Similarity and phonotactics in Arabic[1]

Stefan Frisch, Michael Broe, and Janet Pierrehumbert[2]

Submitted for publication, September, 1997

Abstract

The statistical patterns of language are systematic linguistic data which must be accounted for in linguistic theory. The phonotactics of the Arabic verbal roots are presented as a case study. In Arabic, the acceptability of a verbal root is gradiently dependent on the similarity of homorganic consonant pairs within the root. We propose the stochastic constraint model of phonological patterning, in which the relative frequency of a phonological form provides a measure of its acceptability. The stochastic constraint model can be parameterized to account for gradient or categorical constraints. We also propose a novel similarity metric for phonological segments based on the representational framework of structured specification. Structured specification provides a basis for a similarity metric for phonemes that is sensitive to featural redundancy and contrastiveness. This metric is superior to previous proposals using underspecification. The account of Arabic consonant cooccurrence using similarity and the stochastic constraint provides a more accurate account of the data than the non-quantitative autosegmental account, demonstrating that a quantitative description is necessary to capture the true pattern of the data.

---

[1] Parts of this paper were presented at the 30th regional meeting of the Chicago Linguistic Society, the XIIIth International Conference of the Phonetic Sciences, the 1995 Summer Institute in Logic and Linguistics, the Ohio State University, UCLA, Indiana University, the Max Planck Institute for Psycholinguistics, the Ecole Normale Superieure, LOT, and the Institute de Linguistique et Phonetique Generale et Applique. We would like to thank those audiences for their comments and questions.

This manuscript was originally circulated as Frisch, Broe, and Pierrehumbert (1995). We thank the many people who provided comments on this early work. We would also like to thank Ken de Jong, Dan Silverman, and Richard Wright for providing comments on this paper.

This work was supported in part by NSF Grant No. BNS 9022484 to Northwestern University, NIH Training Grant No. DC00012 to Indiana University, and a fellowship from the John Simon Guggenheim Memorial Foundation to Janet Pierrehumbert. Janet Pierrehumbert would also like to thank the Département Signal, ENST, Paris, for their hospitality during the 1996-97 academic year.

[2] Stefan Frisch (safrisch@indiana.edu), Speech Research Laboratory, Department of Psychology, Indiana University. Michael Broe (mbroe@nwu.edu) and Janet Pierrehumbert (jbp@nwu.edu), Department of Linguistics, Northwestern University.

0. Introduction

In this paper, we present a new model of the OCP-Place constraint in the verbal roots of Arabic. OCP-Place (or the Obligatory Contour Principle for Place of Articulation) refers to a phonotactic constraint which disfavors combinations of homorganic consonants in proximity to each other. The highly elaborated consonantal system of Arabic provides an opportunity to explore in detail the nature of the OCP-Place constraint and the lessons it provides for the theory of phonology. The most comprehensive previous treatment is a series of papers by McCarthy (1986, 1988, 1994). The present treatment is considerably influenced by that in Pierrehumbert (1993).

The cornerstone of a theory of phonology answers the question, "What is a possible word?". In addressing themselves to this issue, phonologists traditionally recognize a distinction between accidental and systematic gaps in the lexicon. In practice, this distinction has been applied in an intuitive and post-hoc fashion. A gap is taken to be systematic if it belongs to a natural class of examples according to the current theory of the investigator, otherwise it is viewed as accidental. Following Greenberg (1950) and McCarthy (1988, 1994), we treat the existing verbal roots of Arabic as a random sampling of the possible roots, and compare the observed counts of consonantal combinations in the lexicon to the counts which would be expected under the null hypothesis that the consonants are not subject to any constraint (i.e. to the hypothesis that consonants freely combine at random). We propose that phonological theory be held responsible not for individual gaps per se, but rather for the pattern of statistical underrepresentation and overrepresentation as computed over the entire system with respect to a fully articulated null hypothesis. Using this approach, we investigate linguistic patterns with statistical tools that are the standard of evidence in other social sciences.

Statistical phonological data can be interpreted in two very different ways. On the one hand, traditional generative phonological theory is entirely symbolic, and the symbolic grammar is to exclude patterns whose underrepresentation is statistically significant. This is the interpretation of underrepresentation that is carried out formally in McCarthy (1988, 1994). Note that the traditional phonological practice is to exclude those systematic patterns which do not occur at all – though a few exceptions are often ignored. Patterns which do not occur at all are the most extreme case of statistical underrepresentation and are thus an endpoint on the continuum of statistical cooccurrence.

In an alternative perspective, proposed in Pierrehumbert (1994) and used elsewhere (Anttila 1996, Broe 1997, Frisch 1996, Hayes and MacEachern 1996, Pierrehumbert 1993), the grammar has a stochastic component and there is a direct account of the observed statistical patterns. That is, knowledge of the degree of underrepresentation or overrepresentation of a form is imputed to the minds of speakers, and the grammar is responsible for generating the observed probabilities of the various outcomes. McCarthy (1994) acknowledges this point of view and discusses the appropriateness of such a 'soft' or gradient constraint for Arabic, but does not carry the point through formally. In this paper, we argue that the correct model of OCP-Place in Arabic must be quantitative. We provide an integrated treatment of statistical underrepresentation and overrepresentation as the necessary formal tools for a quantitative phonological description.

In a quantitative analysis we may find that different patterns may be underrepresented and overrepresented to different degrees. Our analysis of Arabic reveals that variation in the degree of cooccurrence among violations of OCP-Place is not random. We find that these gradient patterns of cooccurrence provide insight into a variety of theoretical issues. In the remainder of the introduction we preview our most general conclusions to orient the reader with respect to the detailed discussion which follows.

0.1. Transparency

As is well-known, Arabic has a non-concatenative morphology, with the consonants and the vowels in the inflected verb contributed by different morphemes. This situation provides a rationale for segregating the consonants and vowels into separate tiers (McCarthy 1979), as in now famous examples such as <u>kutib</u> 'to be written', in (1).

```
(1)   consonant tier:   k       t       b
                        |       |       |
      skeletal tier:    C   V   C   V   C
                            |       |
      vowel tier:           u       i
```

By this device, consonants which are separated by a vowel on the surface are rendered adjacent at a more abstract level of representation: the consonantal tier. The OCP-Place constraint, which appears on the surface to govern nonadjacent consonants separated by a vowel, can actually apply to adjacent consonants at the more abstract level. A general treatment of phonological transparency follows from this formal device, under the assumption that all constraints can and must be stated on elements which are adjacent in the representation. The hypothesis of autosegmental phonology is that it is possible to use tier segregation to define a representation in which all effects on apparently nonadjacent elements reduce to local effects on some tier.

Below, we summarize and examine the autosegmental treatment of OCP effects in considerable detail. We do this not only because the autosegmental treatment is the most fully developed analysis available to date, but also because its limitations reveal the generic limitations of theories using only symbolic mathematics (logic and/or formal language theory). Use of continuous mathematics (such as probabilities and continuously valued functions) is needed to capture the deep regularities whose existence we demonstrate.

Our statistical analysis of Arabic, presented in sections 1 and 2, shows that the autosegmental analysis of transparency utilizing tier separation is untenable. If all effects are reduced to local adjacency, there can be no treatment of phenomena in which an effect is identified, but where it is also influenced or conditioned by intervening material. The essence of the autosegmental approach is that all effects are expressed as adjacency effects, and by definition, nothing can intervene in such a relation. We show that transparency for OCP-Place is a gradient effect (we might speak of translucency instead). The farther apart two consonants are in the root, the less strongly they are affected by OCP-Place. In other words, we show that OCP-Place applies in a truly non-local manner, and further that it is gradiently influenced by distance.

The broader implication is that constraints in general may apply non-locally, and thus that the autosegmental approach to locality must be abandoned.

0.2 Similarity and underspecification

Pierrehumbert (1993) noted that the strength of the OCP-Place effect for each individual place of articulation in Arabic is related to the variety of phonemes at that place. For example, the large class of coronals has weaker OCP-Place effects than the small class of labials. She claimed that, although the strength of the effect varies with place, this variation need not be specified because it is predictable from more general principles. Pursuing an observation made by Lightner (1973), Pierrehumbert proposed that the strength of the OCP-Place effect depends on the similarity of the consonants involved: highly similar homorganic consonant pairs are found less frequently than dissimilar homorganic consonant pairs. Crucially, variation in the strength of the effect by place of articulation follows from the fact that non-contrastive features contribute less to similarity than contrastive features (Tversky 1977). The weakening of the OCP over distance is also accounted for, as intervening material provides interference to the similarity comparison.

Pierrehumbert (1993) used contrastive underspecification (Steriade 1987), and left non-contrastive features underspecified for the computation of similarity. She claimed underspecified features do not affect similarity. The large class of coronals has many contrastive features, which reduce similarity between coronal consonants. The small class of labials needs very few contrastive features, so similarity between labial consonants is greater.

Underspecification raises a number of empirical and formal problems, discussed in section 2. In particular, we show that underspecified features do influence cooccurrence to some extent. We propose an alternative analysis based on the theory of structured specification (Broe 1993). We show in section 3 that structured specification resolves the formal problems introduced by underspecification while allowing the similarity metric to be conditioned by contrastiveness. Due to its simple and elegant representation of contrast, structured specification is a superior representation not just for the computation of similarity, but for segmental phonology in general.

0.3 Stochastic constraints

We claim that the frequency of a phonological form in the lexicon reflects the acceptability of that form in the grammar. In other words, the acceptability of a form is gradient and the acceptability of a form can be measured by its lexical frequency. Berkley (1994b), Pierrehumbert (in press), and Plenat (1996) argue that accounting for differences in lexical frequency for otherwise equivalent forms is impossible in a non-quantitative phonological formalism. We propose that gradient acceptability be accounted for by gradient phonological constraints. Our model of a gradient constraint is analogous to traditional constraints, which can be grossly characterized as either templates or filters. Traditional templates and filters have discrete, categorical boundaries – a form either fits their description or does not. The difference between gradient and categorical constraints is that gradient constraints have continuous, fuzzy boundaries. The OCP-Place constraint is a filter which prohibits a root which contains repeated homorganic consonants, but the constraint boundary is non-categorical. The degree to which the filter is respected depends on the similarity of the consonants involved.

In our account, we predict the frequency of phonological patterns with a mathematical model of a continuous phonological constraint. We use the logistic function, shown in (A). We explain the logistic function in detail in section 4.

(A) $$y = \frac{1}{1 + e^{K+S \cdot x}}$$

The interpretation of the logistic function as a mathematical model of a stochastic constraint is shown in figure 1. The stochastic constraint reflects the traditional categorical distinction between acceptable (good) and unacceptable (bad) forms at the extremes. The transition from one extreme to the other is continuous however, and forms in between have gradient degrees of acceptability.

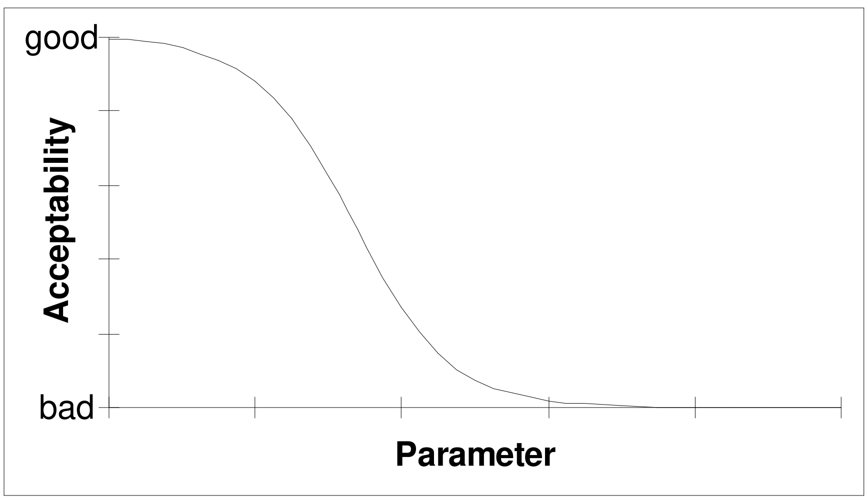

Figure 1: Stochastic constraint model, which predicts the gradient acceptability of a form (reflected in its relative frequency of occurrence) as a function of a linguistic parameter.

In section 4, we introduce the stochastic constraint model, and compare the stochastic constraint model of the Arabic data to the traditional categorical autosegmental model and a soft constraint model which does not employ similarity. We find that the stochastic constraint model provides a more accurate account of the quantitative patterns in the data. We conclude that the OCP-Place effects in Arabic provide strong evidence for gradient constraints which have fuzzy constraint boundaries.

Note that these phonotactic restrictions are part of a native speaker's implicit linguistic knowledge, and not merely artifacts of historical ancestry. OCP-Place effects are reflected in language games (McCarthy 1986), in morphological derivation (Berkley 1994b, Bybee and Slobin 1982, Jespersen 1939, Liberman 1994), and in the interpretation of complex words

(Pierrehumbert 1994). Therefore, gradient constraints should be included in our model of phonological competence as part of the theory of synchronic grammar.

0.4 The OCP constraint family

Cooccurrence constraints reminiscent of OCP-Place in Arabic are also found in many other languages. Due to variation in the exact nature and strength of the effects, it is not possible to say that Universal Grammar has some particular OCP constraint, which always applies in the same way in every language. Rather, Universal Grammar has the abstract potentiality for constraints of this type – an OCP constraint family – and this abstract potentiality can play out differently in different languages. We propose that the OCP constraint family is gradient, and universally based on similarity. The OCP may vary in the dimension of contrast which is targeted (place features, laryngeal features, and tone are attested). In addition, the OCP may vary in the morphological/prosodic domain over which it applies. Cross-linguistic variation in OCP effects and their expression using stochastic constraints is discussed in sections 4 and 5.

The concept of a 'constraint family' also plays a major role in Optimality Theory (Prince & Smolensky 1993). We share with optimality theorists the goal of parameterizing conceptually related constraints in a way which provides insights into language typology as well as the grammars of individual languages. However, the type of constraint families we develop rely crucially on the idea that the attested words constitute a random sampling of the possible forms, with a constraint constituting a bias on the random sampling. This concept of a constraint is not available in Optimality Theory, as we discuss in section 5.

1. The major cooccurrence classes of Arabic

We begin our detailed analysis of the Arabic verbal root morphemes with a review of the original quantitative description given by Greenberg (1950) and the autosegmental account drawn primarily from McCarthy (1986, 1988, 1994). Arabic verbal roots consist of a set of two to four consonants, with the canonical root containing three. Vowels are inserted between the consonants to make word forms, an example of a non-concatenative morphological system. For example, the root /k t b/ has among its word forms <u>katab</u> 'to write' and <u>kutib</u> 'to be written'. In addition to non-concatenative vowel insertion, concatenative morphemes are used for many morphological processes. For example, <u>katab+a</u> 'he wrote' and <u>ma+katiib</u> 'letters'. The Arabic consonant inventory and the feature assignments we use are presented in the appendix.

Greenberg (1950) studied the overall statistical patterning of the triconsonantal verb root morphemes in Arabic, based on Lane's (1863) dictionary of Arabic. The cooccurrence restrictions he describes apply only to the roots themselves, and not to any derived forms. Thus, the restrictions do not apply across concatenative morpheme boundaries, but only to the consonants within the root. Greenberg begins his analysis with the observation that there are no roots which repeat the same consonant in first and second position. Thus, roots like *<u>dadam</u> do not occur. Many verbs are found with identical consonants in the second and third positions of the root. Examples include <u>madad</u> 'to stretch' and <u>farar</u> 'to flee'.

Greenberg shows that, more generally, Arabic consonants divide into 'sections' of homorganic consonants that tend not to cooccur within the same root, apart from the pairs of identical consonants in the second and third position just mentioned. McCarthy (1988, 1994) replicates Greenberg's study, but characterizes the cooccurrence classes in terms of the combination of place of articulation and the major manner feature [sonorant].

The traditional approach to these cooccurrence restrictions is to divide the Arabic consonants into natural classes, with cooccurrence constraints applying within these classes. The major cooccurrence classes discussed by Greenberg and McCarthy are presented in (2). In their analyses, consonants in any one of these classes are claimed to cooccur freely with consonants from any other class, and within any class consonants tend not to cooccur, with two exceptions. First, the velars (2c) cannot cooccur with the uvular approximants {χ, ʁ}, though they can cooccur with the other gutturals. Second, among the coronal obstruents, there are far more roots containing one fricative and one stop than roots containing two fricatives or two stops.

<u>Major cooccurrence classes</u>:

(2) a. Labials = {b, f, m}
 b. Coronal Obstruents = {t, d, T, D, θ, ð, s, z, S, Z, ʃ}
 c. Velars = {k, g, q}
 d. Gutturals = {χ, ʁ, ħ, ʕ, h, ʔ}
 e. Coronal Sonorants = {l, r, n}

Both Greenberg and McCarthy also note that cooccurrence between non-adjacent consonant pairs (in the first and third position of a triliteral root) is less restricted than cooccurrence between adjacent consonant pairs. McCarthy chose to formalize the OCP so that the constraint applies equally to both adjacent and non-adjacent consonants despite this difference. These early findings of gradient differences in the degree of cooccurrence of consonant pairs are problematic for an account based on autosegmental phonology, as we discuss below. It is not clear in the traditional categorical approach what degree of underrepresentation is considered sufficient to warrant positing a constraint. There are many other sub-regularities not captured by these classifications which are equally problematic. We show in section 2 that any categorical classification is insufficient to describe the true pattern of cooccurrence in Arabic. In the next section, we present our own replication of the results of Greenberg and McCarthy, as the foundation for the more detailed analysis to follow.

1.1. Preliminary quantitative description

Our analysis is based on the 2674 native and assimilated triliteral roots from Wehr's 1979 dictionary of modern Arabic (Cowan, 1979). This is the same dictionary used in Pierrehumbert (1993), which is a later edition of the dictionary used in McCarthy's (1986, 1988, 1994) statistical analyses. Following McCarthy, roots with repeated second and third consonants are excluded because they are conceptually biliteral, as discussed below. In addition, roots with four surface consonants are excluded because almost all of them are either reduplications of the form $C_1C_2C_1C_2$, or unassimilated borrowings.

