations: $\theta^*_{1(x)} = -\log(O_{x+}/N)$ for x in the first position of a pair and $\theta^*_{2(y)} = -\log(O_{+y}/N)$ for y in the second position of a pair. Under this condition, the probability of a the pair xy is:

[26]The Harmony and Acceptability values for homorganic consonant pairs were taken from the file posted on-line at http://www.umich.edu/ coetzee/Muna_Arabic/. The values for non-homorganic consonant pairs were calculated from the constraint weights for Arabic reported in CP, (33), p. 326. The corresponding values for Muna were obtained in the same way (see CP, (24), p. 320).

$$\text{(16)} \qquad \Pr_{\mathbf{G},\theta^*}(xy) \propto \exp(-\theta^*_{\text{OCP-Place}}\text{OCP-Place}(x,y)) \cdot \exp(-\theta^*_{\text{OCP}}\text{OCP}(x,y)) \cdot \exp(-\theta^*_{1(x)}) \cdot \exp(-\theta^*_{2(y)})$$
$$= \exp(-\theta^*_{\text{OCP-Place}}\text{OCP-Place}(x,y)) \cdot \exp(-\theta^*_{\text{OCP}}\text{OCP}(x,y)) \cdot \frac{O_{x+}}{N} \cdot \frac{O_{+y}}{N}$$

The second line of (16) shows that the $\exp$ operation of MaxEnt simply undoes the $\log$s in the preceding definitions of $\theta^*_{1(x)}$ and $\theta^*_{2(y)}$. What remains is $(O_{x+}/N)\cdot(O_{+y}/N)$ — which is an estimate of the probability of the pair xy independently of constraints on consonant co-occurrence — multiplied by the same OCP-Place and OCP terms as in (15). Recall that $(O_{x+}/N) \cdot (O_{+y}/N)$ is simply the E_{xy} term value of the O/E statistic (8) divided by N. Hence, our second condition has a closer formal relationship to previous quantitative research on OCP-Place than the first condition (while avoiding the methodological flaws of O/E itself). This makes it all the more striking that the two conditions agree with one another, and to a certain extent overturn the conclusions of previous work based on O/E, as we now show.[27]

According to equations (15) and (16), *the probability of a consonant pair is an exponentially decreasing function of its OCP-Place violation*. This relationship, which we anticipated in §2, is constant across the two treatments of the marginal terms. In the FPB Similarity and Weighted Feature models, the degree of OCP-Place violation is equal to the similarity of the offending consonants, so probability is an exponentially decreasing function of similarity.

Table (a) below presents the results for the Arabic data in the first condition, with the weights of the marginal constraints $\{{}^*r(x)\}$ fit along with the other parameters. Several measures are reported for each model: the Laplace approximation to the log posterior probability of the model ('Laplace'; see (12)); the value of the objective function (11) evaluated at the MAP parameters ('logPost'); and three measures of correlation between the predicted and observed frequencies (*not* O/E values) for each ordered pair of consonants drawn from the Arabic consonant inventory. The correlation measures are Pearson's product-moment (r), Kendall's tau (τ), and Spearman's rho (ρ). Two values are given for each correlation statistic: the first is the correlation taken over all consonant pairs; the second, in parentheses, is the correlation taken over all non-identical homorganic pairs only. For all of the measures considered here, higher values are preferable.

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
FPB Similarity	-25130.62	-24950.39	<u>0.891</u> (<u>0.883</u>)	0.738 (0.626)	0.894 (0.771)
CP Acceptability	-25127.43	-24943.31	<u>0.894</u> (<u>0.856</u>)	0.733 (0.623)	0.888 (0.765)
CP Harmony	-25154.65	-24967.49	0.890 (0.853)	0.726 (0.625)	0.883 (0.767)
CP MaxEnt	-25083.03	**-24875.76**	**0.909** (**0.897**)	**0.750** (**0.646**)	**0.902** (0.791)
Weighted Features	**-25078.92**	-24881.11	0.908 (0.894)	0.749 (0.645)	0.901 (**0.793**)

(a) Arabic: marginal constraint weights fit to the data

The <u>underlined</u> values in Table (a) are the ones most comparable to the results reported by CP (see CP, Table 14, p. 327). In partial agreement with CP's findings, their Acceptability values contribute to a better correlation with the Arabic consonant co-occurrence frequencies than do FPB's Similarity values. However, the difference between the two approaches is not nearly as stark as CP claim: on the Pearson measure, the two approaches explain $.891^2 \approx 72\%$ (FPB) and $.849^2 \approx 79\%$ (CP) of the data variance;

[27]The two conditions considered in the text should be considered as merely convenient proxies for the correct theory of segment/position restriction in phonology (e.g., Beckman 1997, 1999; Steriade 1999; Zoll 2004). Like all previous work on OCP-Place, our proposal leaves the proper analysis of such effects, and their ultimate integration with restrictions on segment co-occurrence, to future research.

and the result reverses when only non-identical homorganic consonant pairs are considered ($.883^2 \approx 78\%$ of the variance explained with FPB Similarity as compared to only $.856^2 \approx 73\%$ with CP Acceptability). Furthermore, the model in which OCP-Place violations are instead determined by CP Harmony is dominated on all measures by FPB Similarity.

The several OCP-Place constraints proposed by CP do have the capacity to systematically outperform FPB Similarity, but only when they are weighted according to MaxEnt principles. Indeed, Table (a) shows that the CP MaxEnt model is preferable to the other four models on nearly all of the criteria considered — with the Laplace-based measure being the crucial exception. This measure, which incorporates a penalty for model complexity as discussed in §3.2.2, assigns a higher value to the Weighted Feature model relative to all others. These results are therefore consistent with our main claim that the definition of similarity for the purpose of OCP-Place violation reduces to a weighted sum of shared features, and that subsidiary features do not vary in weight across places of articulation within a language.

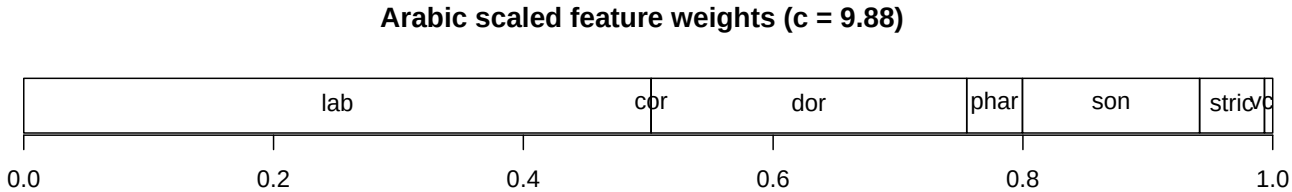
Parallel results are obtained in the second condition, with the weights of the marginal constraints $\{^*r(x)\}$ fixed to the corresponding marginal proportions (i.e., marginal frequencies divided by N), as shown in table (b) below.

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
FPB Similarity	-25019.31	-25015.18	0.882 (0.846)	0.714 (0.589)	0.876 (0.733)
CP Acceptability	-24986.58	-24978.6	0.887 (0.841)	0.723 (0.614)	0.880 (0.755)
CP Harmony	-24988.89	-24978.6	0.887 (0.841)	0.723 (0.615)	0.880 (0.755)
CP MaxEnt	-24916.85	**-24885.07**	**0.907** (**0.894**)	**0.748** (**0.647**)	**0.899** (**0.792**)
Weighted Features	**-24913.05**	-24890.84	0.906 (0.891)	0.747 (0.645)	0.898 (0.791)

(b) Arabic: marginal constraint weights fixed to empirical proportions

As before, the CP Maxent model is dominant on all measures except complexity-sensitive Laplace, which again prefers Weighted Features.[28]

To facilitate comparison of our analysis of Arabic co-occurrence restrictions with previous (and future) analyses, the features weights of our model in the second condition are displayed below. To construct this figure, the weights were normalized: that is, divided by their sum, which is shown as c in the graph. The constant c is a so-called **sensitivity factor** (e.g., Nosofsky 1986) that scales the similarity of all consonant pairs that violate OCP-Place relative to all consonant pairs that do not violate the constraint. To recover segment similarity values as defined in (6), the normalized weights of the shared features are summed and the result is multiplied by c: $sim_{\mathbf{w}}(x, y) = c \sum_{F \in \mathcal{F}} \hat{w}_F \delta_F(x, y)$. (The visual artifacts in the figure are caused by the very small weights of Coronal and [voice]; the former is zero and the latter, depicted at the far right edge, is nearly zero.)



Consistent with previous quantitative work, as well as the non-quantitative proposals of McCarthy

[28] We have found that the qualitative results reported in the text remain unchanged under slight variations of the models, such as a version of FPB Similarity in which OCP-Place violation is a logistic function of similarity. For reasons of space, we omit the quantitative details obtained under these variations.

(1988), Yip (1989), and others, we see that among the subsidiary features [sonorant] has the largest weight and [voice] the smallest, with [stricture] (erstwhile [continuant]) taking on an intermediate value. The unique contribution of the present analysis is the finding that all of the subsidiary features can be assigned constant weights across the four places of articulation. For example, there is no significant advantage to stipulating that [sonorant] is subsidiary to Coronal but not Labial. Rather, the large difference in the weights of those place features themselves ($w_{\text{Label}} \approx 5$ and $w_{\text{Coronal}} = 0$) is sufficient to explain the fact that essentially all Labial-Labial sequences are excluded in Arabic verbal roots whereas Coronal-Coronal sequences are heavily penalized only if they share [sonorant] or [stricture]. (For discussion of how constant subsidiary behavior can be obscured by differences in place weights, see §2.)

Considerations of space prevent a more detailed exegesis of our analysis. We conclude this subsection, therefore, by addressing two points of more general importance that are raised by the analysis of OCP-Place effects in Arabic:

- *Categorical restrictions in quantitative phonology*. If quantitative approaches to phonology are to advance beyond previous qualitative approaches, they must have the capacity to account for (near-) categorical restrictions as well as those that are gradient. This has not been the case in previous quantitative analyses of Arabic. Both FPB and CP acknowledge that their analyses do not rule out all unattested sequences that violate OCP-Place. This can be seen, under FPB's analysis, by sorting consonant pairs on their natural classes similarity values, and, under CP's analysis, by inspecting the weights of the OCP-Place and IDENT-PLACE constraints. For example, the Harmonic Grammar analysis of CP fails to rule out the zero-frequency pair [b-f] (CP, p. 326, note 16). In contrast, our weighted feature analysis heavily penalizes all Labial-Labial sequences, including [b-f], capturing the categorical restriction by assigning a large weight to Labial place. FPB suggest that feature weighting could solve the problems that they identify for the natural classes similarity model (p. 204). We have formally proposed feature weighting, and argued that it is superior to alternative models with respect to the Arabic data even when a penalty for free parameters is taken into account. CP offer only that their results 'may eventually prove to force a different approach to biases and/or phonotactic restrictiveness' (p. 326, note 16). We have offered a concrete alternative approach, founded in weighted constraint interaction, probability theory, and Bayesian learning, that unifies categorical and gradient phonotactics within a single quantitative framework.

- *Homorganicity and secondary place features*. Previous analyses of Arabic differ with respect to whether, and how, secondary place specifications contribute to OCP-Place violation (McCarthy 1994; Bachra 2000, FPB; see also Kenstowicz 1994, pp. 456ff. for a helpful review). Without attempting to resolve this issue, we point out one puzzling aspect of the co-occurrence pattern of Arabic that bears on it. Suppose that the emphatic coronals [tˤ dˤ sˤ zˤ], unlike the plain Coronals, bear a secondary Dorsal place specification in addition to their primary Coronal place (Kenstowicz 1994; Bachra 2000; FBP). If a segment with secondary place P is homorganic to a segment with primary place P, in the sense relevant for OCP-Place violation, then the emphatics are predicted to exhibit strong co-occurrence restrictions with all primary Dorsals. This prediction is supported by the finding that emphatics rarely co-occur with [k,g]. However, as noted by FPB (p. 2000), emphatics seem to co-occur relatively freely with [q], which is also Dorsal in the feature systems of McCarthy (1994), FPB, and others. Revoking the Dorsal feature of [q] (which is also secondarily Pharyngeal) is not a viable solution to this problem, because co-occurrence of [k,g] with [q] is tightly restricted. Therefore, we seem to have a failure of transitivity: in the sense relevant for OCP-Place evaluation, the emphatics are homorganic with [k,g]; and [k,g] are homorganic with [q]; but the emphatics and [q] somehow fail to be homorganic. Since transitivity is inherent to all existing notions of homorganicity, ours included, resolving this apparent inconsistency may shed considerable light on the formal nature of OCP-Place

evaluation in general and the features of Arabic velars, uvulars, and emphatics in particular.

4.2 Muna

CP argue that the consonant co-occurrence pattern of Muna, even more than that of Arabic, provides evidence for the superiority of their Harmonic Grammar model relative to the model of FPB. We have already seen that the difference between the two theories with respect to Arabic is much smaller than previously claimed, and uniformly favors the CP analysis only when it is set within the MaxEnt framework. Surprisingly, the Muna evidence turns out to provide an even weaker evidence for the HG model.

The data set for Muna consists of all ordered pairs of adjacent consonant pairs (i.e., pairs separated only by one or more vowels) in the electronic corpus based on van den Berg and Sidu (1996).[29] The consonants that occur in the corpus are [p b ɓ d̪ t d k g ᵐp ⁿt ⁿd ᵑk ᵑg f s ⁿs ʁ h m n ŋ r l w].[30] The size of the data set (i.e., total frequency of all ordered consonant pairs) is $N = 7,892$.

The five models of the previous subsection were tested on the Muna data under the same two conditions. Note that, because Muna has prenasalized consonants, the subsidiary feature [prenasalized] is active here. However, following CP's discussion (p. 302), we exempt consonant pairs that are identical except with respect to this feature — as well as identical consonant pairs — from OCP-Place. Only identical consonant pairs violate the strict OCP. Note also that, since Muna has only one Pharyngeal consonant ([h]), the weight of this feature is irrelevant. The results of the first condition are shown in table (a) below and those of the second condition in table (b).

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
FPB Similarity	-46678.47	-46486.93	0.919 (0.943)	0.699 (0.672)	0.866 (0.833)
CP Acceptability	-46728.94	-46532.52	0.909 (0.918)	0.697 (0.687)	0.865 (0.849)
CP Harmony	-46662.07	-46464.3	0.921 (0.934)	0.709 (0.721)	0.873 (0.872)
CP MaxEnt	-46647.41	**-46416.65**	**0.931** (**0.956**)	**0.712** (**0.716**)	**0.876** (**0.866**)
Weighted Features	**-46638.05**	-46424.11	**0.931** (0.954)	0.711 (0.708)	0.875 (0.860)

(a) Muna: marginal constraint weights fit to the data

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
FPB Similarity	-46575.26	-46567.71	0.912 (0.924)	0.680 (0.623)	0.852 (0.789)
CP Acceptability	-46632.78	-46620.34	0.901 (0.890)	0.675 (0.626)	0.849 (0.796)
CP Harmony	-46502.54	-46489.13	0.916 (0.927)	0.705 (0.709)	0.872 (0.865)
CP MaxEnt	-46517.25	**-46469.85**	**0.921** (**0.949**)	**0.705** (**0.705**)	**0.872** (**0.855**)
Weighted Features	**-46508.14**	-46477.69	**0.921** (0.948)	0.703 (0.694)	0.871(0.848)

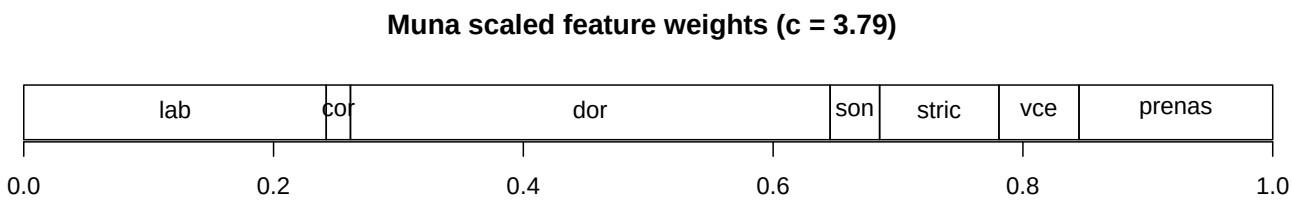
(b) Muna: marginal constraint weights fixed to empirical proportions

[29] We are grateful to Andries Coetzee for making this corpus available to us.

[30] The sound [d̪], which is described by van den Berg (1989, p. 18) as a voiced lamino-dental stop, is misidentified by CP as the coronal implosive [ɗ] (e.g., CP, (4), p. 294 and Appendix A, p. 332). This is no doubt due to van den Berg's convention of writing the labial implosive [ɓ] as ⟨bh⟩ and [d̪] as ⟨dh⟩. Fortunately, this error does not affect the calculation of OCP-Place violations for any of the theories considered in the text, since [d̪] is the only coronal consonant with a positive specification for CP's feature [implosive]. For concreteness, we simply replaced [implosive] with an ad-hoc feature [dental] in our own feature chart. We also corrected some additional errors that would affect natural classes similarity values: CP's Appendix A incorrectly specifies the Muna coronal nasal [n] as having the stricture [-narrow,-intermediate,+wide] and as being [-voice]. We changed the specifications of [n] to [+narrow,-intermediate,-wide] and [+voice], in conformity with the features given for the other two nasals [m ŋ], and recomputed natural classes similarity with the amended coronal feature chart.

On the parametric correlation measure (underlined values), the FPB Similarity model provides a better fit to the Muna data than the CP Acceptability model, contrary to the central claim of CP. (Note that CP Harmony does outperform FPB Similarity, but the opposite was true for Arabic.) Non-parametric correlations, log a posteriori probability, and the Laplace approximation also favor FPB Similarity. However, as in the case of Arabic, CP MaxEnt is superior to FPB Similarity on all measures. Finally, and most importantly for our argument, Weighted Features outperforms all others on the Laplace measure.

The optimized feature weights of our analysis (in the second condition) are shown in the following figure, which was constructed in the same way as the figure for Arabic in the previous subsection.



As in Arabic, Coronal has the smallest weight among the place features. The relationship between Labial and Dorsal is reversed, however. And, consistent with the descriptive analysis of CP (pp. 295-301), the subsidiary features are more evenly weighted in Muna than in Arabic. Note also that the sensitivity constant (c) is smaller here, indicating an overall weaker OCP-Place effect.

The conclusion of our model comparison up to this point is that the statistical evidence from both Muna and Arabic, when properly assessed, is consistent with the claim that the relevance or weight of a subsidiary feature cannot be stipulated on a place-specific basis. To further test this claim, we examined data from two additional languages, Shona and Wargamay, whose OCP-Place effects have not been investigated previously.

4.3 Shona

The data for this case study is based on the verbal forms in Hannan's dictionary of Shona (Hannan 1959, 1981, 1984), which most consistently represents the Zezuru dialect. A partial list of such forms, drawn from the first half of the dictionary (letters A-M) is available from the CBOLD project (http://www.ddl.ish-lyon.cnrs.fr/bdd/cbold/).[31] We supplemented this list with all of the forms from the second half of Hannan (1984) that could be automatically extracted by optical character recognition, and entered many additional forms by hand.[32] Hannan typically lists several suffixed versions of a given verbal root (though prefixes are listed separately).[33] We sought to minimize the influence of the phonological content of the suffixes by analyzing only the first vowel-separated consonant pair of each verbal form. The total number of such pairs is $N = 7,890$.

The Shona consonant inventory contains [p b t̪ d̪ k g ᵐb ⁿd̪ ᵑg m n̪ ɲ ŋ p͡f b͡v t͡s̪ d͡z̪ t͡ʃ d͡ʒ ⁿd͡ʒ f v s̪ z̪ ᵐv ⁿz̪ ʃ ʒ r j w ɦ] and two labialized coronal fricatives, which are standardly written as ⟨sv⟩ and ⟨zv⟩ (Fortune 1955, 1964; Hannan 1959, 1981) and variously transcribed as [s̪ʷ z̪ʷ] (Doke 1931; Maddieson 1990; Sagey 1990), [s̫ z̫] (Bladon et al. 1987; Ladefoged and Maddieson 1996), or [s̯ʷ z̯ʷ] (Shosted 2006). (There is also

[31]We are grateful to Bruce Hayes for making this list available to us. Note that as of the present writing the CBOLD link given in the text is broken, but the list can be downloaded from Hayes's website: http://www.linguistics.ucla.edu/people/hayes/Phonotactics/index.htm

[32]Thanks to Claire Snodgrass for assistance on this part of the project.