In order to quantify overrepresentation and underrepresentation in the lexicon, we follow Pierrehumbert's (1993) use of the ratio of the observed number of occurring pairs of a type (O) to the number that would be expected if consonants combined at random (E). A value of O/E less than 1 indicates that there are fewer observed tokens than would be expected if consonants combined at random – that is, there is a cooccurrence restriction affecting the consonants of the pair. If O/E is greater than 1, the number of pairs observed is greater than the number expected. If O/E equals 1, the number of observed pairs is exactly what is expected by random cooccurrence.

For example, Arabic triconsonantal roots of the form /d t C/ (where C is any consonant) are not found. Given the frequency of roots beginning in /d/ and the frequency of roots with /t/ in second position, one would expect 2.3 such roots if consonants combined at random. In this case, O/E = 0, the strongest degree of underrepresentation. There are 2 roots containing /d s C/, and 2.9 are expected, giving an O/E of 0.69 (underrepresentation). There are 4 roots with /d g C/ and 3.3 expected at random, giving an O/E of 1.21 (overrepresentation).

Given the large number of Arabic consonants, a full matrix of pairwise combinations has 784 cells. However, there are only 2674 different verbal roots. This means that the data are very sparse from a statistical point of view, and a high or low O/E in any individual cell could easily be a consequence of random variation. For example, due to the use of E in the denominator, the measure of O/E is particularly unstable for low frequency consonants. The problem is conceptually equivalent to the traditional notion of an accidental gap. In this paper we base our arguments on aggregate data (for which the observed and expected counts are larger) and on patterns of over or underrepresentation across natural classes. Thus, while our approach is quantitative, we examine natural classes of examples to discover the patterns of cooccurrence much like a traditional phonological analysis.

With these preliminaries, we can now present table 1, which shows the O/E for aggregated sets of consonants in adjacent (C_1C_2 or C_2C_3) and non-adjacent (C_1C_3) position in the root. Observed and expected counts were computed for individual consonant pairs, as in the examples above, and then aggregated in three ways. First, pairs in first and second position and in second and third position have been collapsed into a single set of adjacent consonant pairs. For example, for the root /d s w/, the pairs /d, s/ and /s, w/ are adjacent pairs. The pair /d, w/ is a non-adjacent pair. In addition, the order of consonants within the pairs have been collapsed, so that the consonant pair /d, s/ is treated identically to the pair /s, d/. Finally, the data have been grouped to highlight the cooccurrence restrictions between the major classes and subclasses noted by Greenberg and McCarthy. As in their analyses, we do not examine the cooccurrence patterns of /w/ and /y/, due to uncertainties about their phonological status.

The posited cooccurrence restrictions within the major classes, shaded in gray, can be seen in the low O/E values in table 1. For these classes, the number of observed consonant pairs is far below that expected by chance, indicating that roots containing pairs within these classes are rare. Note, though, that there is quantitative variation in cooccurrence levels in different classes. As noted above, there are strong effects within the coronal stops and fricatives (O/E = 0.14 and 0.04, respectively for adjacent pairs), as compared to the effects between the stops and fricatives (O/E = 0.53). Comparing the adjacent to the non-adjacent values within the major classes shows that the cooccurrence constraint is weaker for non-adjacent pairs. Non-adjacent pairs are still underrepresented, but the degree of underrepresentation is less than for adjacent pairs. We next present the standard analysis of OCP-Place which employs autosegmental phonology. This account can capture only some of the generalizations we have discussed thus far.

		Adjacent						
		Labial	Cor obs		Velar	Guttural		Cor son
		b f m	t d T D	θ ð s z S Z	k g q	χ ʁ	ħ ʕ h ʔ	l r n
Labial	b f m	0.00	1.37	1.31	1.15	1.35	1.17	1.18
Cor obs	t d T D		0.14	0.53	0.80	1.43	1.25	1.23
	θ ð s z S Z			0.04	1.16	1.41	1.26	1.21
Velar	k g q				0.02	0.07	1.04	1.48
Guttural	χ ʁ					0.00	0.07	1.39
	ħ ʕ h ʔ						0.06	1.26
Cor son	l r n							0.06
		Nonadjacent						
		Labial	Cor stop		Velar	Guttural		Cor son
		b f m	t d T D	θ ð s z S Z	k g q	χ ʁ	ħ ʕ h ʔ	l r n
Labial	b f m	0.30	1.08	1.02	1.26	1.25	1.28	1.11
Cor obs	t d T D		0.38	1.06	1.24	1.05	1.02	0.97
	θ ð s z S Z			0.24	1.16	1.35	1.14	1.23
Velar	k g q				0.07	0.68	1.19	1.03
Guttural	χ ʁ					0.25	0.12	1.10
	ħ ʕ h ʔ						0.34	1.13
Cor son	l r n							0.67

Table 1: Cooccurrence of consonant pairs in Arabic, aggregated by major class and distance. Major classes are shaded.

1.2. The autosegmental account

McCarthy (1986, 1988, 1994) and others (Mester 1986, Padgett 1995c, Yip 1989) formalize the cooccurrence restrictions in the Arabic verbal roots using the notation of autosegmental phonology. Autosegmental analyses have attempted to account for two characteristics of the Arabic data. First, McCarthy (1986) proposed an account of the distribution of identical consonant pairs, e.g. madad, but *dadam. Later work (e.g. McCarthy 1994) has tried to account for the underrepresentation of roots with consonant pairs in the major classes in any position.

McCarthy (1979) showed that the non-concatenative morphology of Arabic can be represented by separating the vowels and consonants of the word form onto different tiers. Thus, a typical verb is represented as in (3).

(3)	vowel tier:		u		i		
			|		|		
	skeletal tier:	C	V	C	V	C	
		|		|		|	
	consonant tier:	k		t		b	

In (3), there are three different tiers, each of which contributes to the meaning of the word form. The consonants /k t b/ 'write', the vowels /u i/, which indicates the passive, and the skeletal pattern CVCVC which indicates the infinitive. These three morphological components combine to make the complete form. If one pattern is changed, the resulting word changes. For example, we can change the consonants in (3) to /f ʕ l/, resulting in fuʕil 'to be done'; we can change the vowels to /a a/, resulting in katab 'to write'; or we can change the CV-skeleton to CVVCVC, resulting in kuutib 'to be corresponded with'.

The autosegmental analysis accounts for the cooccurrence restrictions in Arabic by application of the OCP, stated in (4).

(4) OCP: Adjacent identical elements are prohibited

The OCP was applied in McCarthy (1986) to autosegmental representations in which feature bundles for vowels and consonants were connected to C and V nodes on different tiers, as in (2). The OCP was applied to the consonant tier to rule out roots like *dadam with adjacent identical consonants (the 'total OCP').

While Arabic does not allow roots where the first two consonants are identical, like */d d m/, it does permit the second pair of consonants to be identical, as in madad. McCarthy claims the underlying form of madad is /m d/ with only two consonants. When the consonants /m d/ are to be associated with a CVCVC tier, there are two consonants and three C slots, so a one-to-many mapping results. Association in Arabic proceeds from left to right. The resulting representation is given in (5).

(5)

```
C V C V C
|    \ /
m     d
```

Vowels, which are on a different morpheme plane, are inserted normally, giving surface forms where there appear to be two separate /d/s. With left to right association and the OCP, it is impossible to derive a surface form like *dadam. McCarthy thus claims that the idiosyncratic triliteral roots with repeated second and third consonants are conceptually biliteral at the relevant level of abstraction, and do not violate the OCP.

Autosegmental representations became more intricate with the introduction of feature geometry (Clements 1985). In feature geometry, the features of a segment are hierarchically organized into a tree structure, and grouped into place features, laryngeal features, and so on. McCarthy (1988) applied the OCP to individual place feature tiers to rule out roots containing homorganic consonants in any position, as an account of the statistical underrepresentation in the major classes ('OCP-Place'). In this analysis, each privative place feature occupies its own tier, and the OCP applies to each place feature tier. This analysis rules out homorganic consonant pairs whether they are adjacent or not, as their place features are adjacent on the place tier. For example, a hypothetical root like */f t b/ is represented as in (6), with irrelevant feature tiers excluded for clarity. Adjacent identical features on the labial tier violate OCP-Place, marking the structure as ill-formed.

(6)

```
labial tier:    [lab]         [lab]
                  |             |
skeletal tier:   *C      C      C
                         |
coronal tier:          [cor]
```

The analysis of OCP-Place effects requires an additional elaboration: a constraint against one-to-many mappings of individual place features. With this prohibition, a representation like (7), which would circumvent OCP-Place, is disallowed.

(7)

```
labial tier:          [lab]
                      /   \
skeletal tier:   *C     C     C
                        |
coronal tier:         [cor]
```

McCarthy (1988, 1994) assigns primary place of articulation features as in (8) in his analysis of the cooccurrence classes. He accounts for the dual patterning of /χ/ and /ʁ/ with the velars and the gutturals by using two place of articulation specifications for /χ/ and /ʁ/. In addition, the split of the coronals into two major classes is by special reference to the feature [sonorant]. McCarthy (1994) notes that the split among the coronal obstruents could also be accommodated by reference to the feature [continuant], following the analysis of Padgett (1995c). Note, however, that these proposals treat differences in degree of cooccurrence identically. Even though the split between the coronal obstruents and sonorants is much stronger than the split between the stops and fricatives, they are accounted for with the same formal device.

(8) [labial] = {b, f, m}
[coronal] = {t, d, T, D, θ, ð, s, z, S, Z, ʃ, l, r, n}
[dorsal] = {k, g, q, χ, ʁ}
[pharyngeal] = {χ, ʁ, ħ, ʕ, h, ʔ}

Pierrehumbert (1993) pointed out that the separation of individual features onto distinct tiers used in the analysis of OCP-Place creates problems for the original analysis of the total OCP in McCarthy (1986). If we attempt to formalize the total OCP in autosegmental phonology given the tier separation of features required for OCP-Place, our only option is to treat the total OCP as an application of the OCP on all feature tiers simultaneously. Thus a root violates the total OCP only if two segments have adjacent feature specifications on all tiers. Obviously, the root */d d m/ violates the total OCP. The root */f d f/ does not violate the total OCP. Consider the partial representation of */f d f/ in (9).

(9)

obstruent tier:	[obs]	[obs]	[obs]
skeletal tier:	*C	C	C
coronal tier:		[cor]	
labial tier:	[lab]		[lab]

There is intervening material on the obstruent tier, so the [obs] features for the two /f/s are not adjacent and the total OCP is not violated. This root does violate OCP-Place, as the labial features are adjacent.

Formally, OCP-Place subsumes the total OCP for adjacent consonants, as identical consonants are homorganic. However, McCarthy (1988, 1994) maintains the distinction between the total OCP and OCP-Place because the total OCP is stronger than OCP-Place. Adjacent identical consonants are prohibited. Roots with homorganic consonants do occur but they are highly underrepresented. The difference between these two constraints is one of degree. McCarthy (1994) notes the distinction, but provides no formal account of the difference. We have seen other differences in the degree of underrepresentation within the coronals and between adjacent and non-adjacent pairs. In each case, the autosegmental account must decide whether the degree of effect is strong enough to warrant inclusion in the set of cooccurrence classes. These differences in degree motivate our reanalysis of the data and proposals for a quantitative formal account.

2. Beyond the autosegmental account

There is a clear sense in which the OCP effect is cumulative, with total OCP being the limiting case of the OCP-Place constraint. This insight is present in Greenberg's (1950) discussion, but is totally absent from the autosegmental account. Non-place features do have some role to play in the OCP effect: the total OCP and the division of the coronals by manner are evidence of this. In this section, we identify many other subpatterns where non-place features play a role in determining cooccurrence.

Many of the problems with the autosegmental account of OCP-Place were originally pointed out in Pierrehumbert (1993). She showed that the autosegmental model makes incorrect predictions on the effects of distance for the total OCP and OCP-Place. She also demonstrated widespread patterns of sub-classification within homorganic classes which are not accounted for by the autosegmental model. She proposed that an account based on similarity captures the effects of distance and subclassification.

We present additional patterns of cross-classification between the major classes. The autosegmental account proposes four major place classifications, of which only the coronals are subdivided by manner. The patterns of cross-classification show that there are three major place classes in Arabic, each of which is subdivided by manner to some degree. We also present an effect of voicing on cooccurrence between sonorants and obstruents which further demonstrates the advantages of a similarity based account. The effect of voicing provides important insight to the featural representation of the consonants in Arabic as voicing features are often underspecified for sonorant consonants.

In this section, we review each of these additional pieces of data in detail. Together, these data reveal the thoroughly gradient nature of the cooccurrence restriction, providing strong support for a quantitative model.

2.1. Distance effects

Pierrehumbert showed that McCarthy's account makes very specific predictions about the effects of distance in the Arabic data. For non-adjacent consonants, it is very likely that the intervening consonant has at least one feature in common with the flanking consonants which thus blocks the total OCP on that tier (see (9) above). The autosegmental model predicts that non-adjacent identical consonants should pattern with other homorganic consonants under the OCP-Place restriction. Pierrehumbert (1993) also noted that the autosegmental account predicts that distance should have no effect on OCP-Place. Non-adjacent segments with identical place specifications are adjacent on the place tier, where OCP-Place is enforced. Thus, non-adjacent homorganic consonants are subject to the exact same restriction as adjacent homorganic consonants, as intervening consonants with different place specifications are transparent to OCP-Place. In summary, there should be equal underrepresentation for non-adjacent identical consonants, adjacent homorganic consonants, and non-adjacent homorganic consonants, while adjacent identical consonants are prohibited.

We can test these predictions by looking at the degree of cooccurrence (O/E) of identical and homorganic consonant pairs separately in both adjacent and non-adjacent position. Table 2 shows the distribution of consonant pairs in Arabic divided into these four classes. The left side of table 2 shows aggregated O/E for adjacent consonant pairs in the major classes. The right side shows O/E for non-adjacent consonant pairs. For each distance, the data are divided into identical pairs and non-identical homorganic pairs.

	Adjacent		Non-adjacent	
Major Class	Identical	Homorganic, non-identical	Identical	Homorganic, non-identical
Labials	0.00	0.00	0.05	0.42
Coronal Obs	0.03	0.33	0.46	0.70
Velars and χ,ʁ	0.00	0.05	0.24	0.36
Gutturals	0.00	0.08	0.05	0.43
Coronal Son	0.00	0.09	0.08	0.94

Table 2: Cooccurrence of consonant pairs in the major classes, for adjacent and non-adjacent pairs of identical and homorganic non-identical consonants.

For non-adjacent consonants in all major classes, O/E for identical pairs is less than the O/E for non-identical pairs. In other words, non-adjacent identical pairs cooccur relatively less frequently than non-adjacent, non-identical, homorganic pairs. In addition, the rate of cooccurrence is consistently less for adjacent consonant pairs than for equivalent non-adjacent consonant pairs. O/E for adjacent pairs is less than O/E for non-adjacent pairs for both identical pairs and non-identical homorganic pairs. Together, these facts show that the total OCP has an effect on non-adjacent identical pairs, and that distance has a gradient effect on both the total OCP and OCP-Place. In other words, transparency to the OCP is gradient and based on superficial proximity contrary to the predictions of the autosegmental account.

2.2. Sub-classification

There are also significant sub-patterns of cooccurrence within the major classes. Greenberg originally pointed out that the coronal obstruents actually break into two classes, the coronal stops and coronal fricatives, as seen above. There are several other cases of sub-classification. For example, Pierrehumbert (1993) reports that the emphatic coronals /T/, /D/, /S/, and /Z/ have a stronger cooccurrence restriction with each other than they do with the other coronal obstruents. Pierrehumbert also reports that /l/ and /r/ form a subclass of the coronal sonorants, which have stronger cooccurrence restrictions with each other than they do with /n/.

We have already seen that the cooccurrence of coronal consonants depends on their manner of articulation. Pierrehumbert reports that the labial class also shows some evidence of sub-classification by manner. While there are no observed pairs of labial consonants in adjacent position, there are 17 such pairs in non-adjacent position. Of those, 16 involve /m/ with a labial obstruent (/b/ or /f/). There is only one pair with two obstruents, indicating manner does have an effect for labials. The effect of manner for labials is much weaker than the effect of manner for coronals. The coronal split by manner is observed even for adjacent consonant pairs, for which the constraint is stronger. The difference in the degree to which manner has an effect on cooccurrence, and the interaction of the manner effect with distance, is one of the most challenging aspects of the Arabic data.

All of these cases of sub-classification are unexplained in the autosegmental account. Further, all of these sub-classifications are not just partitions of the data into groups which can and cannot cooccur. Rather, the differences in cooccurrence within and between sub-classes of the major classes are differences in degree. The autosegmental analysis used two separate constraints to account for the difference in degree of cooccurrence between identical and homorganic but non-identical consonant pairs. In order to account for these additional cases of sub-classification, as well as the cases of cross-classification in the next section, the autosegmental analysis requires an explosion of seemingly unrelated sub-constraints. Pierrehumbert (1993) proposed that all of the cooccurrence restrictions can be captured by a single constraint, OCP-Place, where the degree of cooccurrence restriction depends on the perceived similarity between homorganic consonants. Pierrehumbert claimed that perceived similarity is a function of the size of the inventory at each place of articulation and the distance between consonants. Pierrehumbert's account not only predicts variation in the degree of cooccurrence, but also predicts that the variation follows from more general principles.

2.3 Cross-classification

In addition to the sub-classification within the previously reported major classes, we have discovered a number of regularities outside of the previously noted cooccurrence classes. These patterns of cross-classification show that the Arabic consonants are divided into three major cooccurrence classes: [labial], [coronal], and [dorso-guttural], rather than the four place of articulation classes proposed in McCarthy (1994). Each of these classes is sub-divided by manner. More generally, the analysis shows that reference to non-place features is not just an ad hoc requirement for handling subregularities among the coronals, but has a systematic conditioning effect for all places of articulation.

Given that there are strong cooccurrence restrictions within the major classes, other consonant combinations outside of the major classes are necessarily statistically overrepresented. The quantitative approach permits us not only to examine the pattern of underrepresentation among relatively ill-formed roots, but also the pattern of overrepresentation among roots that are relatively well-formed. Over the entire lexicon, the total number of expected consonant pairs is the same as the actual number of consonant pairs (5348 adjacent pairs and 2674 non-adjacent pairs). If some pair occurs far less than expected, other pairs must occur more frequently than expected for the totals to balance. Under the autosegmental analysis, we would expect unrestricted consonants to combine freely, so overrepresentation should be randomly distributed among unrestricted pairs. In this section we present patterns of overrepresentation inconsistent with the autosegmental account.