[33]For information on Shona morphology see the Introduction to Hannan 1984 and Beckman 1997, pp. 9-11.

a prenasalized voiced labialized fricative, just as there are prenasalized counterparts for the other voiced stops, fricatives, and affricates except [b͡v ʒ ɦ].)

Many of the consonants listed above have counterparts generally described as labiovelarized: [pk tkw zgw kw ...]. Considerable descriptive and theoretical work has been done on the variable realization of labiovelarization within and across Shona dialects (Doke 1931) and on the proper phonetic and phonological representation of consonants with this property (Sagey 1990, pp. 159-166; Maddieson 1990; Ladefoged and Maddieson 1996, pp. 345-347). However, as in the case of Arabic emphatics we ignored this (presumably) secondary place specification in the present analysis. One clear direction for further research is to assign such secondary place features, as well as additional subsidiary features, their own weights.

For reasons of space, we limit the model comparison here to the two models that faired best on Arabic and Muna: CP MaxEnt and Weighted Features. Results from the same two conditions as in the previous case studies are given in tables (a) and (b) below.

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
CP MaxEnt	-55134.75	**-54715.02**	**0.830** (0.834)	0.526 (0.491)	0.650 (**0.604**)
Weighted Features	**-55131.27**	-54723.2	0.829 (0.834)	0.526 (0.491)	0.650 (0.603)

(a) Shona: marginal constraint weights fit to the data

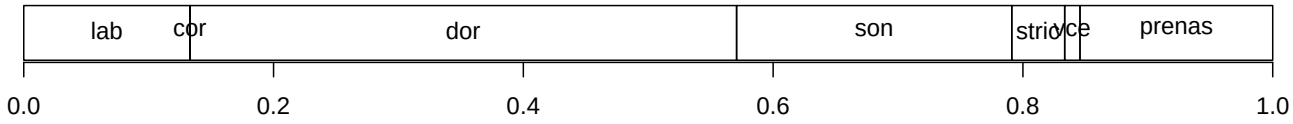
OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
CP MaxEnt	-54764.43	**-54719.96**	0.828 (**0.836**)	0.525 (**0.492**)	0.649 (**0.604**)
Weighted Features	**-54759.65**	-54727.13	0.828 (0.835)	0.525 (0.491)	0.649 (0.603)

(b) Shona: marginal constraint weights fixed to empirical proportions

In terms of the correlation measures (r, τ, ρ), the two models are hardly distinguishable; equivocal evidence such as this should be taken to support the more restrictive theory. The two probability-based measures (logPost, Laplace) do distinguish the models. If fit to the data is the primary criterion (logPost), CP MaxEnt is preferred. However, as in Arabic and Muna, consideration of both data fit and theoretical restrictiveness (Laplace) favors Weighted Features. This is true in both conditions.

The learned feature weights for Shona, displayed in the figure below, are an interesting mixture of those found for the previous two languages. As in Arabic, [sonorant] receives the largest weight among the subsidiary features (with [prenasalized], not relevant for Arabic, a close second, and [stricture] stronger than [voice]). But, as in Muna, Dorsal is weighted more highly than Labial, and the overall strength of the OCP-Place effect (indicated by c) is smaller than in Arabic. (Note that the weight of Coronal is zero, as also in Arabic, so there is a zero-width bar in the figure for this place feature.)

Shona scaled feature weights (c = 1.35)



A further point of empirical contact between Shona and Muna involves the prenasalized consonants. CP (pp. 295ff.) discovered that segments differing only in [prenasalization] were not subject to OCP-Place restriction in Muna, even though this feature is demonstrably subsidiary in the language. Our investigation of Shona (and Muna) converges with this conclusion: we found that the evaluations of the models above,

in which pairs like [b-^{m}b] were exempt from OCP-Place, were systematically better than the evaluations of parallel models in which such pairs violate the constraint. A project for future research is to find further statistical evidence that bears on this treatment of prenasalization, and to discuss what, if any, consequences this evidence has for the representation of prenasalized consonants (see also Herbert 1975, 1986; Feinstein 1979; Padgett 1991, 1995).

4.4 Wargamay

Our final case study focuses on Wargamay. Previous phonological studies of the language (Dixon 1981; Sherer 1994; Hayes 1995; Kager 1995; McGarrity 2002; Hayes and Wilson 2008) have not examined OCP-Place effects. Indeed, Hayes and Wilson (2008) aims to provide a complete phonotactic analysis of Wargamay, but the constraint schema and projections (or tiers) adopted in that paper cannot capture unbounded consonant-to-consonant dependencies (Hayes and Wilson 2008, p. 426). Therefore, the present analysis fills gaps in the study of Wargamay and in the theoretical development of MaxEnt phonotactics.

The consonant inventory of Wargmay is [b d ɟ m n ɲ ŋ r l ɻ j w]. The feature set and list of forms analyzed here are similar to those in Hayes and Wilson (2008). The total number of ordered consonant pairs in the data is smaller than in the other three languages, $N = 1,634$, due to the limited number of forms available.[34]

The results of model comparison are given in the two tables below, which have a by now familiar format.

OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
CP MaxEnt	-7467.61	**-7371.10**	0.957 (0.935)	0.712 (0.574)	0.864 (0.723)
Weighted Features	**-7466.92**	-7371.19	0.957 (0.935)	0.712 (0.574)	0.864 (0.723)

(a) Wargamay: marginal constraint weights fit to the data

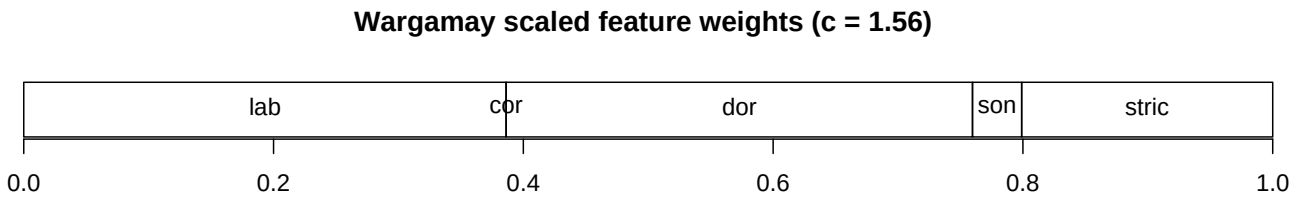
OCP-Place constraint	Laplace	logPost	Pearson r	Kendall τ	Spearman ρ
CP MaxEnt	-7393.90	-7375.39	0.958 (0.933)	0.705 (0.587)	0.858 (0.734)
Weighted Features	**-7393.21**	**-7375.38**	0.958 (0.933)	**0.706** (0.589)	0.858 (0.734)

(b) Wargamay: marginal constraint weights fixed to empirical proportions

As in Shona, correlation measures cannot distinguish the two models. Only when examined with the more sensitive techniques of probability theory (logPost, Laplace) do differences emerge, and only when model complexity is factored into the comparison (Laplace) does Weighted Features fare slightly better. Consistent with the overall perspective of this paper, we interpret small or non-existent *quantitative* differences as support for the *qualitatively* simpler or more restrictive theory of subsidiary features.

The learned feature weights for Wargamay, shown below, are similar to those of the other languages, with Labial and Dorsal weighted much more highly than Coronal (which in fact has a zero weight, as in Arabic and Shona). Perhaps more empirically interesting is the fact that the [stricture] subsidiary feature has a weight greater than that of [sonorant], reversing the relative weighting of these two features in Arabic. (The other two subsidiary features, [voice] and [prenasalized], are otiose here because the Wargamay consonant inventory contains no relevant contrasts: all consonants agree on these features.)

[34] Unlike Hayes and Wilson (2008), who analyzed inflected forms, we considered roots only: that is, the vocabulary items, with borrowings and reduplications excluded, of Dixon (1981).



This concludes our series of case studies, though we briefly discuss generalizations that hold across the learned feature weights of the four languages in the following, final section.

5 Summary and remaining issues

In this paper, we have made a number of contributions to the theory of OCP-Place and the methodology of quantitative phonology. We have argued, on both general grounds of restrictiveness and on the basis of four empirical case studies, against a main assumption of much previous work (quantitative and non-quantitative alike): namely, that subsidiary feature relevance is a matter of place-specific stipulation, parameterization, or weighting. The alternative proposed in §2 and tested in §4 is simpler, consistent with well-known patterns that have been assumed to require place-specific subsidiary behavior, and closely related to other formalizations of similarity in cognitive science.

We have also introduced a method, based on the Laplace approximation, by which quantitative phonological theories that differ in parameterization and functional form can be rationally compared. This method replaces the widely-adopted but mathematically and conceptually flawed practice of correlation with O/E values. Applications of the Laplace method in this paper support neither the most restrictive theory that we considered (the natural classes model of FPB) nor the least restrictive one (the weighted constraint model of CP). Instead, the proper balance between the freedom needed to *describe* language-particular data patterns and the restrictiveness required to *explain* aspects of those patterns is found in a theory in which each place and subsidiary feature contributes an independent weight to OCP-Place evaluation.

We look forward to further simplifications of subsidiary theory and other components of quantitative phonology. The method of model comparison that we have adopted, and the MaxEnt formalism within which we have framed competing theories, are fully general. As for subsidiary theory itself, the feature weights found in §4 reveal some consistencies across our (admittedly small) sample that might be unexpected without further restrictions on the theory. Most notably, the inequality $w_{\text{Labial}}, w_{\text{Dorsal}} > w_{\text{Coronal}}$ holds in all four languages. If borne out by other languages, this inequality could be stated as a universal and integrated with the general theory of Coronal unmarkedness (e.g., Paradis and Prunet 1991; McCarthy and Taub 1992; de Lacy 2006).[35] It can also be observed that the overall share of the total weight is greater, in all cases, for the place features as a group than for the subsidiary features. This is plausibly related to the special 'gating' role of place of articulation that was discussed in §2 and that justifies the name OCP-*Place*, though we do not yet know how to formalize the connection. Finally, [voice] is never the most highly weighted subsidiary feature, suggesting that previous restrictions on the set of subsidiary features, such as Padgett (1991, 1995)'s 'extended articulator group', should be reinstated quantitatively (e.g., with a *prior* that assigns greater penalties to non-zero weights of certain subsidiary features; on the uniform prior assumed here, see §3.2.1).