We can exhibit the structure of overrepresentation using a similar measure to that already employed, O/E. This time, however, we want to compare the ratio of observed pairs to the 'expected overrepresentation' of the pair (O/E_{Ov}). Expected overrepresentation is the degree of overrepresentation we would expect if the increased cooccurrence among unrestricted pairs were distributed randomly among the more commonly occurring consonant pairs. We construct this theoretical random overrepresentation by distributing the excess expected counts from within the major classes (the amount of expected beyond the actual cooccurrence) among cells outside of the major classes.

For example, there are 135 expected adjacent labial-labial pairs, but no actual pairs. In this case, 135 excess expected tokens are distributed among all pairs involving labials outside of the major classes. The excess expected were distributed among the corresponding cells outside of the major classes in proportion to the expected frequency of those cells. Distributing excess expected frequency in this way preserves the overall phoneme frequencies of the consonants so that individual expected phoneme frequencies are correct. For example, there are 178 expected adjacent pairs of a coronal stop with a labial, and 332 expected pairs of a coronal sonorant with a labial. Among the 135 excess expected labial-labial pairs, roughly twice as many are added to the expected overrepresentation of coronal sonorant-labial pairs (37 excess + 332 expected = 369 expected overrepresentation) as coronal stop-labial pairs (19 excess + 178 expected = 197 expected overrepresentation). The remaining 79 pairs were distributed to the coronal fricative-labial pairs (27 excess), velar-labial pairs (19 excess), {χ, ʁ}-labial pairs (8 excess), and {ħ, ʕ, h, ʔ}-labial pairs (25 excess).

		Labial	Cor obs		Velar	Guttural		Cor son
		b f m	t d T D	θ ð s z S Z	k g q	χ ʁ	ħ ʕ h ʔ	l r n
Labial	b f m	-	1.09	1.05	0.94	1.01	0.95	0.95
Cor obs	t d T D		-	-	0.65	1.06	1.01	0.96
	θ ð s z S Z			-	0.95	1.06	1.00	0.95
Velar	k g q				-	-	0.86	1.20
Guttural	χ ʁ					-	-	1.03
	ħ ʕ h ʔ						-	1.01
Cor son	l r n							-

Table 3: Cooccurrence of adjacent pairs outside the major classes, with expected overrepresentation adjusted for major class effects (see text).
Significant deviations from expected overrepresentation are shaded.

Table 3 shows the ratio of observed to expected overrepresentation for adjacent consonant pairs outside of the major classes. Note that in general, O/E_{Ov} is close to 1 for most pairs outside of the major classes. There are three cells in table 3, indicated by gray shading, which have observed counts significantly different from expected overrepresentation ($\underline{p} < 0.05$ on χ^2-test with 1 df).

Notice first that all of these cells involve the velars, /k/, /g/, and /q/. The cooccurrence of the velars and the guttural approximants /ħ/, /ʕ/, /h/, and /ʔ/ is significantly less than the expected amount of overrepresentation ($O/E_{Ov} = 0.86$). While pairs of consonants in these two classes are not underrepresented (153 observed and 147 unadjusted expected, O/E = 1.04), they do show far less overrepresentation than they should according to the autosegmental account. Pairs of consonants from these classes do not share place of articulation in McCarthy's (1994) account. We propose that the velars and the guttural approximant classes share the [dorso-guttural] place feature, and that the division between the velars and the gutturals as major OCP classes is based on manner of articulation. This is analogous to the division of the coronals by manner into coronal obstruents and coronal sonorants, but the split is weaker for the dorso-gutturals.

Table 3 also reveals an effect of secondary place of articulation on consonant cooccurrence. The cooccurrence of the velars with the coronal stops is significantly different from the expected overrepresentation ($O/E_{Ov} = 0.65$). The aggregate data are misleading in this case, however, as the cooccurrence restriction is limited to the emphatic stops /T/ and /D/. The emphatics are actually underrepresented in combination with the velars. The unadjusted O/E for /k/, /g/, and /q/ with /T/ and /D/ is 0.53; while for /k/, /g/, and /q/ with /t/ and /d/, O/E is 1.00. There is an analogous but lesser effect for /S/ and /Z/, which also have secondary [dorso-guttural] articulation. O/E for /k/, /g/, and /q/ with /S/ and /Z/ is 0.77; the O/E for /k/, /g/, and /q/ with the other coronal fricatives is 1.25.

The underrepresentation of emphatics with the velars is statistically significant. We compared corresponding pairs involving emphatic and non-emphatic coronals, like /k, t/ with /k, T/, /g, s/ with /g, S/, and so forth. There are 24 such pairs, and no reason to assume that the O/E values are normally distributed, so a sign test is used. Emphatic coronals are more underrepresented with velars than corresponding non-emphatic coronals ($\underline{n} = 24$ comparisons, 19 have smaller O/E for emphatics, 5 have smaller O/E for non-emphatic coronals, $\underline{p} < 0.05$).

We claim the cross-classification of velars and emphatics is due to the secondary dorso-guttural articulation for the emphatic consonants. Our analysis is consistent with a number of X-ray studies of the emphatic coronals, reviewed in McCarthy (1994), which show that the emphatics have a constriction in the uvular region. Note that McCarthy himself exempts secondary articulations from the OCP, because otherwise the cooccurrence constraint would be far too strong. For example, roots such as /T g m/ would have to be underrepresented to the same degree as the major classes. That is, he had a choice of getting no effect, or too strong of an effect. Using similarity, we can get an intermediate effect to account for the patterns of cross-classification.

The velars have cooccurrence restrictions with each other, the gutturals, and the emphatic coronals, and thus have significant underrepresentation with half of the Arabic consonant inventory. Not surprisingly, the velars are highly overrepresented with the other consonants. In fact, table 3 shows there is a greater than expected overrepresentation between the velars and the coronal sonorants ($O/E_{Ov} = 1.20$). The overrepresentation seen here is the result of the large number of cooccurrence restrictions that the velars have with other consonants.

2.4. Redundant voicing

In addition to the cross-classification between the major classes, we have found an effect of voicing on the cooccurrence of coronal obstruents and coronal sonorants. Sonorants are phonetically voiced, though voicing is phonologically non-contrastive. An account based on similarity predicts that voiceless obstruents should cooccur with sonorants more than voiced obstruents do. This prediction is supported by the data. The aggregate O/E is 1.15 for coronal sonorants and voiced coronal obstruents (239 actual and 207 expected). The aggregate O/E is 1.31 for coronal sonorants with voiceless obstruents (245 actual and 187 expected). There is less cooccurrence of sonorants with voiced obstruents than sonorants with voiceless obstruents, indicating that shared voicing has an effect on cooccurrence.

The effect of redundant voicing is subtle, but statistically significant. Coronal obstruents which contrast only by voicing can be paired with coronal sonorants, and comparisons made between pairs. For example, we can compare /t, n/ with /d, n/, /s, l/ with /z, l/, and so on. There are $\underline{n} = 30$ pairs, no reason to assume a normal distribution, and the sign test is not sensitive enough to be used in this case, so we use the Wilcoxon Signed-Rank test (Devore 1987; Signed-rank sum, $\underline{S}_+ = 305$, $\underline{p} < 0.07$).

The correct model of cooccurrence restrictions in Arabic must include the effects of redundant voicing. In particular, since the non-contrastive voicing feature is often assumed to be underspecified for sonorants, the effect of redundant voicing provides evidence against the underspecification of redundant features in the phonotactic constraints of Arabic.

2.5. Interim summary

We have shown that the degree of cooccurrence restriction in the Arabic verbal roots depends on place of articulation, manner, and voicing features, as well as the distance between consonant pairs. The standard account used separate rules for what were seen as two different effects, one on identical consonants (total OCP) and one on homorganic consonants (OCP-Place). OCP-Place was known to be manner sensitive for coronals, for which a specific sub-rule was proposed. We have shown the effects of manner are much more widespread than previously noted, as some manner effects are found within every place class. However, the degree of a manner effect varies from class to class. The strongest manner effect is found for coronals, as coronal obstruents and sonorants frequently cooccur. A weaker manner effect was found for dorso-gutturals. Dorso-gutturals which differed in manner are not underrepresented, but they had less than the expected degree of overrepresentation. Finally, the labial class had a very weak effect of manner, which was only in evidence in non-adjacent consonant pairs. Each of these cases would require an additional sub-rule in the autosegmental account. Additional sub-cases are required for voicing features as well, resulting in an explosion of seemingly unrelated sub-cases, and no a priori predictions of the strength of the effect for each of these cases. An alternative account which could incorporate all of the sub-cases under a single generalization would be preferred. Following Pierrehumbert (1993), we propose that similarity accounts for the gradient strength of OCP effects and all of the variation we have discussed.

In this section, we have also shown that any model of the OCP-Place effects in Arabic must be fundamentally quantitative in nature. A model based only on categorical restrictions cannot account for any of the gradient variation in cooccurrence patterns. Our account uses a novel model of a gradient linguistic constraint, which predicts cooccurrence based on similarity. In the next section, we present a general metric of similarity for phonemes which we use in our account. In section 4, we return to the phonotactics of Arabic and present a quantitative constraint model which accounts for the gradient cooccurrence patterns.

3. The similarity account

We propose that the OCP effects in Arabic can be accounted for through a single gradient constraint. There is a cooccurrence restriction on homorganic consonant pairs in the Arabic roots in proportion to their perceived similarity. Identical consonants are maximally similar, and have the strongest cooccurrence constraint. Consonants which differ in many features but are still homorganic are subject to weak cooccurrence restrictions. Similar non-adjacent consonants are more difficult to detect than similar adjacent consonants, so the constraint is also weaker for non-adjacent consonants (Pierrehumbert 1993).

In order to make a formal model of the cooccurrence constraints of Arabic, it is necessary to have a general method for computing the similarity between any two consonants. In this section, we propose a novel similarity metric for phonemes. It is standard practice to refer to distinctive features when discussing the similarity of phonemes (e.g. van den Broeke and Goldstein 1980). In general, previous research related to phonemic similarity has made ad hoc assumptions about the similarity of phonemes. For example, van Oijen (1994) assumes that phoneme pairs differing in a single feature are all equally similar, without specifying how the similarity of phonemes may in general be computed. Stemberger (1991) argues that speech error data provide evidence for radical underspecification, but does not develop a general similarity metric. In addition, Frisch's (1996) reanalysis of Stemberger's data shows that the approach used here provides a superior model of the effect of similarity on error rates.

The psychological literature provides some indication of how a similarity metric should behave, but it provides no general solution to the problem. In this literature, the attributes of categories are manipulated orthogonally in controlled experiments, so that the results of the experiments can be analyzed using statistical techniques such as analysis of variance. However, phonemic systems do not display orthogonal combinations of features, because the possible combinations are limited by facts of physics and biology, as well as by the historical development of the language. We would suggest that phonemic systems are in this respect typical of natural category systems. Thus, our proposal for computing similarity over nonorthogonal features promises to have applications outside of the domain of phonology.

A similarity metric based on distinctive features presupposes an adequate system of feature specification. In this section, we first discuss some desirable properties for a similarity metric for phonemes, based on a survey of the cognitive psychology literature. We then evaluate the primary theories of feature specification (full specification, contrastive underspecification, and radical underspecification) for their suitability and conclude that none of these theories provide an adequate basis for computing similarity. We then introduce the theory of structured specification and a similarity metric based on it which satisfies all of our requirements.

3.1. Properties of a similarity metric

Our similarity metric meets the requirements for generality and mathematical coherence which are laid out in Tversky (1977). In addition, we have identified three further properties which a metric should meet in order to provide a well-behaved characterization of natural category systems, and our treatment has all of these properties. They are:

i. Stability under relabeling: The similarity of two objects should depend on the intrinsic properties of these objects, and not on arbitrary decisions about how to label those properties (Goodman 1972).

For example, as pointed out in Jakobson, Fant, and Halle (1952), a bivalent opposition such as [±voice] can also be recast in terms of two monovalent features [+voice] and [+voiceless]. Such a relabeling of the features should not affect the calculated similarity of two phonemes, since their intrinsic nature is unaffected by the relabeling.

Analogously, the description of a phoneme may often be elaborated through the addition of redundant features. For example, in a language in which all voiceless stops are phonetically tense and all voiced stops are phonetically lax, one might view /t/ and /d/ as differing either as [+voiceless], [+voice], or as differing in [+voiceless, +tense], [+voice, + lax]. The addition of such descriptive detail has no effect on the phonological system, however. Thus, we hold that the similarity metric should remain stable under the addition of fully redundant features.

ii. Relativizing contrast to a sub-universe: The influence of a feature on similarity depends on whether the feature contrasts objects or permits objects to be identified in the universe of discourse (Tversky 1977).

The perception of similarity is context sensitive, depending on the universe of discourse in which the similarity comparison is made. For example, at a family reunion, siblings are regarded as very different by members of the family. If those siblings take a trip together to a distant country with a different culture, everyone would find the siblings to be highly similar. At the family reunion, the traits shared by all of the family members are found throughout the universe of discourse and do not distinguish individuals. Hence they have little effect on similarity. In the foreign culture, these shared traits permit the siblings to be distinguished from other individuals, and as a result they strongly influence the perceived similarity of the siblings.

The effects of context on similarity are found in the cooccurrence patterns of Arabic (Pierrehumbert 1993). For example, in the set of labial obstruents in Arabic, [+cont] and [+voice] are not independently contrastive. While /b/ and /f/ exist, there is no /p/ or /v/. However, these features are independently contrastive for the larger set of coronal consonants. Since the coronals have more contrastive features, they should have less similarity to one another than the labials.

We saw above that the coronal stops and fricatives do have greater cooccurrence than other pairs in the major classes. We claim this pattern is predicted by effect of context on similarity.

iii. Synergistic combination: The increase in similarity with features in common is faster than linear (Gluck and Bower 1988, Goldstone 1994, Hayes-Roth and Hayes-Roth 1977).

Experiments in category learning indicate that having two features in common has more of an effect on similarity than twice the effect of having a single feature in common. This effect has been captured either by counting shared feature conjunctions in addition to shared features when determining similarity (Hayes-Roth and Hayes-Roth 1977) or by the 'rich get richer' effect in spreading activation models of similarity, where shared features reinforce each other by mutual excitation (Gluck and Bower 1988, Goldstone 1994).

The properties of stability under relabeling, relativizing contrast to a sub-universe, and synergistic combination also interact with one another. For example, it seems undesirable to have totally redundant features show synergistic effects on similarity, as non-contrastive features generally do not influence similarity much. We require a theory of feature specification and a similarity metric which respects these three properties, and which also provides a straightforward prediction of their interaction.

3.2. Current theories of feature specification

In this section, we examine three standard theories of feature specification and show that none provide a suitable basis for a similarity metric. In our discussion, we also highlight some additional problems with these theories which are not specifically related to similarity. In the next section, we introduce a new theory of feature specification which meets all of our needs while avoiding the pitfalls of the previous theories. We begin our discussion with the theory of full specification.

Full specification is the surface representation theory formalized in Chomsky and Halle (1968). In this theory, an individual language – or possibly Universal Grammar – has a list of bivalent distinctive features. For any positive value of a feature, the negative value is also defined. In addition, all phonemes are defined for all features. Together these two properties imply that the system is closed under complementation of individual features: the complement to any natural class of phonemes defined by [+X] is also a natural class, defined by [-X]. The literature on feature geometry (e.g. Clements 1985, McCarthy 1988) provides abundant evidence that the natural classes of phonology are not closed under complementation. For example, although [+spread glottis] phonemes may constitute a natural class, the complement of this class is not natural, since it would include vowels, sonorant consonants, voiced obstruents, and creaky-voiced phonemes.

Full specification does not provide a foundation for computing similarity because it provides no way to permit features to play a differential role depending on their redundancy or contrastiveness. In this theory, all features are equivalent. Completely redundant features are formally just the same as ones which are contrastive. This theory provides no way to describe the changes in similarity relationships which occur when different sub-universes are considered.

For this reason, Pierrehumbert (1993) prosed to compute similarity between consonants using contrastively specified feature matrices. In contrastive underspecification, feature assignments are made which minimally differentiate the consonant inventory (Steriade 1987). For example, the labial class in Arabic, with only three consonants, requires very few features, while the coronals and dorso-gutturals require many. The similarity metric employed by Pierrehumbert, in common with most used in the psychological literature, counts not just the shared features between two sounds, but the proportion of shared features to differentiating features. As the shared features decrease in proportion to the differentiating features, so the similarity decreases. As a concrete example, consider the pairs /f, m/ and /s, n/ in the Arabic consonant inventory. A feature assignment using full specification makes these pairs equally similar. Using contrastively underspecified features, the similarity of /s/ to /n/ is decreased dramatically by the number of additional features which are needed to differentiate /s/ from the other coronal fricatives. Analogous features are not needed for /f/, as it is the only labial fricative.

The use of contrastive underspecification can relativize similarity to place of articulation, due to the method by which features are chosen as redundant, and thus omitted. Pierrehumbert (1993), following Stevens and Keyser (1989), assigns 'major features' first, like the place features, [±sonorant], and [±continuant]. She then assigns 'minor features', like [±voice], secondary place features like [anterior], and manner features like [nasal], just in case they distinguish consonants within the major classes.

In general, contrastive underspecification makes a poor basis for similarity because it only relativizes contrast to pre-defined sub-universes, based on the major place and sonority features. The problem, though subtle, can be seen as a failure to meet the condition of stability under relabeling. It has long been known that implementing contrastive underspecification requires a decision about which features are the ones which differentiate phonemes, and which are redundant. As pointed out in Stanley (1967), combinatoric theory provides an immense number of different ways of doing this to any given feature matrix. Pierrehumbert (1993) assigns major place and sonority features first. This strategy utilizes an implicit hierarchy of the features, assuming that certain features are more important for classification than others. For example, assigning [±sonorant] before [±voice] represents an a priori decision that sonority is more important than voicing for the classification of consonants in Arabic. A more useful system would be one which can relativize the comparison to any sub-universe, and evaluate the contrastiveness of features within that sub-universe.

These objections are equally applicable to the theory of radical underspecification (Archangeli 1984, 1988). In radical underspecification, one value of each feature (generally the positive value) is the marked one. The unmarked, or default value, is simply omitted in the representation and is eventually supplied by rule during the course of the derivation. In addition, non-contrastive features are omitted as in contrastive underspecification.

Underspecification simply omits minor features which are redundant with major place and sonority features. For example, the there is no specification of voicing for the sonorant /l/. Since redundant information is completely omitted by underspecification, there is in fact no way to tell from the representation itself if /l/ is redundantly voiced or voiceless. The claim being made by employing underspecification as a basis for computing similarity is that 'redundant similarity' does not count as similarity – that is, that /l/ is no more similar to /d/ than to /t/ with respect to voicing. However, as we have seen, this prediction is false. Voicing, though redundant for sonorants, does have an effect on cooccurrence in Arabic. We require a similarity metric which is sensitive to contrastiveness, but does not omit non-contrastive information entirely.

The use of feature blanks to encode redundant or default features leads to a more general representational problem. Underspecification often leads to feature specifications which do not meet the individuation condition discussed in Broe (1995) and Frisch (1996). This property was originally referred to as a lack of distinctiveness by Stanley (1967). For example, in radically underspecified representations, the unvoiced counterpart of /g/, namely /k/, is characterized by a proper subset of the features characterizing /g/. That is, /k/ has all the same features as /g/, except that the feature [+voice] is left out. Since [-voice] is the underspecified default for voicing, the list of features which characterizes /k/ is identical to that characterizing the natural class {g, k}. There is no way to refer to /k/ individually in a similarity metric without referring to the natural class {g, k}. Indeed, in the worst case radical underspecification can leave the similarity of phoneme pairs undefined. For example, if /t/ is taken to have no features (being unmarked for all features), then it is impossible to compute its self-similarity, or its similarity to any other unmarked segment, such as an epenthetic vowel.

To summarize, current theories of feature specification do not have the desired characteristics of stability under relabeling and/or context sensitivity. Full specification has the additional undesirable assumption that all features are bivalent. Theories of underspecification omit information, resulting in a lack of individuation for segments and a loss of information that is relevant to the phonology of Arabic. In addition, note that none of the systems of feature specification provide insight into the effect of synergistic feature combination on similarity or how synergy interacts with these other properties. In the next section, we introduce the theory of structured specification, and show that it provides a framework in which the connection between stability under relabeling, context sensitivity, and synergy is obvious.

3.3. Structured specification

The theory of structured specification (Broe 1993) represents the segment inventory using a hierarchy of the natural classes of segments. We show that this hierarchy is a powerful tool for understanding how features can represent context sensitive dimensions of contrast which are stable under relabeling and combine synergistically in an intuitive way. Despite its explanatory power, the fundamental characteristics of the theory of structured specification are actually quite simple. Defining the possible natural classes has always been the goal of distinctive feature theory. A natural class is simply a set of segments defined by a conjunction features. Structured specification makes the natural classes an explicit part of the representation. Once the set of natural classes is considered directly as the result of a set of feature specifications, a hierarchy among natural classes automatically arises because the natural classes are (partially) ordered by set containment. Larger, more general natural classes contain smaller, more specific ones. See Broe (1993) for a more rigorous introduction to the theory.

By considering all possible combinations of features, we form a list of the entire collection of natural classes that can be defined within a set of phonemes using a given feature scheme. Given a segment inventory and a set of feature assignments, the hierarchy of natural classes for that set of segments given that feature set can be unambiguously determined. We exemplify the algorithm with a simple case, the three vowel inventory {a, i, u}, based on the monovalent feature specifications in (11).

(11)

	/a/	/i/	/u/
[high]		+	+
[low]	+		
[front]		+	
[back]	+		+

To construct a hierarchy for the three vowel inventory, we first consider the set of natural classes which can be denoted by the feature matrix in (11). There are 4 feature values, so there are $2^4 = 16$ possible conjunctions of features. Each feature conjunction, with its corresponding natural class, is given in (12).The symbol ∅ denotes the empty set, which is the natural class created by an incompatible conjunction of features.

(12)
{[]} = {a, i, u}
{[+high]} = {i, u}
{[+low]} = {a}
{[+front]} = {i}
{[+back]} = {a, u}
{[+high], [+low]} = ∅
{[+high], [+front]} = {i}
{[+high], [+back]} = {u}
{[+low], [+front]} = ∅
{[+low], [+back]} = {a}
{[+front], [+back]} = ∅
{[+high], [+low], [+front]} = ∅
{[+high], [+low], [+back]} = ∅
{[+high], [+front], [+back]} = ∅
{[+low], [+front], [+back]} = ∅
{[+high], [+low], [+front], [+back]} = ∅

Out of the 16 possible conjunctions of features, there are 7 distinct sets of segments. These are the natural classes. They are {a, i, u}, {a, u}, {i, u}, {a}, {i}, {u}, ∅. These 7 sets are partially ordered by the following set containment relationships.

(13)
{a, i, u} ⊇ {a, u}, {i, u};
{a, u} ⊇ {a}, {u};
{i, u} ⊇ {i}, {u};
{a}, {i}, {u} ⊇ ∅

Note that not all set containment relationships are given in (13). Relationships which can be deduced by transitivity are omitted. For example, {a, i, u} ⊇ {u}, which can be deduced from {a, i, u} ⊇ {a, u} and {a, u} ⊇ {u}.

A lattice is a partial ordering of the natural classes of segments which are possible given a featural representation. A standard text on lattice theory is Davey and Priestley (1990). Figure 2 shows the lattice of the three vowel inventory graphically. Each node in the lattice is a natural class, and the set of features and segments which the node denotes are shown above and below the node, respectively. Nodes are ordered from top to bottom by size. The top node of the lattice represents the entire inventory, and the bottom node is the empty set. The row of nodes just above the bottom are the singleton classes containing the individual segments. The features which denote these nodes are the features of the segments. Lines connecting nodes indicate set containment. As in (13), not all set containment relationships are indicated by lines, those which can be deduced by transitivity are excluded.

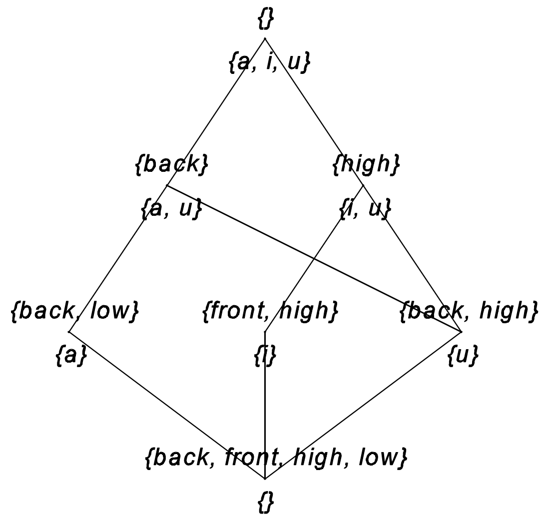

Figure 2: Lattice of the three vowel inventory.

There is a dualism between sets of segments (natural classes) and sets of features in the lattice. The hierarchical set containment relationship between the natural classes corresponds to an inheritance relation for the features that define those natural classes. For example, the natural class {u} is {[+back], [+high]}. It is contained by the natural class {i, u}, which is {[+high]} and by the natural class {a, u} which is {[+back]}. The natural class {u} inherits the feature [+back] from the natural class {a, u}, and it inherits [+high] from {i, u}. Through set containment and feature inheritance, the lattice represents redundancy structurally. For example, {[+front]} = {i} is contained by {[+high]} = {i, u}. Thus, every segment which is [+front] is a member of {[+high]}; in other words [+front] ⇒ [+high]. Featural redundancy can be 'read off of' the lattice.

The key insight in structured specification is that redundancy is a relation between features, and one which derives from the inclusion relation between their denotations. The lattice of natural classes constitutes a 'classification space', in which redundancy relations are explicitly represented. We propose to compute similarity directly over this classification space, and claim that this automatically conditions the similarity metric to redundancy in the appropriate way. Note especially the structural difference between two features {[f], [f']}where [f] implies [f'] (they are partially redundant), as opposed to two features {[f], [g]} where neither implies the other (no redundancy). The features [f] and [f'] each define a natural class, but the class defined by their conjunction is coextensive with the class defined by [f] (see the classes of front vowels and high vowels just discussed in figure 2). The information content of the conjunction is only as great as the information content of the most specific feature [f], and no new class is defined. On the other hand, the conjunction of the orthogonal features [f] and [g] defines a new natural class which is more specific than either (e.g. the back vowels and high vowels). There is an increase in information in the conjunction of non-redundant features which enriches the classification space by creating additional natural classes.

By highlighting the relationships of the natural classes, structured specification resolves the problem of stability under relabeling. This problem arose in part because of a move in phonological classification away from bivalent feature specifications of the form [±coronal], to monovalent feature specifications of the form [coronal], [labial], [dorsal]. This move was motivated by the fact that many negative classes – the non-labials, for example – never form active natural classes in phonological systems. The wholesale recasting of feature systems into monovalent form leads to problems in the computation of similarity in cases where features are indeed complementary, such as [sonorant] and [obstruent]. The problem can be illustrated by comparing /b/ and /m/ under two different feature specifications, as in (14).

(14)

	bivalent	monovalent	
	±son	obs	son
b	-	+	
m	+		+

In the bivalent scheme, classification along the dimension of sonority yields one feature mismatch; in the monovalent scheme, it yields two. The problem, of course, is that in the monovalent scheme the relation of complementation between the two properties is lost, and so cannot condition the computation of similarity. In structured specification, the complementation relation between properties is represented by the structural relations between the natural classes and appropriately conditions the computation of similarity regardless of the way those properties are labeled. Discussion of these issues in generative phonology has been unduly influenced by notational concerns – are pluses and minuses used in the notation or not – as opposed to substantive modeling of the logical relationships involved. In structured specification, the feature labels [-son] versus [+obs] are irrelevant, as the same natural class is defined in either case.

We do not <u>require</u> that every feature defines both a class and its complement. Indeed, we see that this is just the question of redundancy in another guise: if we know that /b/ is obstruent and /m/ is not, it is redundant to be informed that /m/ is sonorant and /b/ is not, just in case [sonorant] and [obstruent] are complementary properties.

Structured specification encodes redundancy without omitting information and thus avoids many of the problems introduced by underspecification theory (see Broe 1993). In structured specification, feature blanks are used solely to represent that a segment is undefined for a particular feature. For example, [±ant] is irrelevant for labials and so is not specified. Redundancy is encoded in the natural classes hierarchy as shown above. Properties which were formerly encoded by blanks are differentiated structurally in the hierarchy.

A lattice just like the lattice of the three vowel inventory can be constructed for the phonemes and features of the Arabic consonant inventory. Since there are 28 Arabic consonants, there are a very large number of natural classes. Displaying the entire lattice of natural classes graphically is not especially informative. Figure 3 shows a simplified lattice which contains the traditional major cooccurrence classes within the Arabic consonant inventory. Each major class is represented by a single consonant from that class. In the complete lattice, the natural classes highlighted in figure 3 are found, as well as many sub-classes of the major classes. Figure 3 also shows some of the cross-classification of classes discussed in section 2. For example, the [coronal] node contains coronal stops (e.g. /d/), coronal fricatives (e.g. /z/) and coronal sonorants (e.g. /l/). The [dorso-guttural] node contains the velars (e.g. /k/) and the gutturals (e.g. /h/).

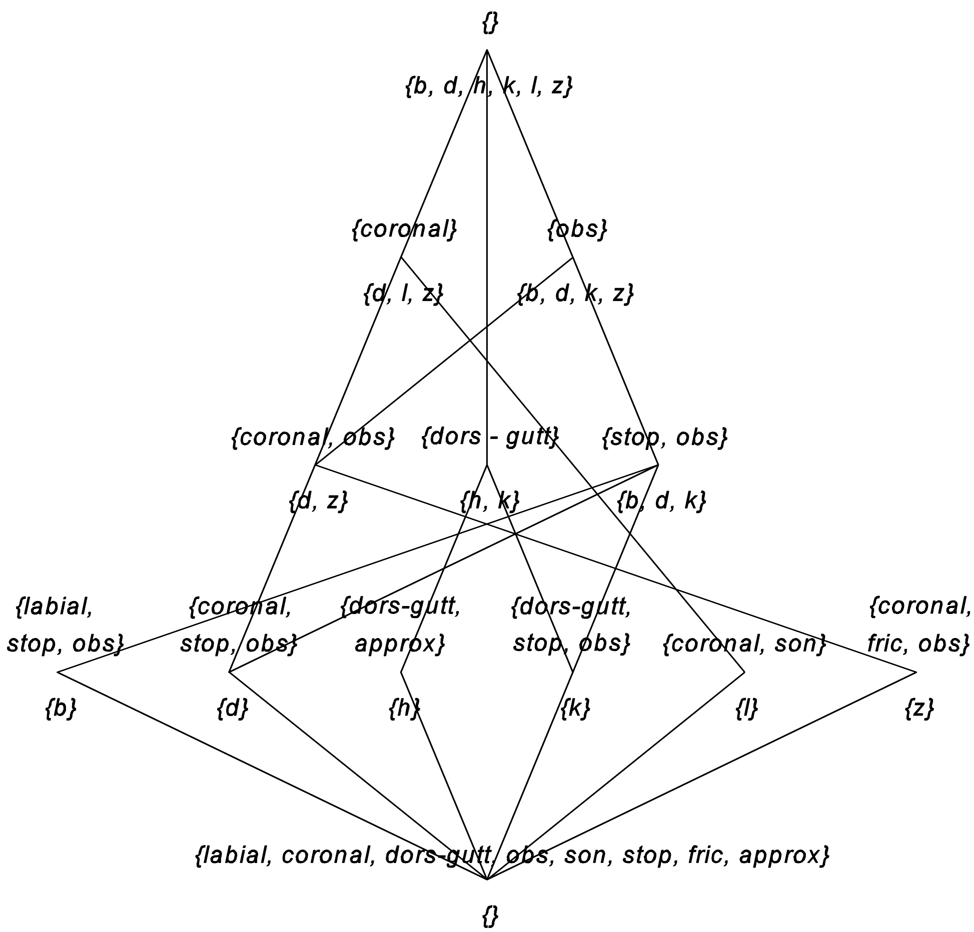

Figure 3: Lattice of selected Arabic consonants, one from each major class.

Some of the natural classes in figure 3 have no cooccurrence restriction. For example, there is no cooccurrence restriction on obstruents (e.g. {b, d, k, z}) in general. There is only a restriction for consonants which share place. We have already noted that the appropriate domain for the computation of similarity for the OCP-Place constraint is the set of sub-universes based on the places of articulation. Relativizing the featural relations to a sub-universe requires no additions to the lattice formalism. The total lattice contains sub-lattices for each sub-universe of consonants which share place of articulation.

Figure 4 shows a complete sub-lattice of the Arabic consonants restricted to the [labial] place of articulation sub-universe. This lattice was created by removing all non-labials from the set of natural classes of Arabic. No features are left out of the lattice, only segments are excluded. Many of the features apply to no segments in the labial lattice, so they are found at the bottom of the lattice, which is the empty set.

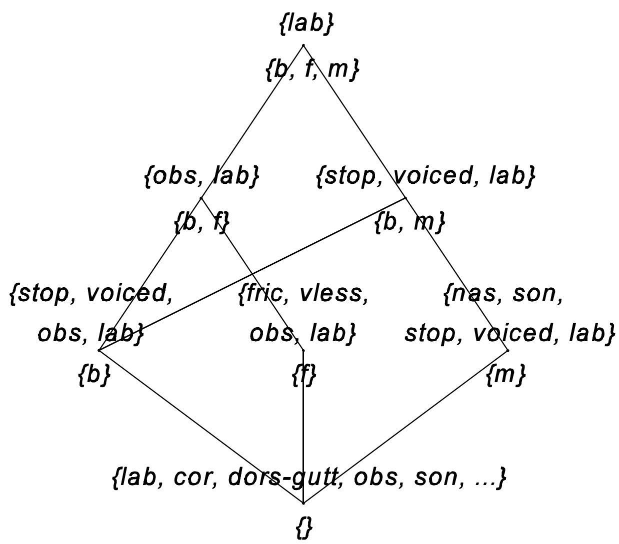

Figure 4: Lattice of the labial consonants of Arabic, {b, f, m}.

Recall that Pierrehumbert (1993) used contrastive underspecification to account for the difference in strength of OCP effects between the pairs /f, m/ and /s, n/. The comparison of /f, m/ used very few features, since the small class of labials could be differentiated by a small number of distinctive features. The comparison of /s, n/ involved many features, since /s/ needed to be differentiated from /θ/, /S/, and /ʃ/, which are all voiceless coronal fricatives. The use of natural classes in the lattice representation captures this distinction without underspecification. The redundant features within the labials do not support additional natural classes but rather they aggregate at existing class nodes, and the structure of the labial lattice is 'simpler' than if there were a greater number of contrasts. We take advantage of structured specification in the computation of similarity by basing our calculation on the lattice of natural classes. As a result, we can achieve the same effects of the size of the inventory and the number of contrasts on similarity without resorting to underspecification. In this way, the similarity of /f, m/ is greater than the similarity of /s, n/, and in general the large classes of [coronal] and [dorso-guttural] consonants split into highly dissimilar subclasses while the labials do not.

3.4. The similarity metric

We compute similarity by comparing the number of shared and unshared natural classes of two consonants, using the equation in (B). We refer to our metric as the <u>natural classes model</u> of similarity in the discussion below.

(B) $$\textit{Similarity} = \frac{\textit{Shared natural classes}}{\textit{Shared natural classes} + \textit{Non-shared natural classes}}$$

Identical consonants, with all natural classes shared, have similarity 1.0, while highly dissimilar consonants share very few natural classes (because they share very few features) and have very low similarity. Since OCP effects only apply to consonants which share major place of articulation features, the lattices used in the similarity computations are the sub-lattices for the sub-universes of each major place class.

We illustrate the metric with sample similarity values based on figure 4. The pair /f, m/ share 1 class, namely {[lab]}. They have 4 non-shared classes: {[obs], [lab]}, {[stop], [voiced], [lab]}, {[fric], [vless], [obs], [lab]}, and {[nas], [son], [stop], [voiced], [lab]}. The similarity of /f, m/ is 1/5 by equation (B). The pair /b, f/ share 2 classes, {[lab]} and {[obs], [lab]}. They have 3 non-shared classes: {[stop], [voiced], [lab]}, {[stop], [voiced], [obs], [lab]}, {[fric], [vless], [obs], [lab]}. The similarity of /b, f/ is 2/5. Intuitively, we identify the two parts of the lattice that dominate each of the two segments, and note the degree of overlap. The greater the overlap, the greater the similarity.