[35]For example, the violation levels of a *Place constraint could be identified with the place weights. If the weights are restricted as in the text — a type of gradient or quantitative underspecification of Coronal relative to Labial and Dorsal — then Coronal $\succ_{*\text{Place}}$ Labial, Dorsal in all languages.Thanks to Luigi Burzio, p.c., for related discussion.

We also acknowledge that the theory we have put forward is *too* simple, or at least not simple in all the right ways. It makes no provision for the effect of phonological distance, which is known to weaken OCP-Place and other co-occurrence restrictions (e.g., FPB; Frisch 2004; Martin 2004, 2007; Rose and Walker 2004; Wayment 2009). And it does not account for attested OCP effects on other features, such as [nasal] (e.g., Zuraw and Lu 2009) and perhaps [stricture] (e.g., Kaisse 1988). See Wayment (2009), which extends ideas of Burzio (2002b,a, 2005), for an attempt to derive many phonological phenomena from a theory of similarity based on quantitative phonological representations. Finally, though for purposes of comparison with previous literature we have analyzed consonant *pairs*, and ignored phonological constraints other than OCP-Place, it is straightforward to extend the model to a full probabilistic grammar of *roots* or even surface forms. Recall that a main motivation for working in the MaxEnt framework is that it can encompass many interacting constraints while remaining tractable and probabilistically sound (see the introduction to §3.2).

By way of conclusion, we briefly address two further issues, one that broadens the scope of our grammatical framework and one that indicates the limits of our approach to model comparison.

5.1 Relating phonotactics and alternations

CP briefly criticize MaxEnt models of the type proposed here on the grounds that 'phonotactics and alternations are given separate accounts' (p. 330), so that these models cannot take advantage of the OT approach to the duplication problem (see Clayton 1976; Goldsmith 1976; Kenstowicz and Kisseberth 1977; McCarthy 2002; Smolensky et al. 2006). The Harmonic Grammar model of CP does provide a uniform account of phonotactics and alternations (though the derivative notion of Acceptability has been applied to phonotactics only), and so could provide solutions to particular duplication problems (though no instances of alternation are analyzed in that paper). However, we have already seen that the particular HG grammars proposed by CP overgenerate (e.g., allowing [b-f] pairs in Arabic) and, more generally, that the HG approach to quantitative phonology lacks the probabilistic foundation enjoyed by MaxEnt (see especially §3.2.3). A remaining question is whether MaxEnt grammars can be deployed in a way that will relate phonotactics and alternations in the desired sense (see also CP, pp. 330-331; Coetzee and Pater 2008a; Hayes and Wilson 2008, pp. 423-424).

There is in fact a simple architecture in which static and dynamic phonological generalizations are accounted for with a single MaxEnt grammar. Let $\mathbf{M}$ stand for the Markedness part of the grammar (the Markedness constraints and their weights) and $\mathbf{F}$ stand for the Faithfulness part of the grammar (the Faithfulness constraints and their weights). All of the constraints used in the analyses up to this point qualify as Markedness. As CP point out (p. 318), Faithfulness constraints are crucial to the analysis of alternations in constraint-based phonology; for present purposes, we need not commit to a particular theory of what these constraints are. Further assume an operator $\oplus$ such that $\mathbf{G} = \mathbf{M} \oplus \mathbf{F}$, where $\mathbf{G}$ is the entire grammar.

The proposal is to factor the joint probability of a pair $(input, output)$ as follows:

(17) $$\Pr_{\mathbf{G}}(input, output) = \Pr_{\mathbf{M}}(input) \cdot \Pr_{\mathbf{M}\oplus\mathbf{F}}(output|input)$$

That is, the probability of the pair $(input, output)$ according to grammar $\mathbf{G}$ is equal to the product of (i) the probability of the input according to the Markedness part of the grammar and (ii) the probability of the output conditional on the input according to the entire grammar (i.e., the Markedness and Faithfulness parts). Note that in $\Pr_{\mathbf{M}}(input)$ it is the input that is evaluated by Markedness constraints, whereas in $\Pr_{\mathbf{M}\oplus\mathbf{F}}(output|input)$ the Markedness part of the grammar evaluates the output.

An important limiting case of (17) arises when all Faithfulness constraints are weighted much more highly than Markedness, prohibiting alternation. In this limit, equation (17) reduces to:

(18) $$\Pr_{\mathbf{G}}(input, output) = \Pr_{\mathbf{G}}(output) = \begin{cases} \Pr_{\mathbf{M}}(input) & \text{if } output \text{ is fully faithful to } input \\ 0 & \text{otherwise} \end{cases}$$

All of the analyses proposed in this paper fit without alteration into this 'high Faithfulness' regime, which suits our focus on phonotactics. (To simplify exposition, we have assumed that, if there is more than one fully faithful output for a given input, either **M** or the candidate generator restricts all probability mass to one of them. Nothing hinges on this assumption.)

To extend the analysis to account for alternations, the limit or idealization of alternation-free phonology can be relaxed by lowering the weights of one or more Faithfulness constraints (in which case the marginal probability of each output, $\Pr_{\mathbf{G}}(output)$, becomes more difficult to compute; see Jarosz 2006b). Importantly, the same Markedness constraints and weights that account for phonotactic restrictions would then also drive Faithfulness violation. We do not claim to know that this architecture is correct, but we do claim that, conceptually at least, it resolves the duplication problem in the same sense that OT and HG do. Of course, if the relationship between phonotactics and alternations is not as tight as has been claimed, then (17) could be weakened to allow the $\mathbf{M}$ component to evaluate inputs and outputs differently (e.g., with different weights), just as OT and HG could be weakened in various ways.[36]

5.2 Simplicity, restrictiveness, and constraint induction

Finally, the discussion throughout has assumed that the criteria of *simplicity* and *restrictiveness* (in the typological sense) are in accord. This is correct for theories of Universal Grammar (UG) that have a finite number of rules, principles, or constraints that are not derived from more fundamental entities and operations — as is characteristic, for example, of most work in OT (cf. Smolensky 1995; Hayes 1999). However, notions of simplicity and restrictiveness diverge if UG provides only the primitives from which rules/principles/constraints are constructed, rather than tightly delimiting the set of possible combinations of those primitives (e.g., Johnson 1993, 2008; Gildea and Jurafsky 1996; Albright and Hayes 2003; Heinz 2005; Boersma and Pater 2007a; Hayes and Wilson 2008; Albright 2009; see also FPB, pp. 190, 192, 215-218, for discussion of the idea that phonotactics are induced, in a way that is not clearly specified, from the lexicon). A UG that provides only features and constraint schema, as in Hayes and Wilson (2008), may be simple in comparison to a finite list of constraints, but much less restrictive with respect to allowed typological variation.

The fully general model selection problem for quantitative phonology, then, is to meaningfully compare theories that differ not only in the number and type of parameters (broadly construed) but also in whether a fixed, finite set of parameters is stipulated at all. The Laplace approximation cannot provide a full solution to this problem, as it assumes that a finite-length parameter vector is part of the definition of each model. (The approximation could be applied after a constraint induction model has learned a finite set of constraints, but this would overlook the fact that the constraints are not part of the model specification.)

In principle, the most comprehensive approach to this problem is offered by the Minimum Description Length framework (MDL; e.g., Rissanen 1999, 2001; Grünwald 2000; Grünwald et al. 2005; for a linguistic application, see Goldsmith 2001). In MDL, complexity of the model and failure to fit the data are penalized in common units (bits or nats). Therefore, if a model that provides only the resources for inducing

[36]The architecture in (17) may at first appear inconsistent with the spirit, if not the technical definition, of the Richness of the Base principle (Prince and Smolensky 2004; Davidson et al. 2004). However, the two are actually deeply connected: (17) states that all cross-linguistic differences in the probability distribution over inputs must be due to grammatical differences (specifically, differences in the weights of the Markedness constraints).

constraints is less complex than one that stipulates a fixed set of constraints, MDL would prefer the former unless it (unlike the latter) fails to account for significant aspects of the data. Typological restrictiveness is still favored in the way outlined in §3.2.2: a more restrictive model fits data patterns that it predicts under many parameter settings, so fit parameters can be specified with lower precision (fewer bits) without sacrificing descriptive adequacy.

Application of MDL is not completely straightforward in practice, however, because in its 'ideal' form the framework requires the most concise description of each model, and of the data in terms of each model, to be determined — in general, such computations are intractable (e.g., Li and Vitanyi 1997; Grünwald 2000; Grünwald 2005, pp. 6-7). The search for suitable approximations is an active research area (e.g., Grünwald et al. 2005; Myung et al. 2006) and the Laplace approximation may be one component of the result (e.g., Grünwald 2005, p. 50). Even if MDL has not yet reached the desired level of objectivity and tractability, we think its close connection to Bayesian inference and fundamental notions of informational complexity (e.g., Li and Vitanyi 1997; Vitanyi and Li 2000), as well as to evaluation metrics of the type that have been proposed in phonology (e.g., Chomsky and Halle 1968), suggest that it will develop into a useful tool for model comparison in general and for quantitive phonology in particular.

References

Ackely, D. H., Hinton, G. E., and Sejnowski, T. J. (1985). A learning algorithm for Boltzmann Machines. *Cognitive Science*, 9:147–169.

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley & Sons, Hoboken, NJ, second edition.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrox, B. N. and Caski, F., editors, *Second international symposium on information theory*, pages 267–281, Budapest. Akademiai Kiado.

Albright, A. (2009). Feature-based generalization as a source of gradient acceptability. *Phonology*, 26(1):9–41.

Albright, A. and Hayes, B. (2003). Rules vs. analogy in English past tenses: a computational/experimental study. *Cognition*, 90:119–161.