We can get some idea of how similarity for pairs like /s, n/ differs from the similarity of /f, m/ due to the larger space of contrasts in the coronals by considering a second, hypothetical labial inventory. Suppose the Arabic inventory were to contain two additional labials, /p/ and /v/. The lattice of the expanded inventory, based on the same features as figure 4, is shown in figure 5. Notice the increase in complexity with the addition of these two segments, even though there is no change in the features used. In the expanded inventory, the pair /f, m/ still share only 1 class: {[lab]}. They now have 8 non-shared classes. The four extra classes are: {[stop], [lab]}, {[fric], [obs], [lab]}, {[voiced], [lab]}, {[vless], [obs], [lab]}. The similarity of /f, m/ in the expanded inventory is 1/9 by equation (B). Using a similarity metric based on natural classes, the difference in the strength of OCP effects in the major classes can be captured without resorting to underspecification. The larger number of natural classes among the coronals and dorso-gutturals decreases the perceived similarity between class members that share few features, due to the contrastiveness of many of the non-shared features. In a small class, like the Arabic labials, these non-contrastive features do not contribute as independently to dissimilarity.

Using natural classes in the similarity computation provides an simple understanding of the interaction between stability under relabeling, context sensitivity, and synergy of features. In the natural classes similarity model, there is an increased similarity which is created by multiple feature matches which gives them greater effect than two feature matches would have individually, capturing the effect of synergy. For example, if two consonants match in the features [coronal] and [obstruent], then they share three natural classes: {[coronal]}, {[obstruent]}, and {[coronal], [obstruent]}. If two consonants match on three features, then they share as many as seven natural classes if the features are orthogonal. But synergy is only found if the phoneme inventory is rich enough to support all of the contrasts with distinct natural classes, so context sensitivity is also captured. Recall that the two lattices of labials in figures 4 and 5 are based on the same set of features. Multiple feature matches have a synergistic effect on the number of shared natural classes, but similarity also depends on the system of contrasts. Completely redundant features in these lattices do not create additional natural classes, so the similarity metric is also stable under relabeling. Additional features increase similarity only when they are contrastive in the sub-universe under consideration, so they can have different effects when different sub-universes are considered.

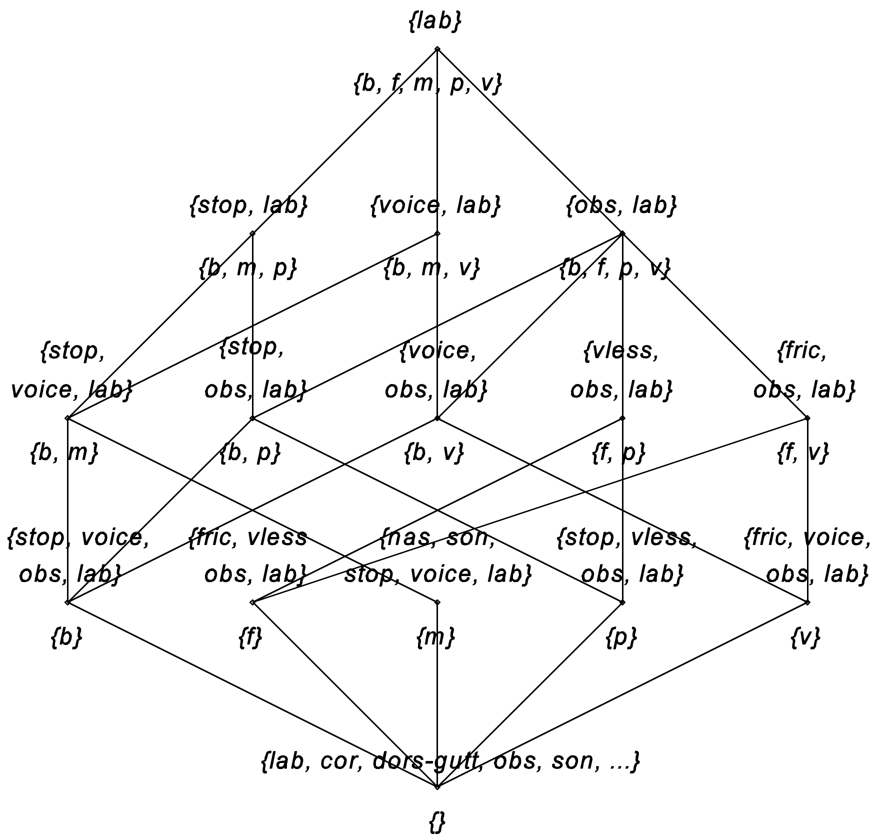

Figure 5: Lattice of hypothetical expanded inventory of labial consonants of Arabic, {b, f, m, p, v}.

In this section, we have proposed a similarity metric based on the theory of structured specification. Our approach to feature specification and similarity has a number of advantages over alternative methods. Completely redundant features define no new natural classes and so do not effect similarity, maintaining stability under relabeling. The lattice formalism does not omit redundant features, it encodes them in the structure of the natural classes, so the influence of features is context sensitive and can be relativized to any sub-universe. Orthogonal features combine to make distinct natural classes, so orthogonal feature matches have synergistic effects on similarity.

To examine the effect of similarity on cooccurrence in Arabic, we computed the similarity of consonant pairs based on shared and non-shared natural classes using the feature specifications for the Arabic consonants given in the appendix. The computations were made on three separate sub-lattices, similar to those in figure 5, one for each major place of articulation: [labial], [coronal], and [dorso-guttural]. Pairs involving the emphatic coronals, which are found in both the coronal and dorso-guttural lattices, were given the maximal similarity value computed by the coronal and dorso-guttural lattices. Consonant pairs which were not found on the same lattice, and thus do not share place of articulation, were given similarity zero, to reflect their non-participation in the OCP-Place constraint (note that a non-homorganic consonant pair has no shared nodes on any place lattice). As discussed in section 2.3 above, such pairs are necessarily overrepresented; therefore they act as if they were even less similar than homorganic pairs having a very small number of features in common. A complete table of the similarity values for each consonant pair can be found in Frisch (1996); we present aggregate data in next section.

4. The stochastic constraint model of OCP-Place

Given the similarity metric developed in the previous section, the next step is to find an appropriate model of the Arabic cooccurrence restrictions based on similarity. The ideal model predicts the actual number of occurrences for each consonant pair, in both adjacent and non-adjacent positions, and thus is based on both similarity and distance. We find that a mathematical model is suggested by the data, when the data are aggregated over similarity values. Table 4 shows O/E in the verbal roots when consonant pairs are aggregated by similarity. We aggregated identical pairs and non-homorganic pairs into distinct classes, with similarity of 1.0 and 0, respectively. Homorganic but non-identical pairs were aggregated into equal similarity intervals of width 0.1. There are no consonant pairs in Arabic with similarity in the range (0.7,1) as computed by the natural classes model, so no data is reported for that range.

Figure 6 shows O/E values for the aggregated data from table 4. Each point in figure 6 represents one group of consonants aggregated by similarity. Each point has an x coordinate of the mean similarity of consonant pairs in that group, and y coordinate the aggregate O/E for that group. The left panel of figure 6 shows the O/E for adjacent pairs, and the right panel of figure 6 shows the O/E for non-adjacent pairs. Clearly, there is an effect of similarity on cooccurrence in the Arabic data. That the effect weakens with distance is also seen when the left panel is compared to the right. The left panel shows an overall S-shaped pattern to the data, with an upper and lower asymptote on O/E. The data in the right panel are less uniform, but an S-shaped curve does not look inappropriate.

Similarity	Adjacent O/E	Non-adjacent O/E
0	1.27	1.13
0-0.1	1.23	1.08
0.1-0.2	0.89	1.09
0.2-0.3	0.58	0.77
0.3-0.4	0.22	0.75
0.4-0.5	0.08	0.73
0.5-0.6	0.07	0.32
0.6-0.7	0.00	0.32
0.7-1	-	-
1	0.01	0.15

Table 4: Cooccurrence of consonant pairs aggregated by natural classes similarity, for adjacent and non-adjacent pairs.

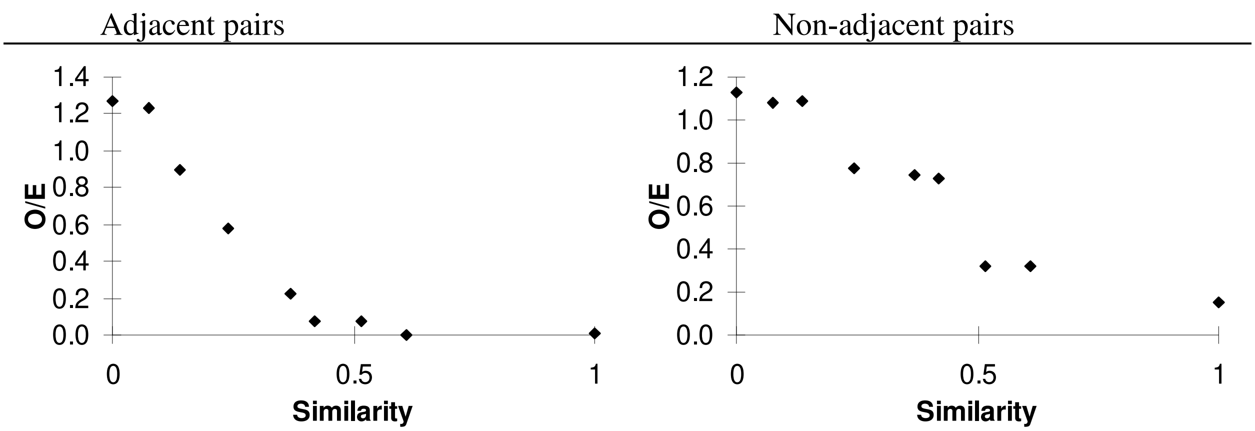

Figure 6: Aggregate O/E for adjacent pairs (left panel) and non-adjacent pairs (right panel) based on similarity as computed in the natural classes model.

The pattern of data in figure 6 is reminiscent of the S-shaped curves found in identification experiments demonstrating so-called categorical perception (see Repp 1984 for a review). In these experiments, synthetic stimuli on a continuum between two linguistic contrasts, for example a VOT continuum between /t/ and /d/ or an F2 continuum between /d/ and /g/, is identified categorically, with a sharp boundary between categories. We view phonotactic acceptability judgements, as reflected in the pattern of lexical items in Arabic, as a type of categorical perception. In this case the category is 'phonotactically acceptable word'. It is important to note that extending our investigation and analysis to gradient data does not alter much of the fundamental nature of linguistic generalizations. We maintain the basic categorical structure of language, grounded in distinctive features and natural classes, but adopt a more psychologically realistic formalization in which phonological type frequency plays an important role in determining acceptability and in which constraint boundaries may be non-categorical.

The sharpness of the boundary in categorical perception depends on the type of contrast examined. VOT boundaries and consonant place of articulation cues have sharp boundaries while vowel cues show a much more gradual boundary. Adjacent consonant pairs in Arabic show a sharp boundary. Non-adjacent pairs have a smoother boundary. Conceptually, our account of the Arabic data is to claim that the observed patterns are the result of a gradient linguistic constraint, based on the perceived similarity of homorganic consonant pairs. Perceived similarity is a combination of the (paradigmatic) similarity discussed in the previous section, conditioned by interference due to (syntagmatic) temporal distance (discussed in detail in section 4.2).

The true test of an account is whether it is the best known account of the entire set of data under discussion. We therefore compare our model with previous models on their ability to fit all of the Arabic data. This requires a precise quantitative model of cooccurrence in Arabic. In this section, we develop the stochastic constraint model of OCP-Place, compare its ability to fit the Arabic data with a variety of other models, and show that it provides a better account of the data. The following sections contain some mathematical detail which may not be of interest to all readers. We have made an effort to structure the exposition so that the major theoretical points can be appreciated without absorbing all of the mathematical detail.

4.1. The logistic function

We begin constructing an explicit model of cooccurrence in Arabic with a mathematical function appropriate for the data. For data like the Arabic data in figure 6, which asymptotically approach two limits, and vary smoothly in between in an S-shaped pattern, a logistic function is an appropriate mathematical model. The equation of the logistic is repeated in (A).

(A) $$y = \frac{1}{1 + e^{K + S \cdot x}}$$

$\underline{K}$ and $\underline{S}$ are parameters of the logistic which determine its exact shape. These parameters will be fixed in modeling the precise pattern of data in Arabic, as discussed below. If $\underline{S} > 0$, the logistic function asymptotically approaches $\underline{y} = 1$ as $\underline{x}$ approaches negative infinity (high acceptability for low values of the parameter $\underline{x}$), and it asymptotically approaches $\underline{y} = 0$ as $\underline{x}$ approaches positive infinity (low acceptability for high values of the parameter $\underline{x}$).

For the Arabic data, and OCP effects in general, the variable $\underline{x}$ is based on similarity. The variable $\underline{y}$ is the abstract acceptability of a form. The stochastic constraint model of Arabic is shown in (C). Informally, if similarity is small, then the denominator in (C) is small, and acceptability is high. If similarity is large, then the denominator is large, and acceptability is low. In the next section, we discuss applying this model to the Arabic data to determine the values of the parameters $\underline{K}$ and $\underline{S}$ which best fit the pattern of cooccurrence in Arabic.

(C) $$Acceptability = \frac{1}{1 + e^{K + S \cdot Similarity}}$$

We use the logistic function as the mathematical basis for a gradient constraint model which we call the stochastic constraint model. The cooccurrence restrictions between adjacent consonants in Arabic are quite strong, but other languages, like English, have weak OCP-Place effects (Berkley 1994a, Frisch 1996). Since phonological patterns can be gradient to different degrees, a general model of gradient constraints needs to be able to capture many different strengths of constraint. We can model different strengths of constraints, from highly gradient to categorical, by altering the parameters K and S. Figure 7 shows three different stochastic constraint models of increasing strength. The leftmost panel in figure 7 shows a highly gradient constraint (K = -4, S = 8). The center panel shows a gradient constraint with a sharp boundary (K = -10, S = 20). The rightmost panel shows a categorical constraint (K = -5,000, S = 10,000).

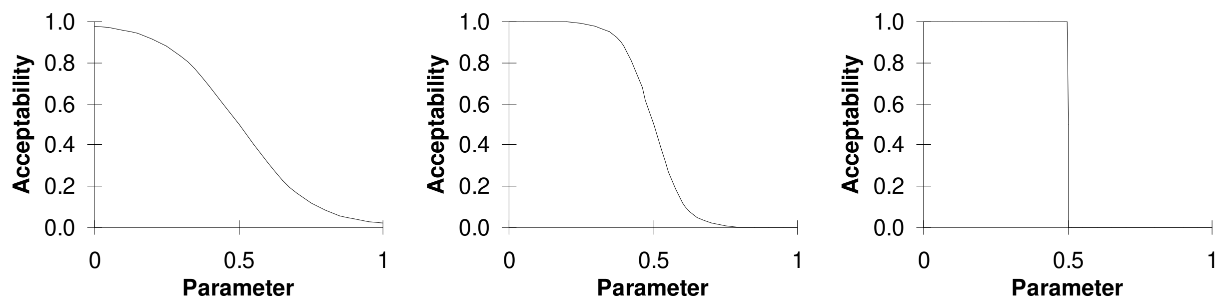

Figure 7: Three stochastic constraint models. Left panel: Weak constraint appropriate for highly gradient data. Center panel: Strong gradient constraint. Right panel: Categorical constraint.

The parameters of the stochastic constraint control two characteristics of the gradient constraint which we refer to as the constraint boundary and the constraint sharpness. The stochastic constraint model uses gradient degrees of acceptability and so does not generally have a categorical constraint boundary like standard phonological constraints. The midpoint of the acceptability range (Acceptability = 0.5) is interpreted as the constraint boundary. The midpoint of the acceptability range is found at x = -K/S. Constraint sharpness refers to the steepness of the boundary between acceptable and unacceptable forms. Sharpness is controlled directly by the parameter S. For a given sharpness (fixed S), the location of the boundary is controlled the parameter K. The sample stochastic constraints in figure 7 have increasing sharpness, with K adjusted to maintain the constraint boundary at x = 0.5.

Figure 8 shows four logistics which demonstrate graphically the possibilities for modeling cross-linguistic variation based on altering the parameters K and S The first stochastic constraint, on the upper left, has K = -5 and S = 20. This constraint has its boundary at x = 0.25. We can shift the location of the constraint boundary by changing K. The second stochastic constraint, on the upper right of figure 8, has K = -15 and S = 20. The change in K has shifts the constraint boundary rightward to 0.75 while maintaining the same sharpness. The third stochastic constraint, on the lower left of figure 8, has K = -5 and S = 10. This stochastic constraint is more gradient as the sharpness is reduced. The more gradual slope of the constraint also shifts the boundary to 0.5. We could maintain the boundary at 0.25 by changing K to -2.5.

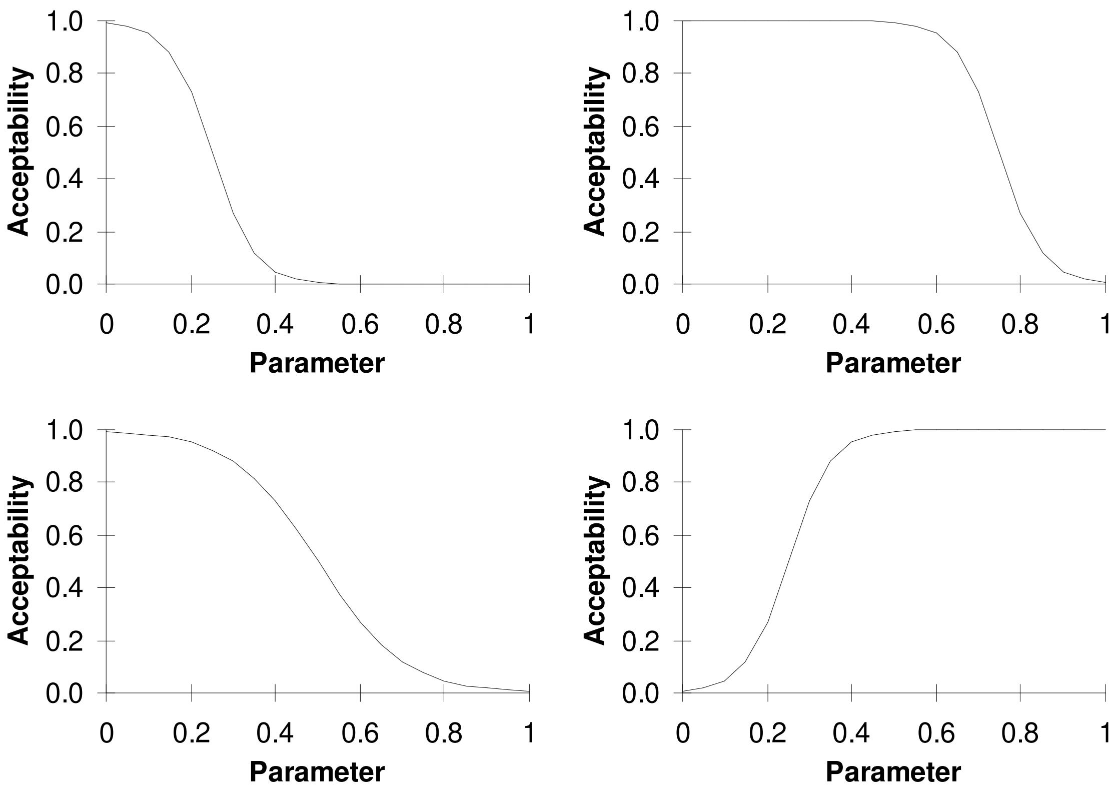

Figure 8: Adjustment of individual parameters of the stochastic constraint. Top left: K = -5, S = 20. Top right: constraint boundary shift, K = -15, S = 20. Bottom left: reduction in sharpness, K = -5, S = 10. Bottom right: assimilatory constraint K = 5, S = -20.

The stochastic constraint based on similarity can also be used to model assimilatory phenomena. In this case, high similarity is acceptable, and low similarity is unacceptable. This is captured in the stochastic constraint by reversing the signs on K and S. The fourth stochastic constraint, on the lower right of figure 8, shows an assimilatory stochastic constraint with K = 5 and S = -20. Figure 8 shows that stochastic constraints which require different degrees of either similarity or dissimilarity can easily be modeled by altering the parameters K and S.

In the discussion above, the stochastic constraints modeled acceptability (on the y-axis) against similarity (on the x-axis). However, the logistic model as a stochastic constraint can be applied to model acceptability of a form to any parameter x, and thus could be applied to any category defined by variation on a single dimension.

4.2. The model of cooccurrence

In order to compare the ability of the stochastic constraint model to predict cooccurrence in the Arabic data to previous accounts, we must elaborate the stochastic constraint model in two ways. First, we need to incorporate the effects of distance on the perception of similarity, to account for differences in cooccurrence between adjacent and non-adjacent consonants. Second, we need to translate the abstract acceptability of a form to actual patterns of cooccurrence.

We claim that distance reduces the effect of similarity on cooccurrence, weakening the constraint for non-adjacent consonants. Effects of temporal spacing on perceived similarity in the relatively short time scale of speech, on the order of one to two seconds, is found both in visual and auditory perception (Ericksen and Shultze 1973, Massaro 1970, Pisoni 1973). We assume that as temporal distance increases, extreme similarity and dissimilarity are more difficult to judge, so moderate or neutral similarity values result from the similarity comparison of distant objects (Pierrehumbert 1993). We take neutral similarity to be the mean similarity over all consonant pairs, which is 0.15 for Arabic. In order to capture the effects of distance on cooccurrence, our model of Arabic cannot be based directly on the similarity of homorganic consonants as computed using the natural classes model. Rather, we use perceived similarity which includes the effects of distance on similarity. We propose to compute perceived similarity using (D).

(D) $$\textit{Perceived similarity}(x,y) = \textit{Mean similarity} + \frac{\textit{Similarity}(x,y) - \textit{Mean similarity}}{\textit{Distance}}$$

This equation has the desired distance effects. As distance becomes large, perceived similarity approaches mean similarity.

The revised model of OCP-Place is given in (E).

(E) $$\textit{Acceptability} = \frac{1}{1 + e^{K + S \cdot \textit{Perceived similarity}}}$$

Our account of gradient cooccurrence connects relative frequency (O/E) to acceptability. The similarity based OCP predicts that highly similar consonants are unacceptable, and thus do not cooccur (underrepresentation). Highly dissimilar consonants are acceptable, and cooccur frequently (overrepresentation). We claim that relative frequency is an index of acceptability, which can be expressed as in (F1), or its equivalent (F2), where C is a constant.

(F) 1. $O/E = C \cdot \textit{Acceptability}$
2. $O = E \cdot C \cdot \textit{Acceptability}$

In order to predict O/E from acceptability, we next derive the value of the constant C via our claim that neutral similarity results in random cooccurrence. We predict highly similar consonant pairs are underrepresented, and highly dissimilar consonants are overrepresented. For consonant pairs which are neither similar nor dissimilar, we predict random cooccurrence (O/E = 1, or O = E). In other words, O approaches E as similarity approaches mean similarity. Given that distance causes perceived similarity to approach mean similarity, we also predict that O should approach E as distance becomes large. Since we assume O = E when perceived similarity equals mean similarity, we can substitute mean similarity for perceived similarity in equation (E), and use the resulting value of acceptability in equation (F2) to find the constant C. This gives (G).

(G) $$O = E \cdot C \cdot \frac{1}{1 + e^{K + S \cdot \mathit{Mean\ similarity}}}$$

Given (G), O = E only if C = 1 + $e^{K+S \cdot \mathrm{Mean\ similarity}}$. Since the value of C is now known, we substitute it in (F2) and use the equation for acceptability in (E) to create an equation relating O/E to perceived similarity. Thus, the model we propose to predict Arabic cooccurrence is given in (H). We claim this model captures all of the predictions of the similarity account.

(H) $$O = E \cdot (1 + e^{K + S \cdot \mathit{Mean\ similarity}}) \cdot \frac{1}{1 + e^{K + S \cdot \mathit{Perceived\ similarity}}}$$

The final step is to determine precise values for the parameters of the logistic function, K and S. We determine the appropriate values for these parameters by finding the values which provide the best fit to the Arabic data using non-linear regression. The model was fit to individual consonant pairs in the triliteral roots. There was one point for each pair of consonants in first and second, second and third, and first and third position in the triliteral root. Note that we fit the model to actual occurrence of consonant pairs rather than to O/E so that small differences between actual and expected cooccurrence would have minimal influence on the model, regardless of the magnitude of the number of expected pairs. For small expected values, O/E is unstable, so portions of the data which are sparse could have a disproportionate influence on the model fit.

In our model, distance, which is a factor in computing perceived similarity in (D), is found to be 1 for adjacent pairs, and 4 for non-adjacent pairs. A 4:1 ratio of distances between non-adjacent pairs and adjacent pairs provided the best model fit to the data among the ratios 2:1, 3:1, 4:1, and 5:1. This distance roughly corresponds to the number of intervening phonemes between pairs of consonants in a typical verb (e.g. katab). This is compatible with the distance used in Berkley (1994a) to study the effects of distance on the OCP in English.

Figure 9 shows predicted O/E plotted against similarity for the Arabic data. Predicted O/E is derived from equation (H), with the best fit regression parameters K = -3.22, S = 13.73, and distance of 1 and 4 for adjacent and non-adjacent pairs, respectively. The left panel shows predicted O/E for adjacent pairs. The right panel shows predicted O/E for non-adjacent pairs. The curves compare well against the aggregate data plots in figure 6.

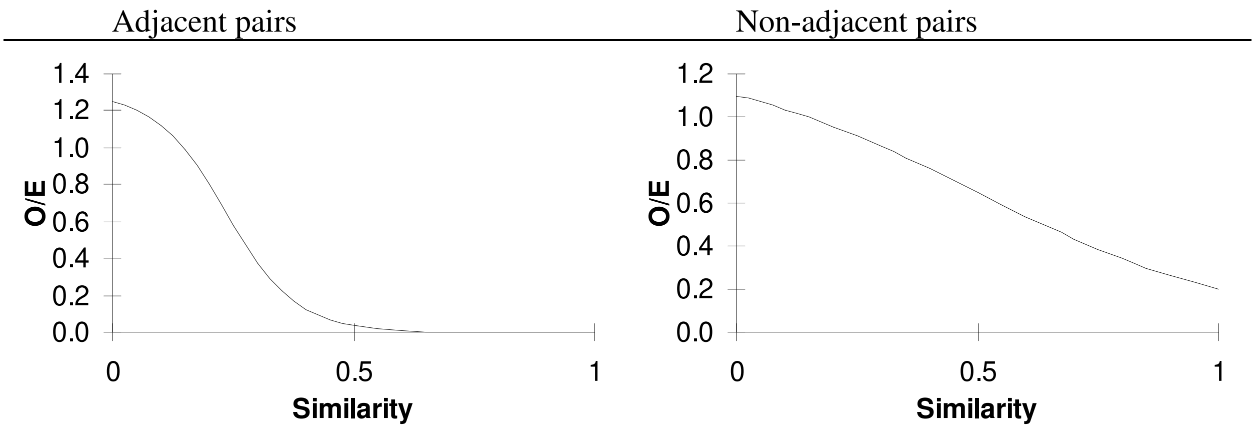

Figure 9: Predicted O/E for the stochastic constraint model using natural classes similarity.

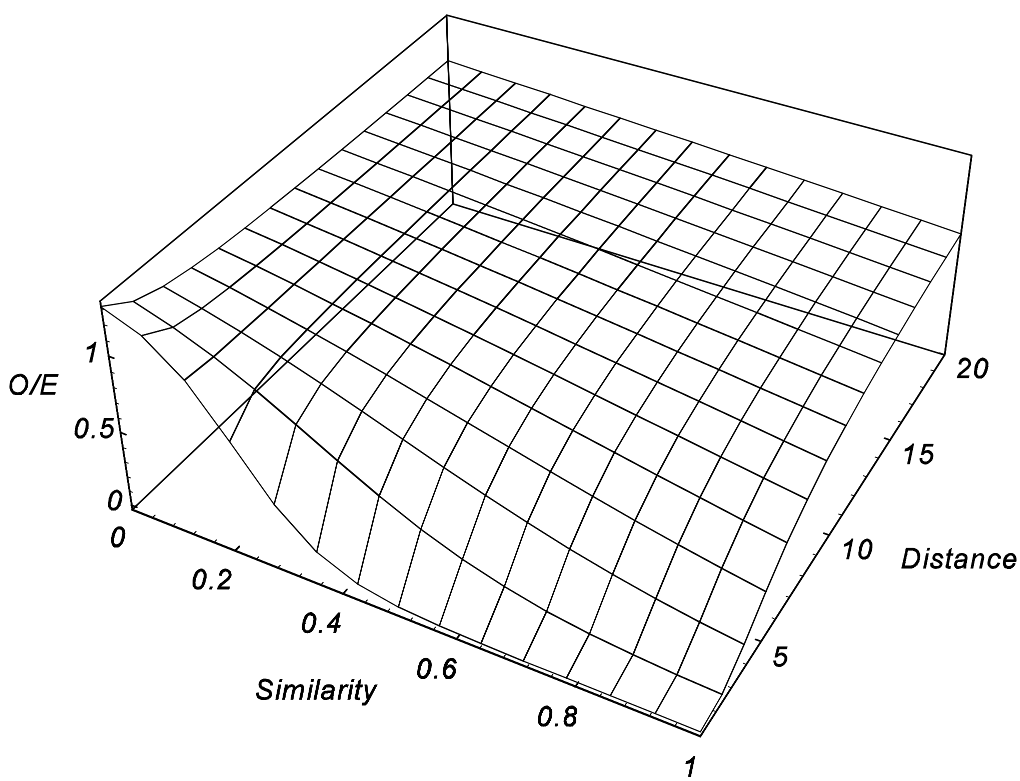

Figure 10: Three dimensional plot of predicted O/E versus similarity and distance.

We claim that distance reduces the strength of OCP-Place effects by interfering with the perception of similarity. Our prediction can be visualized using a three dimensional plot of similarity, distance, and predicted O/E for the model, shown in figure 10. As distance increases, the effects of similarity on predicted O/E is reduced, and the curve is nearly flat with an O/E of 1 at distance 20. As distance approaches 1, the logistic curve becomes quite steep, reflecting the strong cooccurrence constraint for adjacent consonants. If there were more positions within the Arabic roots, the hypothesized effect of distance could be tested in greater detail. This is an important issue left for future research.

4.3. Other models

Now that we have a specific model of OCP-Place effects in Arabic, which makes quantitative predictions about the distribution of the data, we can test our model empirically. In order to test our model, we compare its ability to predict the occurrences of each type of consonant pair in each position in the verbal roots of Arabic against four other models of OCP-Place. Our model has two distinctive components: the pattern of cooccurrence is predicted by a stochastic constraint based on similarity, and similarity is computed using natural classes. We refer to our model as the <u>natural classes model</u>.

An alternative model of Arabic might accept the hypothesis that cooccurrence is predicted by a stochastic constraint, but reject the use of structured specification and our similarity metric. We created a second set of similarity values for the Arabic consonants based on the same features as our model, but which uses a computation of similarity over features, rather than natural classes. This model is analogous to Pierrehumbert's (1993) model, but without the use of contrastive underspecification (which we have independently argued against). This model is included to test the hypothesis that structured specification provides a better representational system for determining similarity than a feature based metric. We refer to this model as the <u>feature model</u>.

We might also consider a model which accepts the similarity metric based on natural classes, but models stochastic constraints using some other function. Given the overall S-shape of the data, the differences between an alternative function and the logistic function we use is likely to be small. We believe that there is not sufficient data in the Arabic lexicon to test this point empirically. In any case, our model is based on the conception of a gradient linguistic category with a fuzzy boundary, so an alternative S-shaped curve which fit the data would suit our purposes equally well. The choice of the logistic as the particular mathematical function is not crucial, and was made for simplicity and convenience.

We created two other models of the data based on the autosegmental account of McCarthy (1994). The first model, which we call the <u>categorical model</u>, uses a categorical cooccurrence restriction on homorganic consonants, whether adjacent or non-adjacent. This model predicts no pairs of homorganic consonants in adjacent or non-adjacent position. This model represents a strict interpretation of the formal account of the total OCP and OCP-Place effects in McCarthy (1994). Under this model, the total OCP duplicates OCP-Place effects, as adjacent identical consonants are also homorganic, and thus are categorically ruled out by OCP-Place without the need for a total OCP. We note that this represents a very literal interpretation of the formalism of autosegmental theory, including weaknesses which are already admitted in the discussion in the text of various papers. However, we evaluate it because it provides an important baseline for other formal models. The fact that the categorical account was not intended to make quantitative predictions is seen below in the extremely poor fit of the model.

We also implemented a second autosegmental model, which more faithfully reflects the intentions of McCarthy's analysis. This model uses a categorical cooccurrence restriction on identical consonants in adjacent position, and also a soft constraint against homorganic consonant pairs in adjacent and non-adjacent positions. The soft constraint is implemented as a constant predicted O/E value for all homorganic pairs. We call this model the soft model.

In implementing a model based on the autosegmental analysis, the issue of what consonants count as homorganic arises. Recall that the coronal obstruents and sonorants are split by manner into different major cooccurrence classes. McCarthy's account treats these two groups as non-homorganic. Our account asserts that they are subject to OCP-Place, but only weakly. In addition, we argued for a single [dorso-guttural] place class while McCarthy used separate [dorsal] and [pharyngeal] classes.

For each of the two autosegmental models we fit two different versions representing the different assumptions of homorganicity in our account and McCarthy's. The first version, based on our analysis, uses an OCP-Place cooccurrence constraint for all homorganic consonant pairs based on the feature assignments in the appendix. This version has cooccurrence restrictions between manner sub-classes and for the secondary articulations of the coronal emphatics. This model has cooccurrence constraints over the same consonant pairs as the natural classes model and the feature model. The second version uses an OCP-Place constraint only within the traditional major classes, as in McCarthy's original analysis. This version has a division between coronal obstruents and sonorants, another within the dorso-guttural class, and does not have cooccurrence restrictions based on secondary articulations for the coronal emphatics.

Finally, we include for comparison a fifth model which does not take into account OCP-Place effects at all. We refer to this model as the frequency model, as consonant cooccurrence is predicted based only on the random combination of consonant pairs. In other words, O = E and cooccurrence is predicted to be entirely random. This model includes no provisions for the OCP-Place constraint, and thus serves to highlight the effect of implementing the constraint in the other models. In addition, this model makes explicit the effect of including frequency as a base line measure of cooccurrence in the other models. All of the other models predict OCP-Place effects as a deviation from expected cooccurrence.

4.4. Model comparison

In this section, we discuss the details of the quantitative comparison of model fits, and show that the natural classes model provides the best fit. We first present the measures of model fit which we apply, and then compare the models in detail. We review results and discuss the theoretical implications in the summary section which follows.

Since a good model of OCP-Place predicts strong underrepresentation within the major classes, the predicted cell counts in many cases are low, and the absolute goodness of fit of the models to the data cannot be determined. In other words, we cannot test the significance of the model fits. We can still compare models using relative measures of goodness of fit. We use two such measures, the residual sum of squares and the R^2 statistic.

We compare all of the models based on their best fits to the data. For the two models which utilize similarity, the feature model and the natural classes model, we use the best fit parameters <u>K</u> and <u>S</u> in the stochastic constraint model in equation (H). For computing perceived similarity in the feature model, mean similarity is 0.22, and the best ratio of distances between non-adjacent and adjacent pairs was 4:1. We implemented the two autosegmental models using constant predicted O/E values for each distinct sub-case of cooccurrence in the model. For the categorical model, O/E for OCP-Place violations was 0, and a best fit constant value was used for non-violations. For the soft model, O/E was 0 for total OCP violations, a best fit constant for OCP-Place violations, and a best fit constant for non-violations. The frequency model had no best fit parameters, O = E for all consonant pairs.

We can compare the models overall using the R^2 measure and the residual sum of squares. In addition, we present two breakdowns of the residual sum of squares for each model. The first is the residual sum of squares over the homorganic consonant pairs, as defined by the feature matrices in the appendix and used in the natural classes model. The second is the residual sum of squares over the consonant pairs within the major classes. The homorganic consonant pairs, and in particular those within the major classes, are the consonant pairs which are crucially involved in the OCP-Place constraint and thus provide important data for comparing the models.