Alderete, J. (1997). Dissimilation as local conjunction. In Kusumoto, K., editor, *Proceedings of the North East Linguistic Society 27*, pages 17–31, Amherst, MA. GLSA.

Alderete, J. (2004). Dissimilation as local conjunction. In McCarthy, J. J., editor, *Optimality Theory in Phonology: A Reader*, pages 394–406. Blackwell, Malden, MA.

Aldrete, J. D. and Frisch, S. A. (2007). Dissimilation in the grammar and the lexicon. In de Lacy, P., editor, *The Cambridge Handbook of Phonology*. Cambridge University Press, Cambridge.

Anttila, A. (1997). Deriving variation from grammar. In Hinskens, F., van Hout, R., and Wetzels, W. L., editors, *Variation, Change, and Phonological Theory*, pages 35–68. John Benjamins, Amsterdam.

Anttila, A. (2002). Morphologically conditioned phonological alternations. *Natural Language & Linguistic Theory*, 20:1–42.

Anttila, A. (2008). Gradient phonotactics and the Complexity Hypothesis. *Natural Language & Linguistic Theory*, 26:695–729.

Bachra, B. N. (2000). *The Phonological Structure of the Verbal Roots in Arabic and Hebrew*. Brill, Leiden.

Bailey, T. M. and Hahn, U. (2001). Determinants of wordlikeness: Phonotactics or lexical neighborhoods. *Journal of Memory and Language*, 44:568–591.

Bailey, T. M. and Hahn, U. (2005). Phoneme similarity and confusability. *Journal of Memory and Language*, 52:339–362.

Beckman, J. (1997). Positional faithfulness, positional neutralisation and Shona vowel harmony. *Phonology*, pages 1–46.

Beckman, J. (1999). *Positional Faithfulness: An Optimality Theoretic Treatment of Phonological Asymmetries*. Garland, New York.

Berent, I. and Shimron, J. (1997). The representation of Hebrew words: Evidence from the obligatory contour principle. *Cognition*, 64:39–72.

Berent, I., Steriade, D., Lennertz, T., and Vaknin, V. (2007). What we know about what we have never heard: Evidence from perceptual illusions. *Cognition*, 104:591–630.

Berger, A. L., Della Pietra, S. A., and Della Pietra, V. J. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71.

Berkley, D. M. (1994a). The OCP and gradient data. *Studies in the Linguistic Sciences*, 24(1/2):59–72.

Berkley, D. M. (1994b). Variability in obligatory contour principle effects. In Beals, K., Denton, J. M., Knippen, R., Melnar, L., Suzuki, H., and Zeinfeld, E., editors, *Papers from the 30th Regional Meeting of the Chicago Linguistics Society, vol. 2: The Parasession on Variation in Linguistic Theory*, pages 1–12, Chicago. Chicago Linguistic Society.

Berkley, D. M. (2000). *Gradient OCP Effects*. PhD thesis, Northwestern University.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley, xxx.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Bladon, A., Clark, C., and Mickey, K. (1987). Production and perception of sibilant fricatives: Shona data. *Journal of the International Phonetic Association*, 17(1):39–65.

Bod, R., Hay, J., and Jannedy, S. (2003). *Probabilistic Linguistics*. MIT Press, Cambridge, MA.

Boersma, P. (1997). How we learn variation, optionality, and probability. In *Institute of Phonetic Sciences, University of Amsterdam, Proceedings 21*, pages 43–58.

Boersma, P. and Hayes, B. (2001). Empirical tests of the Gradual Learning Algorithm. *Linguistic Inquiry*, 32:45–86.

Boersma, P. and Pater, J. (2007a). Constructing constraints from language data: The case of Canadian English diphthongs. Talk presented at NELS 38, University of Ottawa, October 7.

Boersma, P. and Pater, J. (2007b). Testing gradual learning algorithms. Ms., University of Amsterdam and University of Massachusetts Amherst.

Boersma, P. and Pater, J. (2008). Convergence properties of a gradual learner in Harmonic Grammar. Ms., University of Amsterdam and University of Massachusetts Amherst.

Boersma, P. and Weenink, D. (2009). Praat: doing phonetics by computer (version 5.1.11) (computer program).

Broe, M. B. (1993). *Specification theory: the treatment of redundancy in generative phonology*. PhD thesis, University of Edinburgh.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44:108–132.

Buckley, E. (1997). Tigrinya root consonants and the OCP. In *Penn Working Papers in Linguistics*, volume 4.3, pages 19–51.

Burzio, L. (2002a). Missing players: Phonology and the past-tense debate. *Lingua*, 112:157–199.

Burzio, L. (2002b). Surface-to-surface morphology: When your representations turn into constraints. In Boucher, P., editor, *Many Morphologies*. Cascadilla Press.

Burzio, L. (2005). Sources of paradigm uniformity. In Downing, L., Hall, T. A., and Raffelsiefen, R., editors, *Paradigms in Phonological Theory*. Oxford University Press.

Chater, N., Tenenbaum, J. B., and Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10(7):287–291.

Chen, S. F. and Rosenfeld, R. (1999). A Gaussian prior for smoothing Maximum Entropy models. Technical report, CMU-CS-99-108-99-108.

Chen, S. F. and Rosenfeld, R. (2000). A survey of smoothing techniques for ME models. *IEEE Transations on Speech and Audio Processing*, 8(1):37–50.

Chen, Z. and Haykin, S. (2002). On different facets of regularization theory. *Neural Computation*, 14:2791–2846.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. MIT Press, Cambridge, MA.

Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and Log-Linear models. *Computational Linguistics*, 33(4):493–552.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, xxx:xxx–xxx.

Clayton, M. L. (1976). The redundance of underlying morpheme-structure conditions. *Language*, 52:295–313.

Coetzee, A. W. (2004). *What It Means to Be a Loser: Non-Optimal Candidates in Optimality Theory*. PhD thesis, University of Massachusetts, Amherst.

Coetzee, A. W. (2005). The OCP in the perception of English. In Frota, S., Vigario, M., and Freitas, M. J., editors, *Prosodies*, pages 223–245. Mouton de Gruyter, New York.

Coetzee, A. W. (2006a). Lexically determined grammar: The nature of lexical organization. Talk presented at the Workshop on Current Perspectives in Phonology, Phonology Fest 2006, Bloomington, Indiana.

Coetzee, A. W. (2006b). Variation as accessing "non-optimal" candidates. *Phonology*, 23(3):337–385.

Coetzee, A. W. (2008). Grammaticality and ungrammaticality in phonology. *Language*, 84(2):218–257.

Coetzee, A. W. and Pater, J. (2005). Gradient phonotactics in Muna and Optimality Theory. Talk presented at the 13th Manchester Phonology Meeting, May 2005, Manchester, England.

Coetzee, A. W. and Pater, J. (2006). Lexically ranked OCP-Place constraints in Muna. ROA-842.

Coetzee, A. W. and Pater, J. (2008a). The place of variation in phonological theory. In Goldsmith, J., Riggle, J., and Yu, A., editors, *Handbook of Phonological Theory*, pages xxx–xxx. Blackwell, 2nd edition.

Coetzee, A. W. and Pater, J. (2008b). Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language & Linguistic Theory*, 26:289–337.

Coleman, J. and Pierrehumbert, J. B. (1997). Stochastic phonological grammars and acceptability. In Coleman, J., editor, *Third Meeting of the ACL Special Interest Group in Computational Phonology: Proceedings of the Workshop*, pages 49–56, East Stroudsburg, PA. Association for Computational Linguistics.

Coon, J. and Gallagher, G. (2008). Distinguishing total and partial identify: Evidence from Chol. *Natural Language & Linguistic Theory*, xxx:xxx–xxx.

Cowan, J. M. (1979). *Hans Wehr: a dictionary of modern written Arabic*. Otto Harrasowitz, Wiesbaden.

Cutting, J. E., Bruno, N., Brady, N. P., and Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, 121:364–381.

Davidson, L. (2003). *The Atoms of Phonological Representation: Gestures, Coordination and Perceptual Features in Consonant Cluster Phonotactics*. PhD thesis, Johns Hopkins University, Baltimore, MD.

Davidson, L. (2006). Phonology, phonetics, or frequency: influences on the production of non-native sequences. *Journal of Phonetics*, 34:104–137.

Davidson, L., Smolensky, P., and Jusczyk, P. (2004). The initial and final states: Theoretical implications and experimental explorations of richness of the base. In Kager, R., Pater, J., and Zonneveld, W., editors, *Constraints in phonological acquisition*. Cambridge University Press, Cambridge.

de Bruijn, N. G. (1958). *Asymptotic methods in analysis*. North-Holland, Amsterdam.

de Lacy, P. (2006). *Markedness: Reduction and Preservation in Phonology*. Cambridge University Press, Cambridge.

Della Pietra, S., Della Pietra, V. J., and Lafferty, J. D. (1997). Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393.

Dietterich, T. G. (1997). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924.

Dixon, R. M. W. (1981). Wargamay. In Dixon, R. M. W. and Blake, B. J., editors, *Handbook of Australian Languages, Volume II*, pages 1–144. John Benjamins, Amsterdam.

Dmitrieva, O. and Anttila, A. (2008). The gradient phonotactics of English CVC syllables. In *Laboratory Phonology 11*.

Doke, C. M. (1931). *A Comparative Study in Shona Phonetics*. The University of Witwatersrand Press, Johannesburg.

Doya, K., Ishii, S., Pouget, A., and Rao, R. P. N. (2007). *Bayesian Brain: Probabilistic Approaches to Neural Coding*. MIT Press, Cambridge, MA.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley & Sons, New York.

Eisner, J. (2001). Expectational semirings: Flexible EM for finite-state transducers. In van Noord, G., editor, *Proceedings of the ESSLLI Workshop on Finite-State Methods in NLP (FSMNLP)*.

Eisner, J. (2002). Parameter estimation for probabilistic finite-state transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Elmedlaoui, M. (1995). Géométrie des restrictions de co-occurrence de traits en semitique et en berbere: Synchronie et diachronie. *Canadian Journal of Linguistics*, 40:39–76.

Ernestus, M. and Baayen, R. H. (2003). Predicting the unpredictable: Interpreting neutralized segments in Dutch. *Language*, 79:5–38.

Ernestus, M. and Neijt, A. (2008). Word length and the location of primary word stress in Dutch, German, and English. *Linguistics*, 46(3):507–540.

Feinstein, M. (1979). Prenasalization and syllable structure. *Linguistic Inquiry*, 10(2):245–278.

Fischer, M. (2005). A Robbins-Monro type learning algorithm for an entropy maximizing version of stochastic Optimality Theory. Master's thesis, Humboldt University, Berlin.

Fortune, G. (1955). *An Analytical Grammar of Shona*. Longmans, Green & Co., New York.

Fortune, G. (1964). *Elements of Shona (Zezuru Dialect)*. Longman Rhodesia.

Frisch, S. A. (1996). *Similarity and Frequency in Phonology*. PhD thesis, Northwestern University.

Frisch, S. A. (2000). Temporally organized representations as phonological units. In Broe, M. B. and Pierrehumbert, J. B., editors, *Papers in Laboratory Phonology V: Acquisition and the Lexicon*. Cambridge University Press.

Frisch, S. A. (2004). Language processing and segmental OCP effects. In Hayes, B., Kirchner, R., and Steriade, D., editors, *Phonetically-Based Phonology*, pages 346–371. Cambridge University Press, Cambridge.

Frisch, S. A., Broe, M. B., and Pierrehumbert, J. B. (1997). Similarity and phonotactics in Arabic. Bloomington, IN and Evanston, IL: Indiana University and Northwestern University, ROA-223.

Frisch, S. A., Large, N. R., and Pisoni, D. B. (2000). Perception of wordlikeness: Effects of segment probability and length on the processing of nonwords. *Journal of Memory and Language*, 42:481–496.

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. (2004). Similarity avoidance and the OCP. *Natural Language & Linguistic Theory*, 22(1):179–228.

Frisch, S. A. and Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language*, 77:91–106.

Gafos, A. I. (2003). Greenberg's asymmetry in Arabic: A consequency of stems in paradigms. *Language*, 79(2):317–355.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman & Hall, Boca Raton, second edition.

Gildea, D. and Jurafsky, D. (1996). Learning bias and phonological rule inducation. *Computational Linguistics*, 22:497–530.

Gluck, K., Bello, P., and Busemeyer, J. (2008). Special issue on model comparison. *Cognitive Science*, 32(8):1245–1424.

Goldrick, M. and Daland, R. (2009). Linking speech errors and phonological grammars: Insights from Harmonic Grammar networks. *Phonology*, 26(1):147–185.

Goldsmith, J. (1976). *Autosegmental Phonology*. PhD thesis, MIT, Cambridge, MA.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–198.

Goldsmith, J. and Riggle, J. (2007). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony.

Goldsmith, J. and Xanthos, A. (2009). Learning phonological categories. *Language*, 85(1):4–38.

Goldstein, L. (1994). Possible articulatory bases for the class of guttural consonants. In Keating, P. A., editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 234–241, Cambridge. Cambridge University Press.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a Maximum Entropy model. In Spenader, J., Eriksson, A., and Dahl, O., editors, *Proceedings of the Stockholm Workshop on Variation within Optimality Theory*, pages 111–120.

Goodman, B. (1992). Takelman dissimilation and the form of the OCP. *Working Papers of the Cornell Phonetics Laboratory*, 7:41–63.

Goodman, J. (2004). Exponential priors for maximum entropy models. In *Proceedings of HLT-NAACL-2004: Main Proceedings*.

Graff, P. and Jaeger, T. F. (2009). Modeling OCP effects in the Javanese lexicon. Talk given at the UCLA-UC Berkeley Conference on the Languages of Southeast Asia.

Greenberg, J. H. (1950). The patterning of root morphemes in Semitic. *Word*, 6:162–181.

Greenberg, J. H. and Jenkins, J. J. (1964). Studies in the psychological correlates of the sound system of American English. *Word*, 20:157–177.

Grünwald, P. (2005). Minimum Description Length Tutorial. In Grünwald, P. D., Myung, I. J., and Pitt, M. A., editors, *Advances in Minimum Description Length: Theory and Applications*, pages 23–80. MIT Press, Cambridge, MA.

Grünwald, P. D. (1998). *The minimum description length principle and reasoning under uncertainty (ILLG Dissertation Series D5 1998-03)*. PhD thesis, Centrum voor Wiskunde en Informatica, Amsterdam.

Grünwald, P. D. (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, 44:133–170.

Grünwald, P. D., Myung, I. J., and Pitt, M. A. (2005). *Advances in Minimum Description Length: Theory and Applications*. MIT Press/Bradford Books, Cambridge, MA.

Guion, S. G., Clark, J. J., Harada, T., and Wayland, R. P. (2003). Factors affecting stress placement for English nonwords include syllabic structure, lexical class, and stress patterns of phonologically similar words. *Language and Speech*, 46(4):403–427.

Halle, M. (1992). Phonological features. In Bright, W., editor, *International Encyclopedia of Linguistics*, pages 207–212. Oxford University Press.

Halle, M. (1995). Feature geometry and feature spreading. *Linguistic Inquiry*, 26:1–46.

Halle, M. (2005). Palatalization/velar softening: What it is and what it tells us about the nature of language. *Linguistic Inquiry*, 36:23–41.

Hammond, M. (1999). *The phonology of English: a prosodic Optimality-theoretic approach*. Oxford University Press, Oxford.

Hammond, M. (2004). Gradience, phonotactics, and the lexicon in English phonology. *International Journal of English Studies*, 4:1–24.

Hannan, M. (1959). *Standard Shona dictionary*. St Martin's Press, New York.

Hannan, M. (1981). *Standard Shona Dictionary*. The Literature Bureau, Harare, 2nd edition.

Hannan, M. (1984). *Standard Shona Dictionary*. College Press Publishers Pvt Ltd, revised edition.

Hansson, G. O. (2001). *Theoretical and typological issues in consonant harmony*. PhD thesis, University of California, Berkeley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Hay, J., Pierrehumbert, J., and Beckman, M. (2003). Speech perception, well-formedness, and the statistics of the lexicon. In Local, J., Ogden, R., and Temple, R., editors, *Papers in Laboratory Phonology VI*, pages 58–74, Cambridge. Cambridge University Press.

Hayes, B. (1995). *Metrical Stress Theory: Principles and Case Studies*. University of Chicago Press, Chicago.

Hayes, B. (1999). Phonetically-driven phonology: the role of Optimality Theory and inductive grounding. In Darnell, M., Moravcsik, E., Noonan, M., Newmeyer, F., and Wheatley, K., editors, *Functionalism and Formalism in Linguistics, volume I: General Papers*, pages 243–285. John Benjamins, Amsterdam.

Hayes, B. (2004). Phonological acquisition in Optimality Theory: the early stages. In Kager, R., Pater, J., and Zonneveld, W., editors, *Fixing Priorities: Constraints in Phonological Acquisition*, pages xxx–xxx. Cambridge University Press.

Hayes, B. and Wilson, C. (2008). A Maximum Entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Heinz, J. (2005). Learning phonotactic patterns from surface forms. In Baumer, D., Montero, D., and Scanlon, M., editors, *Proceedings of The 25th West Coast Conference on Formal Linguistics*, pages 186–194. Cascadilla Proceedings Project.

Herbert, R. K. (1975). Reanalyzing prenasalized consonants. *Studies in African Linguistics*, 6:105–123.

Herbert, R. K. (1986). *Language Universals, Markedness Theory, and Natural Phonetic Processes*. Mouton de Gruyter, New York.

Hinton, G. E. and Sejnowski, T. J. (1986). Learning and relearning in Boltzmann Machines. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, pages 282–317. MIT Press, Cambridge, MA.

Hocking, R. R. (1996). *Methods and Applications of Linear Models: Regression and the Analysis of Variance*. Wiley Series in Probability and Statistics. Wiley.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. John Wiley & Sons, New York.

Hopfield, J. J. (1984). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 81:3088–3092.

Hornik, K., Stinchcombe, M., and Halbert L. White, J. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366.

Hyman, L. (1995). Nasal consonant harmony at a distance: The case of Yaka. *Studies in African Linguistics*, 24(1):5–30.

Ito, C. (2007). Morpheme structure and co-occurrence restrictions in Korean monosyllabic stems. *Studies in Phonetics, Phonology and Morphology*, 13(3):373–394.

Ito, J. (1984). Melodic dissimilation in Ainu. *Linguistic Inquiry*, 15(3):505–513.

Ito, J. and Mester, A. (2003). On the sources of opacity in OT: coda processes in German. In Féry, C. and van de Vijver, R., editors, *The Syllable in Optimality Theory*, pages 271–303. Cambridge University Press, Cambridge.

Jacobs, A. M. and Grainger, J. (1994). Models of visual word recognition — sampling the state of the art. *Journal of Experimental Psychology: Human Perception and Performance*, 29:1311–1334.

Jarosz, G. (2006a). A Probabilistic Unsupervised Algorithm for Learning Optimality Theoretic Grammars. Talk presented at the 80th Annual Meeting of the Linguistic Society of America, Albuquerque, New Mexico.

Jarosz, G. (2006b). *Rich Lexicons and Restrictive Grammars: Maximum Likelihood Learning in Optimality Theory*. PhD thesis, Johns Hopkins University.

Jeffreys, W. H. and Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, 80:64–72.

Jelinek, F. (1999). *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA.

Jesney, K. (2007). The locus of variation in weighted constraint grammars. Gradience and Frequency in Phonology, Stanford University.

Johnson, M. (1993). The complexity of inducing a rule from data. In *The Proceedings of the Eleventh West Coast Conference on Formal Linguistics*. CLSI, Stanford, CA.

Johnson, M. (2008). Where do the rules come from? Talk presented at the University of Geneva, UIUC and the University of Massachusetts.

Johnson, M., Geman, S., Canon, S., Chi, Z., and Riezler, S. (1999). Estimators for stochastic "unification-based" grammars. In *The Proceedings of the ACL 1999*.

Jongman, A., Herd, W., and Al-Masri, M. (2007). Acoustic correlates of emphasis in Arabic. In *ICPhS XVI*, pages 913–916.