Table 5 shows the models, their parameters, and the evaluations of goodness of fit. For each model, best fit parameters are given in the rightmost column. R^2 statistics and the overall residual sum of squares are shown in columns two and three. Columns four and five show the breakdown of the residual sum of squares for homorganic consonant pairs (Hom), and for non-homorganic consonant pairs (Non-Hom). Columns six and seven show the breakdown of the residual sum of squares for pairs within the major classes (Maj), and pairs between the major classes (Non-Maj). Recall that two different versions of the autosegmental models were fit with two different definitions of homorganicity for the purposes of OCP-Place. For these models, residual sum of square breakdowns are only given for the versions with the corresponding definitions of homorganic. Fits for the mismatched versions (e.g. fits over the major classes for a model with best fit parameter over all homorganic pairs) is by definition worse.

We first compare R^2 statistics for the various models. The natural classes model provides the best overall fit to the data, accounting for 80% of the variation found in the lexicon. The frequency model shows that a significant percentage of the variation in the data is accounted for by frequency alone, emphasizing the relevance of frequency as a base predictor of cooccurrence rate (Pierrehumbert 1994, Frisch 1996). The model with the worst fit is the categorical model applied to all homorganic pairs. This model rules out all homorganic pairs, including for example all combinations of coronal obstruents and coronal sonorants, which is much too strong of a constraint. This model reflects the effect of a categorical non-quantitative OCP-Place, without special elaborations, and without allowing gradience of any kind. It has long been known that such a model is inappropriate, and that fact is reflected in the quantitative fit.

Model	R^2	Residual SS	Homorganic Hom	Non-Hom	Major classes Maj	Non-Maj	Model Parameters
Frequency	0.55	17,990	12,720	5,270	10,910	7,080	O = E
Categorical (v1, all homorg)	0.46	21,554	17,718	3,836	-	-	O/E = 0 for homorganic, O/E = 1.22 otherwise
Categorical (v2, maj class)	0.74	10,230	-	-	4,790	5,440	O/E = 0 for homorganic, O/E = 1.21 otherwise
Soft Model (v1, all homorg)	0.71	11,744	7,908	3,836	-	-	O/E = 0 for adjacent ident, O/E = 0.75 for homorg, O/E = 1.22 otherwise
Soft Model (v2, maj class)	0.78	8,745	-	-	3,305	5,440	O/E = 0 for adjacent ident, O/E = 0.43 for homorg, O/E = 1.21 otherwise
Feature Model	0.78	8,978	5,123	3,855	3,855	5,123	S = 11.61, K = -3.90
Natural Classes	0.80	8,101	4,222	3,879	2,307	5,794	S = 13.73, K = -3.22

Table 5: Model comparison. Residual sum of squares is shown separately for homorganic (Hom), non-homorganic (Non-Hom), major class (Maj), and non-major class (Non-Maj) data.

We next compare residual sum of squares, first focusing on the differences in fit between homorganic and non-homorganic consonant pairs. Note first that much of the lack of fit for the frequency model comes from the homorganic pairs, which are subject to the OCP-Place constraint (residual SS = 12,720). Turning to the other models, we see that they perform equally well on the non-homorganic pairs (residual SS ≈ 3,850). This implies that the difference in overall fit is due to entirely to differences among homorganic pairs. As the reader will recall, nonhomorganic pairs were all assigned similarity zero and so all models which include an OCP constraint predict a uniform and high degree of overrepresentation for these pairs. The appropriateness of this prediction is seen from the relatively small contribution of the non-homorganic pairs to the residual sum of squares (especially in view of the great numbers of such combinations). Among the homorganic pairs, the natural classes model has nearly half of the residual sum of squares (residual SS = 4,222) of the soft model (residual SS = 7,908). Because the categorical model and soft model cannot differentiate homorganic similar consonants from homorganic dissimilar consonants, they are unable to fit the homorganic consonant data very well.

Finally, comparing the residual sum of squares for the major classes shows that the natural classes model still has far superior performance (residual SS = 2,307). This is particularly striking as the autosegmental models perform worse even though they have parameters that are specifically fit to just this subset of the data (residual SS = 3,312 for the soft model and residual SS = 4,790 for the categorical model). The natural classes model parameters come from an overall fit of all consonant pairs. The major classes data also highlight the advantage of the natural classes similarity model over the feature similarity model (residual SS = 3,855). The feature model does not appropriately relativize similarity to the different place classes, and thus is unable to appropriately account for the different degrees of sub-classification and cross-classification for different places of articulation.

4.5. Summary

The natural classes model outperforms all other models. We take this as evidence for both the stochastic constraint model and the natural classes similarity metric. The natural classes similarity model fit better than a similarity model based on a feature counting metric, showing that the representations of structured specification provide a superior basis for the similarity metric. Structured specification allows the similarity metric to be context sensitive, stable under relabeling, and synergistic. These properties are reflected in the patterns of cooccurrence. Thus, we conclude that structured specification is a superior representation of the phonological inventory.

The similarity models also fit better than the models based on an autosegmental analysis. The similarity models use the stochastic constraint which permits gradient degrees of cooccurrence. The autosegmental models only allow one or two degrees of variation and so do not account for the full range of data. Not only do the similarity models allow for variation in degree, they also predict this variation from the independent principles of similarity and distance. The autosegmental analysis provides no explanation of the variation in degree which it does allow.

The quantitative models, which contained soft constraints of some type, were far better than the categorical model. This shows that a quantitative model is required to accurately predict the patterns of cooccurrence in the Arabic lexicon. In fact, the frequency model, which contained no provisions for the OCP-Place constraint at all, accounted for nearly half of the variation in occurrence of consonant pairs and fit better than the most general categorical model (applied to all homorganic pairs). This shows that any reasonable model of cooccurrence must take into account the expected frequency of cooccurrence as a starting point.

5. Discussion

Pierrehumbert (1993) proposed a gradient OCP-Place constraint for Arabic which disfavors combinations of similar homorganic consonants. In implementing this proposal in detail we have raised a number of issues of general importance. In this section we discuss the broader implications of our analysis.

We propose that the OCP-Place constraint is a member of an abstract OCP constraint family of Universal Grammar which is sensitive to similarity. This constraint family can promote similar or dissimilar segments over a domain, along a particular feature dimension. In addition, the strength of the gradient effects can be parameterized on a language particular basis, as we showed in the previous section using the stochastic constraint model. In this section, we sketch out some implications of this proposal for phonological theory and universal grammar. First, we survey a range of similarity effects found cross-linguistically which may be amenable to analysis with this constraint family. Second, we discuss the natural classes similarity model based on structured specification, and the potential for cross-linguistic variation from a single similarity metric. Third, we discuss the nature of domain restrictions for this constraint and the implications of our analysis for transparency and non-local phonological phenomena. Finally, we consider the analysis of OCP-Place within current theories of generative grammar, highlighting the problematic nature of gradient, quantitative phenomena for these theories.

5.1. The role of similarity in phonology

We have proposed a general constraint family based on similarity. This constraint family relates similarity to lexical type frequency, and provides a formal analysis of cooccurrence in Arabic. We believe there are numerous cases in which phonological constraints can be formally described as conditions on similarity, either to increase (assimilation) or decrease (dissimilation) perceived similarity in surface patterns. Moreover, these patterns exhibit properties of gradience and variability, both across languages and within languages. The following examples give some indication of the type and range of data involved.

Phonotactic constraints on consonant place like those in Arabic are found in other Semitic languages (Bender and Fulass 1978, Buckley 1997, Greenberg 1950, Hayward and Hayward 1989, Koskinen 1964), as well as in English (Berkley 1994a), French (Plenat 1996), Javanese (Mester 1986), Ngbaka (Broe 1995), Russian (Padgett 1992), and other languages (Yip 1989). Formally related constraints are found for tone (Leben 1973, Goldsmith 1979, Odden 1986, Pierrehumbert and Beckman 1988) and for laryngeal features (Carré, Bordeau, and Tubach 1995; Ito and Mester 1986; MacEachern 1997; Steriade 1982). Thus, the patterns we observe in Arabic can be found in a variety of unrelated languages. Additional implications for phonological theory are likely to come to light when these languages are studied in greater detail using the tools developed here.

With respect to dissimilatory behavior, Leben (1973) initiated a line of active research in this area when he observed that in the African tone languages he studied, identical adjacent tones were impossible. That is, potential sequences of high and low tones such as HHL, HLL and HL were never contrasted: all melodies in which high preceded low had the same predictable syllabic alignment. Pierrehumbert and Beckman (1988) discuss the problems posed for OCP-Tone by the accent and intonation systems of Japanese and English. Adjacent identical tones are possible in both of these languages, but only if they are affiliated to different types of prosodic nodes (such as the syllable or word as opposed to the intonation phrase). A mechanical account of this phenomenon is to arbitrarily restrict the domain of comparison in evaluation of the OCP. We propose to view structural position as a dimension of similarity. This proposal requires a similarity metric which is more sophisticated than the one we present here. With a structurally sensitive similarity metric, we can exempt tones in different prosodic domains from the OCP in a motivated way. We return to this point in the discussion of transparency and non-local effects.

Odden (1986), surveying a number of African languages of the type that motivated the OCP originally, cites numerous exceptions to OCP-Tone. He notes that most of these concern sequences of identical but heteromorphemic tones. Odden argued that the OCP should be restricted to the role of a well-formedness constraint on underlying representations, whereas derived and surface forms were not subject to such a constraint. Again we observe that adjacent identical tones are possible just in case they are affiliated to different structural positions; it is just that here the relevant structure is morphological rather than prosodic. A structurally sensitive notion of similarity is also applicable in this case.

With respect to assimilation, the very common process of vowel harmony forces all vowels in some domain to become more similar on one or more dimensions. In numerous African languages for example, all vowels in the word must have either an advanced or retracted tongue root (ATR harmony). Like the dissimilation of homorganic consonants in Arabic, many assimilations are sensitive to vowels which share a particular featural dimension. Cole and Trigo (1989) discuss a variety of cases of 'parasitic' vowel harmony. In Yawelmani, for example, a target vowel is forced to agree with a trigger vowel in rounding just in case it already agrees in height. These harmony processes are believed to be categorical, though they have not been examined quantitatively.

There are also harmony systems which are known to have non-absolute but statistically robust patterns of cooccurrence. Carré, Bordeau, and Tubach (1995), employing an acoustically motivated classification of vowels based on perturbation theory (Mrayati, Carré, and Guerin 1988) show that French exhibits a statistical tendency toward vowel harmony. Karlsson (1971) reports that in Finnish, in addition to the much studied and almost absolute front-back harmony, there also exists a significant tendency to rounding harmony. The front-back harmony controls the vocalic quality of suffixes, while rounding harmony is confined to the stem.

Harmony may also occur among consonants, where it is also generally confined to an increase in similarity along one featural dimension (see Shaw 1991 for an overview). We note here that strident consonants are especially prone to harmony: Chumash, Quechua, Kinyarwandi, and Navaho all exhibit forms of 'sibilant harmony' among the stridents /s/ and /ʃ/, but none of the other coronals. Strident fricatives have two sources of aperiodic noise, arising first from the channel turbulence due to the narrowing of consonantal stricture, and second, the wake turbulence caused by the resulting jet encountering the obstruction of the teeth (Shadle 1985). Stridents thus have the greatest degree of acoustic salience of any fricative, and the preponderance of sibilant harmony can be explained as assimilation of features with the highest degree of perceived similarity.

In summary, we wish to highlight the explanatory potential of similarity in all of these processes, and to suggest that a similarity based description and quantitative analysis may reveal previously undiscovered patterns, as was found in the Arabic case. In the remainder of the section, we discuss the implications of our methods for phonological theory in general.

5.2. Structured specification

In our analysis we have found empirical evidence to support a classification system for the consonant inventory based on natural classes. This system has two advantages over other theories of feature specification. First, redundancy is encoded structurally in the hierarchy of natural classes. Redundant features are not omitted, but are identified as redundant by the hierarchy. Second, the hierarchy makes the system of contrasts in the inventory explicit, which takes into account the role of features in different sub-universes of the inventory on a context dependent basis.

Our analysis of Arabic showed the interaction of context dependence and redundancy to be relevant in determining the strength of OCP effects. We use full specification of all monovalent features, following Frisch (1996). Rather than leaving a redundant feature as a blank, the status of a redundant feature is encoded structurally in a lattice (Broe 1993). A redundant feature, such as [voiced] for sonorants, is implied by the subset relation among natural classes. The set of voiced segments is a superset of the set of sonorant segments. The addition of a feature which is totally redundant within a domain has no additional effect on similarity. These predictions for similarity effects have been supported in another analysis. Frisch (1996) applied the natural classes similarity metric to English speech error data, and found it superior to feature counting similarity metrics.

We have seen that the lattice representation provides a formal mechanism which allows similarity to be conditioned by contrastiveness within a sub-universe. Selecting a sub-universe within the segmental inventory requires no additional formal machinery. A sub-universe of the inventory is a sub-lattice in which some features may no longer be contrastive. We believe that relativizing to a sub-universe can be observed in many of the dissimilation and harmony processes reviewed in section 5.1, which often target segments which match along some dimension. In addition, the notion of a sub-universe applies to cases of inventory constraints in different prosodic positions, as in contextual neutralization and licensing (see Steriade 1995 for a recent discussion).

The representations of structured specification also make explicit the dimensions of contrast in the segment inventory. While the features play an important role in defining the natural classes, the lattice of natural classes crucially depends on the language particular inventory. The same set of features applied to two different inventories defines two different natural class hierarchies. Cross-linguistic differences in similarity effects may be found using the same (universal) similarity metric proposed here. Recent work by Padgett (1995a, 1995b) and Homer (1995) converge on analogous notions. Padgett's analysis of color harmony and Homer's analysis of nasal assimilation crucially depend on the contrastiveness of segmental targets, and not on the individual features involved. Traditional descriptions of these processes as operations involving feature spreading are unable to account for the failure of spreading to apply when segmental contrasts would not be maintained.

5.3. Transparency and non-local effects

There is recent work in the literature discussing problems with the autosegmental treatment of transparency (Cole and Kisseberth 1995a, Padgett 1995a). These authors develop alternatives to the autosegmental treatment of long-distance assimilations (such as vowel harmony). These models, both of which utilize Optimality Theory, assume that the harmonizing features are assigned to continuous spans of segments and that what were formerly characterized as transparency effects arise because of constraints on the expression of features. In Cole and Kisseberth (1995a, 1995b), conditions on the phonetic cooccurrence of features can, if highly ranked, block expression of features that would otherwise be apparent in the outcome. Padgett (1995a, 1995b) extends this suggestion by also bringing into play contrast constraints (of the type proposed by Lindblom 1983, Flemming 1995, and Silverman 1997). These proposals, though provocative, fail to incorporate one critical aspect of the autosegmental treatment. In the autosegmental treatment, transparency for assimilation and transparency for dissimilation were two sides of the same coin. But the Cole and Kisseberth and Padgett proposals offer no insights into the nature of transparency under dissimilation. Though our model is based primarily on dissimilation constraints, it offers an integrated treatment of assimilation and dissimilation. Our model manipulates the degree of similarity and the influence of distance on the availability of comparisons, thus providing common underpinnings for both assimilatory and dissimilatory phenomena.

The inadequacy of tier segregation to account for the effects of distance on the OCP shows that phonological phenomena can apply in a truly non-local manner, to non-adjacent objects. This finding reopens the issue of whether the OCP applies to root morphemes (consisting of consonants only) or to surface forms in which the vowels are already interleaved. Might not the vowels also prove to be translucent rather than transparent to the OCP? Since we have followed Greenberg and McCarthy in carrying out calculations on root morphemes only, addressing this issue is beyond the scope of the paper. However, we note that other cases are known in which OCP effects must apply to surface representations. Berkley (1994b) shows that English OCP effects apply across morpheme boundaries and are gradiently sensitive to distance. Recall that OCP-Tone effects are also sensitive to prosodic and morphological structure (Odden 1986, Pierrehumbert and Beckman 1988). We therefore conclude that the OCP is capable of applying over a fully structured surface representation, and that the domain of the constraint and effect of intervening material on the OCP are dependent on structural attributes as well as on the featural attributes employed in this paper. More generally, since OCP effects can apply in a non-local manner, through intervening material and phonological and morphological boundaries, other phonological processes can, in principle, do so as well.

Other linguistic phenomena, such as vowel harmony, which have been analyzed using autosegmental tier separation and transparency, should be examined in a statistically rigorous manner to determine if there is a systematic and predictable effect of distance and intervening material. In addition, similarity phenomena should be examined to determine the range of possible domain effects cross-linguistically.

5.4. OCP-Place in generative grammar

The quantitative model of OCP effects we propose incorporates patterns of sub and cross-classification between the major classes. The smoothly gradient nature of the cooccurrence restriction provides further support for a model which bases the constraint on a gradient phenomenon like similarity. We have shown that the autosegmental account cannot explain the detailed patterns of cooccurrence. Extending the autosegmental account to include these regularities requires feature cooccurrence rules of such complexity that they would in essence duplicate the computation of similarity. More recent analyses proposed in Yip (1989), Padgett (1995c), and Lamontagne (1992) to account for sub-classification of coronal consonants in Arabic and other languages are indeed a step in this direction. However, these approaches still rely on a categorical restriction, and so cannot account for the range of variation found in the Arabic data. We claim the cooccurrence restrictions in Arabic get stronger as more features match because the effect is based on similarity.

An analysis of OCP-Place based on Optimality Theory ultimately will suffer from the same problems as the autosegmental account. While OCP-Place in Arabic has not been analyzed, dissimilarity phenomena have been examined, and the two accounts we are aware of use dissimilatory constraint families. In one approach, dissimilarity is based on conjunctions of markedness constraints (e.g. *Place(labial)^2, as in Alderete 1997). The other approach uses conjunctions of features in an explicit similarity constraint hierarchy to invoke dissimilarity effects (MacEachern 1997); an account in spirit much like our own. These constraint families suffer from the same combinatorial explosion of discrete sub-cases as the autosegmental account. More importantly, the generally accepted formulation of Optimality Theory predicts a single winning candidate from an input candidate, much like an autosegmental constraint either admits or rules out a structure. While the constraints in Optimality Theory are violable due to constraint ranking, the input to output mapping is still categorical. Optimality Theory is unable to allow for gradient acceptability of a phonological type, and thus is unable to make predictions about the statistical distribution of forms. In the analysis of Arabic for example, a single consonant pair would be either acceptable or unacceptable in all cases (Berkley 1994b, Pierrehumbert 1994, Plenat 1996).

Anttila (1996) and Hayes and MacEachern (1996) present extensions of OT in which statistical variable constraint ranking is used to model statistically variable outcomes. These proposals maintain the distinction drawn in OT in general between the lexicon (the set of input forms) and the grammar (which is formalized using constraint rankings). As a result, all input forms which share a given phonological characterization are predicted to share the same pattern of variation in the outcome. This approach can reproduce patterns which were treated in the previous literature using variable rules, whether allophonic or allomorphic, applying across the board to forms with certain properties. It does not provide a treatment of underrepresentation or overrepresentation in the lexicon. Given the separation between the lexicon and the grammar in this architecture, no stochastic augmentation of the grammar can bias the way that the lexical forms sample the space of possible forms.

Probabilistic rules like the stochastic constraint model have been proposed for cases of phonetic implementation, which is often divorced from phonology proper (e.g. Keating 1984, Pierrehumbert and Beckman 1988). The phonetics-phonology division was motivated by a desire to separate the symbolic phonological system from the probabilistic and gradient nature of real speech. We conclude that the OCP reflects implicit linguistic knowledge about the possible words in a language so gradient phenomena must be incorporated within phonology proper (Pierrehumbert 1994). It has also been found that rules of phonetic implementation are language specific, which undermines the existence of a dividing line between phonology and phonetics (Pierrehumbert 1990). We believe that unification of phonological and phonetic knowledge into a single system allows gradient effects within the two domains to be accounted for with analogous mechanisms. Our proposal for a quantitative phonology is a step in this direction.

6. Conclusion

In this paper, we have shown that linguistic formalism must be expanded to include quantitative constraints, in order to model gradient cooccurrence constraints in Arabic. Incorporating these constraints into phonology requires replacing the general use of categorical constraints with stochastic constraints. We have proposed a particular model of a stochastic constraint, based on the logistic function. The logistic provides a smoothly gradient function which can be used to model positive or negative constraints based on a quantitative parameter. The coefficients of the logistic can be altered to model either categorical or highly gradient effects.

We have argued for a hierarchical representation of the segment inventory. In this representation, the natural classes of phonemes play a key role in encoding feature contrastiveness and redundancy. We have provided evidence for the influence of natural classes in the determination of similarity between Arabic consonants. In addition, we have eliminated the problematic use of contrastive underspecification from the previous similarity model of the Arabic inventory.

Together, the use of stochastic constraints based on the logistic and the use of structured specification as a representation of the phoneme inventory demonstrate that quantitative effects can be modeled in a formally precise manner. In addition, the quantitative patterns have revealed additional insights into the general issues of transparency and underspecification. The data analyzed here are naturally occurring and fully reflective of linguistic competence, and therefore extend the range of data which must be accounted for by any theory of grammar.

References

Alderete, John: 1997, 'Dissimilation as local conjunction', in K. Kiyomi (ed.), Proceedings of the North East Linguistics Society 27, pp. 17-32.

Anttila, Arto: 1996, 'Deriving variation from grammar', Ms. submitted for publication, Stanford University, Stanford, CA.

Archangeli, Diana: 1984, Underspecification in Yawelmani phonology and morphology, unpublished Ph.D. Dissertation, MIT.

Archangeli, Diana: 1988, 'Aspects of underspecification theory', Phonology 5, 183-208.

Bender, M. Lionel and Hailu Fulass: 1978, Amharic verb morphology. Michigan State University African Studies Center, East Lansing, MI.

Berkley, Deborah Milam: 1994a, 'Variability in obligatory contour principle effects', in Papers from the 30th regional meeting of the Chicago Linguistics Society, University of Chicago Press, Chicago, pp. 1-12.

Berkley, Deborah Milam: 1994b, 'The OCP and gradient data', Proceedings of FLSM V. Studies in the Linguistic Sciences 24(1/2), 59-72.

Broe, Michael: 1993, Specification theory: the treatment of redundancy in generative phonology, upublished Ph.D. dissertation, Edinburgh.

Broe, Michael: 1995, 'Specification theory and Ngbaka co-occurrence constraints', paper presented at the Mid-Continental Workshop on Phonology, Columbus, OH.

Broe, Michael: 1997, 'Stochastic constraints in phonological analysis', paper presented at Indiana University, Bloomington, IN.

van den Broeke, Marcel and Louis Goldstein: 1980, 'Consonant features in speech errors', in V. Fromkin (ed.), Errors in linguistic performance: slips of the tongue, ear, pen, and hand. Academic Press, New York.

Buckley, Eugene: 1997, 'Tigrinya root consonants and the OCP', paper presented at the 16th annual Penn Linguistics Colloquium, to appear in Penn Working Papers in Linguistics.

Bybee, Joan and Dan Slobin: 1982, 'Rules and schemas in the development and use of English past tense', Language 59, 251-270.

Carré, R., M. Bourdeau, and J-T. Tubach: 1995, 'Vowel-vowel production: the distinctive region model DRM and vowel harmony', Phonetica 52.

Chomsky, Noam and Morris Halle: 1968, The sound pattern of English, Harper and Row, New York.

Clements, G. N.: 1985, 'The geometry of phonological features', Phonology Yearbook 2, 225-252.

Cole, Jennifer and Charles W. Kisseberth: 1995a, 'An optimal domains theory of harmony', Proceedings of FLSM V. Studies in the Linguistic Sciences 24(1/2), 101-114.

Cole, Jennifer and Charles W. Kisseberth: 1995b, 'Paridoxical strength conditions in harmony systems', in J. Beckman (ed.), Proceeings of the North East Linguistics Society 25, pp. 17-30.

Cole, Jennifer and Loren Trigo: 1989, 'Parasitic harmony', in H. van der Hulst and N. Smith (eds.), Features, segmental structure and harmony processes, Foris, Dordrecht, pp. 19-38.

Cowan, J.: 1979, Hans Wehr: a dictionary of modern written Arabic, Otto Harrasowitz, Wiesbaden, Germany.

Davey, B. A. and H. A. Priestley: 1990, Introduction to lattices and order, Cambridge University Press, Cambridge.

Devore, J. L.: 1987, Probability and statistics for engineering and the sciences, 2nd ed., Brooks/Cole, Monterey, CA.

Eriksen, C. and D. Schutze: 1978, 'Temporal factors in visual information processing: a tutorial review', in J. Requin (ed.), Attention and Performance VII, Earlbaum, Hillsdale, NJ.

Flemming, Edward: 1995, Auditory representations in phonology, unpublished Ph.D. dissertation, UCLA.

Frisch, Stefan: 1996, ‘Similarity and frequency in phonology’, unpublished Ph.D. dissertation, Northwestern University.

Frisch, Stefan, Michael Broe, and Janet Pierrehumbert: 1995, The role of similarity in phonotactic constraints, Ms. Northwestern University, Evanston, IL.

Gluck, M. and G. Bower: 1988, ‘From conditioning to category learning: an adaptive network model’, Journal of Experimental Psychology: General 117, 227-247.

Goldsmith, John: 1979, Autosegmental phonology, Garland: New York.

Goldstone, Robert: 1994, ‘Similarity, interactive activation, and mapping’, Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 3-28.

Goodman, N.: 1972, ‘Seven strictures on similarity’, in N. Goodman (ed.), Problems and projects, Bobbs-Merrill, New York.

Greenberg, Joseph: 1950, ‘The patterning of root morphemes in Semitic’, Word 5, 162-181.

Hayes, Bruce and Margaret MacEachern: 1996, ‘Folk verse form in English’, Ms., UCLA. Available from the Rutgers Optimality Archive (http://ruccs.rutgers.edu/roa.html).

Hayes-Roth, Barbara and Frederick Hayes-Roth: 1977, ‘Concept learning and the recognition and classification of exemplars’ Journal of Verbal Learning and Verbal Behavior 16, 321-338.

Hayward, K. and R. Hayward: 1989. ‘“Guttural”: arguments for a new distinctive feature’, Transactions of the Philological Society 87, 179-193.

Homer, Molly: 1995, ‘Segment sets account for opacity in Applecross Gaelic’, paper presented at the Midcontinental Workshop in Phonology, Ohio State University, Columbus, OH.

Ito, Junko and R. Armin Mester: 1986, ‘The phonology of voicing in Japanese’, Linguistic Inquiry 17, 49-73.

Jakobson, Roman, Gunnar Fant, and Morris Halle: 1952, Preliminaries to speech analysis, MIT Press, Cambridge.

Jespersen, Otto: 1939, Growth and structure of the English language, Blackwell, Oxford.

Karlsson, F.: 1971, ‘Finskans rotmorfemstruktur: en generativ beskrivning’, Turun Yliopston Fonetiikan Laitoksen Julkaisuja No 10, Turku.

Keating, Patricia: 1984, ‘Phonetic and phonological representation of stop consonant voicing’, Language 60, 286-319.

Koskinen, Kalevi: 1964, ‘Kompatibilitat in den dreikonsonantigen hebraishen wurzeln’, Zeitschrift der Deutschen Morgenlandischen Gesellschaft 144, 16-58.

Lane, E.: 1863, An Arabic-English lexicon, Williams and Norgate, London.

Lamontagne, Gregory: 1992, Syllabification and consonant cooccurrence restrictions, unpublished Ph.D. dissertation, University of Massachusetts-Amherst.

Leben, William: 1973, Suprasegmental phonology, unpublished PhD dissertation, MIT.

Liberman, Mark: 1994, Phonological optionality in Latin clitics, Penn review of linguistics 18, 87-102.

Lightner, Theodore: 1973, ‘Against morpheme structure rules and other things’, in Issues in phonological theory, Mouton, The Hague, pp. 53-60.

Lindblom, Bjorn: 1983, ‘Economy of speech gestures’, in P. MacNeilage (ed.), The production of speech, Springer-Verlag, New York, pp. 217-245.

MacEachern, Margaret: 1997, ‘Laryngeal similarity effects in Quechua and Aymara’, Proceedings of the West Coast Conference on Formal Linguistics 15, 337-351.

Massaro, Dominic: 1970, 'Retroactive interference in short-term recognition memory for pitch', Journal of Experimental Psychology 83, 32-39.

McCarthy, John: 1979, Formal problems in Semitic phonology and morphology, Garland: New York.

McCarthy, John: 1986, 'OCP effects: gemination and antigemination', Linguistic Inquiry 172, 207-263.

McCarthy, John: 1988, 'Feature geometry and dependency: a review', Phonetica 43, 84-108.

McCarthy, John: 1994, 'The phonetics and phonology of Semitic pharyngeals', in P. Keating (ed.), Papers in laboratory phonology III, Cambridge University Press, Cambridge, pp. 191-283.

Mester, R. Armin: 1986, Studies in tier structure, unpublished PhD dissertation, University of Massachusetts-Amherst.

Mrayati, M., R. Carré and B. Guerin: 1988, 'Distinctive regions and modes: a new theory of speech production', Speech Communication 7, 257-286.

Odden, David: 1986, 'On the obligatory contour principle', Language 62, 353-83.

van Oijen, Brit: 1994, The processing of vowels and consonants, Holland Academic Graphics: The Hague.

Padgett, Jaye: 1992, 'OCP subsidiary features', in Proceedings of the North East Linguistics Society 22.

Padgett, Jaye: 1995a, 'Partial class behavior and nasal place assimilation', in Proceedings of the Arizona phonology conference: workshop on features in Optimality Theory. Coyote Working Papers, University of Arizona, Tucson.

Padgett, Jaye: 1995b, 'Feature classes', Ms. University of California, Santa Cruz, CA. Available from the Rutgers Optimality Archive (http://ruccs.rutgers.edu/roa.html).

Padgett, Jaye: 1995c, Stricture in feature geometry, CSLI Publications, Stanford, CA.

Pierrehumbert, Janet: 1990, 'Phonological and phonetic representation', Journal of Phonetics 18, 375-394.

Pierrehumbert, Janet: 1993, 'Dissimilarity in the Arabic verbal roots', in Proceedings of the North East Linguistics Society 23, pp. 367-381.

Pierrehumbert, Janet: 1994, 'Syllable structure and word structure', in P. Keating (ed.) Papers in laboratory phonology III, Cambridge University Press, Cambridge, 168-190.

Pierrehumbert, Janet: In press, 'Comments on the functionalist and formalist phonology papers', in Proceedings of the 1996 Milwaukee Conference on Formalism and Functionalism in Linguistics.

Pierrehumbert, Janet and Mary Beckman: 1988, Japanese Tone Structure, Linguistic Inquiry Monograph 15, MIT Press, Cambridge.

Pisoni, David: 1973, 'Auditory and phonetic memory codes in the discrimination of consonants and vowels', in Perception and Psychophysics 13, pp. 253-260.

Plenat, Marc: 1996, 'De l'interaction des contraints: une etude de case', in J. Durand and B. Laks (eds.), Current trends in phonology: Models and methods, University of Salford Press, Salford, pp. 585-617.

Prince, Alan and Paul Smolensky: 1993, Optimality theory: constraint interaction in generative grammar, Technical reports of the Rutgers University Center for Cognitive Sciences 2, RUCCS, New Brunswick, NJ.
Repp, Bruno: 1984, 'Categorical perception: Issues, methods, findings', in N. J. Lass (ed.), Speech and language: Advances in basic research and practice, vol. 9, Academic Press, New York.
Shadle, Christine: 1985, The acoustics of fricative consonants, unpublished PhD Dissertation, MIT.
Shaw, Patricia: 1991, 'Consonant harmony systems', in C. Paradis and J.-F. Prunet (eds.), Phonetics and Phonology 2: the special status of coronals, Academic Press, San Diego, CA.
Silverman, Dan: 1997, Phasing and recoverability, Garland, New York.
Stanley, Richard: 1967, 'Redundancy rules in phonology', Language 43, 393-436.
Stemberger, Joseph: 1991, 'Radical underspecification in language production', Phonology 8, 73-112.
Steriade, Donca: 1982, Greek prosides and the nature of syllabification, Unpublished Ph.D. dissertation, MIT.
Steriade, Donca: 1987, 'Redundant values', in Papers from the 23rd regional meeting of the Chicago Linguistics Society, University of Chicago Press, Chicago, pp. 339-362.
Steriade, Donca: 1995, 'Underspecification and markedness', in J. Goldsmith (ed.), The handbook of phonological theory, Blackwell, Oxford, pp. 114-174.
Stevens, Kenneth and Samuel J. Keyser: 1989, 'Primary features and their enhancement in consonants', Language 65, 81-106.
Tversky, Amos: 1977, 'Features of similarity', Psychological Review 84, 327-352.
Yip, Moira: 1989, 'Feature geometry and cooccurrence restrictions', Phonology 6, 349-374.

Appendix: The consonants of Arabic

Arabic has a large set of phonemic consonants, including most consonants of English, as well as a set of 'guttural' consonants, which are produced in the rear of the vocal tract, between the velum and the larynx. The major places of articulation for Arabic consonants are:

[labial]: /b, f, m/
[coronal]: /t, d, T, D, θ, ð, s, z, S, Z, ʃ, l, r, n/
[dorso-guttural]: /T, D, S, Z, k, g, q, χ, ʁ, ħ, ʕ, h, ʔ/

Following McCarthy (1994) these features represent orosensory target regions, rather than active articulators. We provide evidence for three place of articulation categories in Arabic based on the cooccurrence restrictions discussed in section 2.3.

The particular features of the guttural consonants (/χ/, /ʁ/, /ħ/, /ʕ/, /h/, /ʔ/) given below are based on McCarthy's (1994) extensive review. The pharyngeal /ʕ/ and larygeal /ʔ/ are usually accompanied by glottalization, and so the laryngeal feature [constricted glottis] is used instead of [voice]. There is another set of consonants, called the ‘emphatics’, which deserve special mention. These are the coronal obstruents /T/, /D/, /S/, and /Z/. These consonants are similar to the familiar English consonants /t/, /d/, /s/, and /z/, but they contain a second constriction, at the uvula, in addition to the one at the alveolar ridge. Thus, these consonants have both [coronal] place of articulation and [dorso-guttural] place of articulation.

In this paper, we use all monovalent feature values for consistency in the computation of similarity (see section 3).

i. Labials

	Place	Manner					Laryngeal	
	labial	son	nas	obs	stop	fric	voice	vless
b	+			+	+		+	
f	+			+		+		+
m	+	+	+		+		+	

son = [sonorant], nas = [nasal], obs = [obstruent], fric = [fricative], vless = [voiceless]

ii. Coronals

	Place								Manner							Laryngeal	
	cor	inter	dent	alv	pal	dors-gutt	gutt	lo-uv	son	lat	rhot	nas	obs	stop	fric	voice	vless
t	+		+										+	+			+
d	+		+										+	+		+	
T	+			+		+	+	+					+	+			+
D	+			+		+	+	+					+	+		+	
θ	+	+											+		+		+
ð	+	+											+		+	+	
s	+		+										+		+		+
z	+		+										+		+	+	
S	+			+		+	+	+					+		+		+
Z	+			+		+	+	+					+		+	+	
ʃ	+				+								+		+		+
l	+								+	+						+	
r	+								+		+					+	
n	+								+			+		+		+	

cor = [coronal], inter = [interdental], dent = [dental], alv = [alveolar], pal = [palatal], dors-gutt = [dorso-guttural], dors = [dorsal], lo-uv = [low-uvular], lat = [lateral], rhot = [rhotic]

iii. Dorso-Gutturals

	Place										Manner				Laryngeal		
	dors-gutt	dors	gutt	vel	hi-uv	lo-uv	phar	lar	cor	alv	app	obs	stop	fric	voice	vless	conglot
T	+		+			+			+	+		+	+			+	
D	+		+			+			+	+		+	+		+		
S	+		+			+			+	+		+		+		+	
Z	+		+			+			+	+		+		+	+		
k	+	+		+								+	+			+	
g	+	+		+								+	+		+		
q	+	+			+							+	+			+	
χ	+	+			+						+					+	
ʁ	+	+			+						+				+		
ħ	+		+				+				+					+	
ʕ	+		+				+				+						+
h	+		+					+			+					+	
ʔ	+		+					+			+						+

gutt = [low-guttural], vel = [velar], hi-uv = [high-uvular], phar = [pharyngeal], lar = [laryngeal], app = [approximant], conglot = [constricted glottis]