Kager, R. (1995). The metrical theory of word stress. In Goldsmith, J., editor, *The Handbook of Phonological Theory*, pages 367–402. Blackwell, Oxford.

Kager, R., Boll-Avetisyan, N., and Ao, C. (2008). Gradient phonotactic constraints for speech segmentation in a second language. In *BUCLD 2008*.

Kaisse, E. (1988). Modern Greek continuant dissimilation and the OCP. Ms., University of Washington.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90:1311–1334.

Katamba, F. (2006). Phonetically motivated word formation. In *The Encyclopaedia of Language and Linguistics*, pages 411–414. Elsevier, 2nd edition.

Kawahara, S., Ono, H., and Sudo, K. (2006). Consonant co-occurrence restrictions in Yamato Japanese. In *Japanese/Korean Linguistics*, volume 14, pages 27–38. CSLI Publications, Stanford.

Kelly, M. H. (1991). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, 99:349–364.

Kelly, M. H. and Martin, S. (1994). Domain-general abilities applied to domain-specific tasks: Sensitivity to probabilities in perception, cognition, and language. *Lingua*, 92:105–140.

Kenstowicz, M. (1986). Multiple linking in Javanese. In *Proceedings of NELS 16*, pages 230–248.

Kenstowicz, M. (1994). *Phonology in Generative Grammar*. Blackwell, Cambridge, MA.

Kenstowicz, M. and Kisseberth, C. (1977). *Topics in Phonological Theory*. Academic Press, New York.

Kessler, B. and Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language*, 37:295–311.

Kim, W., Navarro, D. J., Pitt, M. A., and Myung, I. J. (2004). An MCMC-based method of comparing connectionist models in cognitive science. In *Advances in Neural Information Processing Systems*, volume 16, pages 937–944.

Kirby, J. P. and Yu, A. C. L. (2007). Lexical and phonotactic effects on wordlikeness judgments in cantonese. In Trouvain, J. and Barry, W. J., editors, *Proceedings of the 16th International Congress of Phonetics Science*, pages 1389–1392, Dudweiler, Germany. Pirrot.

Klein, D. and Manning, C. (2003). Maxent models, conditional estimation, and optimization, without the magic. Tutorial presented at NAACL-03 and ACL-03.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7):320–326.

Ladefoged, P. (1969). The measurement of phonetic similarity. In *Proceedings of the conference on Computational Linguistics*, Sweden. Sang Saby.

Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Blackwell, Cambridge, MA.

Lamontagne, G. A. (1993). *Syllabification and Consonant Cooccurrence Conditions*. PhD thesis, University of Massachusetts, Amherst.

Leben, W. (1973). *Suprasegmental Phonology*. PhD thesis, MIT, Cambridge, MA.

Lee, Y. and Goldrick, M. (2008). The emergence of sub-syllabic representations. *Journal of Memory and Language*, 59:155–168.

Legendre, G., Miyata, Y., and Smolensky, P. (1990a). Harmonic grammar – a formal multi-level connectionist theory of linguistic well-formedness: An application. In *Proceedings of the Cognitive Science Society*, volume 12, pages 884–891.

Legendre, G., Miyata, Y., and Smolensky, P. (1990b). Harmonic grammar – a formal multi-level connectionist theory of well-formedness: Theoretical foundations. In *Proceedings of the Cognitive Science Society*, volume 12, pages xxx–xxx.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York, second edition.

Li, M. and Vitanyi, P. (1997). *An Introduction to Kolmogorov Complexity and Its Applications*. Springer-Verlag, New York, 2nd edition.

Lightner, T. M. (1973). Against morpheme structure rules and other things. In Kenstowicz, M. J. and Kisseberth, C. W., editors, *Issues in Phonological Theory: Proceedings of the Urbana Conference on Phonology*, pages 53–60. Mouton, The Hague.

Liu, D. C. and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3, (Ser. B)):503–528.

Lunn, D. J., Thomas, A., Best, B., and Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:3250–337.

Ma, W. J., Beck, J. M., and Pouget, A. (2006). Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9(11):1432–1438.

MacEachern, M. R. (1999). *Laryngeal Cooccurrence Restrictions*. Garland, New York.

MacKay, D. J. C. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge.

Maddieson, I. (1990). Shona velarization: Complex consonants or complex onsets? In *Working Papers in Phonetics*, number 74, pages 16–34. Departiment of Linguistics, UCLA.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the Sixth Conference on Natural Language Learning (CoNLL-2002)*, pages 49–55.

Martin, A. (2004). The structural notion of locality: gradient distance effects in navaho sibilant harmony. Paper presented at the LSA Annual Meeting.

Martin, A. T. (2007). *The Evolving Lexicon*. PhD thesis, University of California, Los Angeles.

McCarthy, J. (1979). *Formal Problems in Semitic Phonology and Morphology*. PhD thesis, MIT.

McCarthy, J. (1988). Feature geometry and dependency: a review. *Phonetica*, 45:84–108.

McCarthy, J. (2002). *A Thematic Guide to Optimality Theory*. Cambridge University Press, Cambridge.

McCarthy, J. and Prince, A. (1995). Faithfulness and identity in Prosodic Morphology. In Beckman, J., Dickey, L. W., and Urbanczyk, S., editors, *University of Massachusetts Occasional Papers in Linguistics 18*, pages 249–384. GLSA, Amherst, MA.

McCarthy, J. and Prince, A. (1999). Faithfulness and identity in prosodic morphology. In Kager, R., van der Hulst, H., and Zonneveld, W., editors, *The Prosody-Morphology Interface*, pages 218–309. Cambridge University Press, Cambridge.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12:373–418.

McCarthy, J. J. (1986). OCP effects: Gemination and antigemination. *Linguistic Inquiry*, 17:207–263.

McCarthy, J. J. (1994). The phonetics and phonology of Semitic pharyngeals. In Keating, P. A., editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 191–233. Cambridge University Press.

McCarthy, J. J. and Taub, A. (1992). Review of Carole Paradis and Jean-Francois Prunet, eds., *The special status of coronals: Internal and external evidence*. *Phonology*, 9:363–370.

McClelland, J. L. and Wyk, B. C. V. (2006). Graded constraints on English word forms. Ms., Department of Psychology and Center for the Neural Basis of Cognition, Carnegie Mellon University.

McConvell, P. (1988). Nasal cluster dissimilation and constraints on phonological variables in Gurindji and related languages. *Aboriginal Linguistics*, 1:135–165.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Chapman and Hall, New York, second edition.

McGarrity, L. W. (2002). On the typological predictions of fixed vs. complementary rankings of stress constraints. In Aguwele, A. and Park, H., editors, *Online Proceedings of the 2002 Texas Linguistics Society*. Texas Linguistic Forum, University of Texas, Austin.

Medin, D. L. and Shaeffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, 85:207–238.

Mester, R. A. (1986). *Studies in Tier Structure*. PhD thesis, University of Massachusetts, Amherst.

Mester, R. A. (1988). *Studies in Tier Structure*. Outstanding dissertations in linguistics. Garland, New York.

Miller, G. A. and Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, 27(2):338–352.

Minsky, M. L. and Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press, Cambridge, MA.

Mohri, M. (1998). General algebraic frameworks and algorithms for shortest-distance problems. Technical report, Technical Memorandum 981210-10TM, AT&T Labs Research.

Mohri, M. (2002). Semiring frameworks and algorithms for shortest-distance problems. *Journal of Automata, Languages and Combinatorics*, 7(3):321–350.

Moreton, E. (2008). Modelling modularity bias in phonological pattern learning. In Abner, N. and Bishop, J., editors, *Proceedings of the 27th West Coast Conference on Formal Linguistics*, pages 1–16, Somerville, MA. Cascadilla Proceedings Project.

Myers, J. and Tsay, J. (2005). The processing of phonological acceptability judgments. In *Proceedings of the Symposium on 90-92 NSC Projects*, pages 25–54, Taipei, Taiwan.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, 44:190–204.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, 47:90–100.

Myung, I. J. and Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, 4(1):79–95.

Myung, J. I., Navarro, D. J., and Pitt, M. A. (2006). Model selection by normalized maximum likelihood. *Journal of Mathematical Psychology*, 50:167–179.

Nagy, N. and Reynolds, B. (1997). Optimality Theory and word-final deletion in Faeter. *Language Variation and Change*, 9:37–55.

Navarro, D. J. and Griffiths, T. L. (2008). Latent features in similarity judgments: a nonparametric Bayesian approach. *Neural Computation*, 20(11):2597–2628.

Navarro, D. J. and Lee, M. D. (2004). Common and distinctive features in stimulus representation: A modified version of the contrast model. *Psychonomic Bulletin & Review*, 11:961–974.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal Bayesian decision process. *Psychological Review*, 113(2):327–357.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Oaksford, M. and Chater, N., editors (1998). *Rational Models of Cognition*. Oxford University Press, Oxford.

Ohala, J. J. and Ohala, M. (1986). Testing hypotheses regarding the psychological reality of morpheme structure constraints. In Ohala, J. J. and Jaeger, J. J., editors, *Experimental Phonology*, pages 239–252. Academic Press, San Diego, CA.

Onnis, L. and Christiansen, M. H. (2009). Lexical categories at the edge of the word. *Cognition (to appear)*, xxx:xxx–xxx.

Orbán, G., Fiser, J., Aslin, R. N., and Lengyel, M. (2008). Bayesian learning of visual chunks by human observers. *Proceedings of the National Academy of Sciences*, 105(7):2745–2750.

Padgett, J. (1991). *Stricture in Feature Geometry*. PhD thesis, University of Massachusetts, Amherst.

Padgett, J. (1995). *Stricture in Feature Geometry*. Dissertations in Linguistics. CSLI Publications, Stanford, CA.

Padgett, J. (2002a). Feature classes in phonology. *Language*, 78(1).

Padgett, J. (2002b). Russian voicing assimilation, final devoicing, and the problem of [v]. *Natural Language & Linguistic Theory*, xxx:xxx–xxx.

Padgett, J. (2008). Glides, vowels, and feature. *Lingua*, xxx:xxx–xxx.

Paradis, C. and Prunet, J.-F., editors (1991). *The special status of coronals: Internal and external evidence*, volume 2 of *Phonology and Phonetics*. Academic Press, New York.

Pater, J. (2008a). Gradient phonotactics in Harmonic Grammar and Optimality Theory. Ms., University of Massachusetts, Amherst.

Pater, J. (2008b). Gradual learning and convergence. *Linguistic Inquiry*, 29(2):334–345.

Pater, J. (2009). Weighted constraints in generative linguistics. *Cognitive Science*, xxx(xxx):xxx–xxx.

Pierrehumbert, J. B. (1993). Dissimilarity in the Arabic verbal roots. In Schafer, A., editor, *Proceedings of the North East Linguistic Society 23*, pages 367–381, Amherst, MA. GLSA.

Pierrehumbert, J. B. (1994). Syllable structure and word structure: a study of triconsonantal clusters in English. In Keating, P., editor, *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology III*, pages 168–188. Cambridge University Press, Cambridge.

Pierrehumbert, J. B. (2001). Stochastic phonology. *GLOT 5*, pages 1–13.

Pierrehumbert, J. B. (2003). Probabilistic phonology. In Bod, R., Hay, J., and Jannedy, S., editors, *Probabilistic Linguistics*, pages 177–228. MIT Press.

Pierrehumbert, J. B. (2006). The statistical basis of an unnatural alternation. In Goldstein, L. and Best, D. H. W. C., editors, *Laboratory Phonology VIII, Varieties of Phonological Competence*, pages 81–107. Mouton de Gruyter, Berlin.

Pitt, M. A., Kim, W., Navarro, D. J., and Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, 113(1):57–83.

Pitt, M. A. and Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, 6(10):421–425.

Pitt, M. A., Myung, I. J., and Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, 109(3):472–491.

Pitt, M. A. and Navarro, D. J. (2005). Tools for learning about computational models. In Cutler, A., editor, *Twenty-First Century Psycholinguistics: Four Cornerstones*, chapter 21, pages 347–362. Lawrence Erlbaum Associates.

Pozdniakov, K. and Segerer, G. (2007). Similar place avoidance: A statistical universal. *Linguistic Typology*, 11(2):307–348.

Prince, A. and Smolensky, P. (1993/2004). *Optimality Theory: Constraint interaction in generative gram-*

mar. Blackwell, Cambridge, MA. [Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick, NJ, April 1993].

Prince, A. and Tesar, B. (2004). Learning phonotactic distributions. In Kager, R., Pater, J., and Zonneveld, W., editors, *Fixing Priorities: Constraints in Phonological Acquisition*, pages 245–291. Cambridge University Press, Cambridge.

R Development Core Team (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org. ISBN 3-900051-07-0.

Riggle, J. (2004). *Generation, Recognition, and Learning in Finite State Optimality Theory*. PhD thesis, University of California, Los Angeles.

Riggle, J. (2009). Generating contenders. Rutgers Optimality Archive 1044.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14:465–471.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42:40–47.

Rissanen, J. (1999). Hypothesis selection and testing by the mdl primciple. *The Computer Journal*, 42:260–269.

Rissanen, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, 47:1712–1717.

Roberts, S. and Pashler, H. (107). How persuasive is a good fit? a comment on theory testing. *Psychological Review*, pages 358–367.

Rohde, D. L. T. (1999). LENS: The light, efficient network simulator. Technical Report CMU-CS-99-164, Pittsburgh, PA: Carnegie Mellon University.

Rose, S. and Walker, R. (2004). A typology of consonant agreement as correspondence. *Language*, 80:475–531.

Rosenfeld, R. (1996). A Maximum Entropy approach to adaptive statistical language modeling. *Computer, Speech and Language 1996*, 10:187– 228. [Long version: Carnegie Mellon Tech. Rep. CMU-CS-94-138].

Rumelhart, D. E., Durbin, R., Golden, R., and Chauvin, Y. (1996). Backpropogation: The basic theory. In Smolensky, P., Mozer, M. C., and Rumelhart, D. E., editors, *Mathematical perspectives on neural networks*, chapter 15, pages 533–566. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propogation. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processsing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, chapter 8, pages 318–362. MIT Press, Cambridge, MA.

Sagey, E. (1986). *The representation of features and relations in nonlinear phonology*. PhD thesis, MIT, Cambridge, MA.

Sagey, E. (1990). *The Representation of Features in Non-Linear Phonology: The Articulator Node Hierarchy*. Outstanding dissertations in linguistics. Garland, New York.

Scholes, R. (1966). *Phonotactic Grammaticality*. Mouton, The Hague.

Sherer, T. (1994). *Prosodic Phonotactics*. PhD thesis, University of Massachusetts, Amherst.

Shosted, R. K. (2006). Just put your lips together and blow? The whistled fricatives of Southern Bantu. In Yehia, H. C., Demolin, D., and Laboissiere, R., editors, *Proceedings of ISSP 2006: 7th International Seminar on Speech Production*, pages 565–572, Belo Horizonte. CEFALA.

Simonoff, J. S. (2003). *Analyzing Categorical Data*. Springer-Verlag, New York.

Smolensky, P. (1986). Information processing in dynamical systems: foundations of Harmony Theory. In Rumelhart, D. E., McClelland, J. L., and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1: Foundations, pages 194–281. MIT

Press/Bradford Books, Cambridge, MA.

Smolensky, P. (1995). On the internal structure of the constraint component Con of UG. Talk presented at UCLA, April.

Smolensky, P. (1996). Overview: statistical perspectives on neural networks. In Smolensky, P., Mozer, M. C., and Rumelhart, D. E., editors, *Mathematical Perspectives on Neural Networksa*, pages 453–496. Lawrence Erlbaum, Mahwah, NJ.

Smolensky, P. and Legendre, G. (2005). *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammars*. MIT Press, Cambridge, MA.

Smolensky, P., Legendre, G., and Tesar, B. (2006). Optimality Theory: The structure, acquisition, and use of grammar. In Smolensky, P. and Legendre, G., editors, *The Harmonic Mind: from Neural Computation To Optimality-Theoretic Grammar*, volume Volume 1. Cognitive Architecture, chapter 12, pages 453–544. MIT Press, Cambridge, MA.

Spall, J. C. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. Wiley, Hoboken, NJ.

Steriade, D. (1995). Underspecification and markedness. In Goldsmith, J., editor, *The Handbook of Phonological Theory*, pages 114–174. Blackwell, Oxford.

Steriade, D. (1999). Alternatives to syllable-based accounts of consonantal phonotactics. In Fujimura, O., Joseph, B., and Palek, B., editors, *Proceedings of the 1998 Linguistics and Phonetics Conference*, pages 205–245. The Karolinum Press, Prague.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions [with discussion]. *Journal of Royal Statistical Society, Series B*, pages 111–147.

Suzuki, K. (1998). *A Typological Investigation of Dissimilation*. PhD thesis, University of Arizona, University of Arizona.

Szeliski, R. (1986). Regularization in neural networks. In Smolensky, P., Mozer, M. C., and Rumelharts, D. E., editors, *Mathematical perspectives on neural networks*, chapter 14, pages 497–532. Lawrence Erlbaum Associates, Mahwah, New Jersey.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and bayesian inference. *Behavioral and Brain Sciences*, 24:629–640.

Tesar, B. and Smolensky, P. (1998). Learnability in Optimality Theory. *Linguistic Inquiry*, 29(2):229–268.

Tesar, B. and Smolensky, P. (2000). *Learnability in Optimality Theory*. MIT Press, Cambridge, MA.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84:327–352.

ud Dowla Khan, S. (2006). Similarity avoidance in East Bengali fixed-segment echo reduplication. In Brainbridge, E. and Agbayani, B., editors, *Proceedings of the thirty-forth Western Conference On Linguistics (WECOL 2006)*, volume 17, pages 257–271, Fresno. Department of Linguistics, California State University, Fresno.

van den Berg, R. (1989). *A Grammar of the Muna Language*. Foris, Dordrecht.

van den Berg, R. and Sidu, L. O. (1996). *Muna-English dictionary*. KITLV Press, Leiden.

Vitanyi, P. and Li, M. (2000). Minimum description length, bayesianism, and kolmogorov complexity. *IEEE Transactions on Information Theory*, 46:446–464.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., and Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, 48:28–50.

Wagenmakers, E.-J. and Waldorp, L. (2006). Special isssue on model selection: Theoretical developments and applications. *Journal of Mathematical Psychology*, 50(2):99–214.

Walter, M. A. (2007). *Repetition Avoidance in Human Language*. PhD thesis, MIT.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer, New York.

Wayment, A. (2009). *Subsymbolic Phonology: Assimilation as Attraction at a Distance*. PhD thesis, Johns Hopkins University.

Wickens, T. D. (1989). *Multiway Contingency Tables Analysis for the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Wright, R. (2004). A review of perceptual cues and cue robustness. In Hayes, B., Kirchner, R., and Steriade, D., editors, *Phonetically-Based Phonology*, pages 346–371. Cambridge University Press, Cambridge.

Wyk, B. C. V. (2006). *Graded Constraints on English Word Forms, Nonword Goodness Ratings, and Durations of Spoken Rimes*. PhD thesis, Carnegie Mellon University.

Yang, T. and Shadlen, M. N. (2007). Probabilistic reasoning by neurons. *Nature*, 447:1075–1080.

Yip, M. (1989). Feature geometry and co-occurrence restrictions. *Phonology*, 6:349–374.

Zeroual, C., Hoole, P., Fuchs, S., and Esling, J. H. (2007). Ema study of the coronal emphatic and non-emphatic plosive consonants of Moroccan Arabic. In *ICPhS XVI*, pages 397–400, Saarbrücken.

Zoll, C. (2004). Positional asymmetries and licensing. In McCarthy, J. J., editor, *Optimality Theory in Phonology: A Reader*, pages 365–378. Blackwell, Malden, MA.

Zuraw, K. (2000). *Patterned Exceptions in Phonology*. PhD thesis, UCLA.

Zuraw, K. (2007). The role of phonetic knowledge in phonological patterning: Corpus and survey evidence from Tagalog. *Language*, 83:277–316.

Zuraw, K. and Lu, Y.-A. (2009). Diverse repairs for multiple labial consonants. *Natural Language & Linguistic Theory*, 27:197–224.