

# Locality and Feature Specificity in OCP Effects: Evidence from Aymara, Dutch, and Javanese\*

Peter Graff and T. Florian Jaeger  
Massachusetts Institute of Technology and University of Rochester

## 1 Introduction

We compare different sets of constraints on the co-occurrence of similar consonants within words (the Obligatory Contour Principle, henceforth OCP; generalized from Leben 1973) utilizing Maximum Likelihood fitted logistic regression. We present evidence i) that the featural similarity of consonants matters in predicting whether a certain combination of consonants occurs as a word in a given language ii) that OCP constraints need to reference phonological features individually and iii) that OCP constraints need to be able to distinguish among feature matches between adjacent or V-adjacent (i.e., only separated by a vowel) consonants and consonants separated by another consonant (e.g., Pierrehumbert 1993). Our findings support formulations of OCP constraints that treat individual features as free parameters in the similarity computation (e.g., Coon & Gallagher 2007). Similarly, our findings do not support models of phonological similarity that tie the effect of different feature matches to the organization of phonological contrast within a given language (Frisch *et al.* 2004). The statistical approach we employ has several advantages over modeling observed-over-expected ratios and allows us to systematically compare different models of the OCP—relative to the complexity introduced by their constraint sets—across the three languages.

## 2 Preliminaries

### 2.1 Similarity avoidance

Similar adjacent elements are dispreferred in many cognitive domains, including phonology (e.g. Rose & Walker 2004, Coetzee & Pater 2008, Gallagher 2009 for some recent studies). Various languages exhibit a bias against similar consonants within a word, such that fewer similar sounds co-occur than would be expected *a priori*. This suggests that similarity avoidance shapes the phonological lexicon.

Phonological phenomena attributed to similarity include not only the gradient under-attestation of certain consonant combinations in lexical roots in comparison to the co-occurrences one would expect from their independent probabilities (Coetzee & Pater 2008, Frisch *et al.* 2004) but also the complete absence of certain combinations of consonants, whether adjacent or not (e.g. McCarthy 1986, MacEachern

---

\*We would like to thank Adam Albright, Edward Flemming, Gillian Gallagher, Jeremy Hartman and especially Donca Steriade for discussion and comments throughout the different stages of this project. We would also like to thank David Somach and Luis Peña for invaluable help with data preparation as well as audiences at the UCLA-UC Berkeley Conference on the Languages of Southeast Asia, CUNY22 and CLS45 for helpful feedback. All remaining errors are our own.

1999, Gallagher 2009). In this study we focus on the selection of certain words over others from a set of logically possible lexical items. We show that this selection is in part driven by co-occurrence restrictions on similar consonants.

## 2.2 Generative potential

Prior to processing or communicative considerations one might expect that any given language makes use of its full generative potential given an inventory of phonemes and templatic restrictions on possible shapes of words (cf. the related notion of Richness of the Base, Prince & Smolensky 2004). This means that there is no inherent generative reason to assume that the lexicon of a language that allows for words of the general shape CVCVC should not contain words exhibiting every possible permutation of consonants in the C-slots of this template. In fact, one might expect a language to have as many words as possible of a particular allowed shape, given that word-length otherwise has to increase to maintain non-homophonous words. Nonetheless, the majority of combinatorial possibilities are unattested. The goal of this paper is to test whether the attestation of particular words over others is in part driven by feature-based constraints that penalize the co-occurrence of similar consonants within words.

## 2.3 Languages

Our long-term project is the investigation of OCP effects in a large number of languages, the objective being to contribute in this way to a better understanding of universal and language-specific aspects of phonological similarity. We begin our investigation with three unrelated languages that have been reported to exhibit under-representation of similar consonants within words: Aymara (MacEachern 1999), Dutch (Boll-Avetisyan & Kager 2008), and Javanese (Mester 1986). The data for our study are corpora of logically possible ordered consonant triplets. We generated one corpus per language containing every logically possible triplet by permuting all consonants in the inventory of the language. This gives a total of (number of consonants)<sup>3</sup> triplets per language. We then chose the most populated tri-consonantal CV-template for each language (CVCVC in Javanese and Dutch; CVCCV in Aymara) and annotated each logically possible ordered consonant triplet for 1 (=attested) if there is at least one word fitting the CV-template containing these three consonants in order in the language or 0 (=unattested) otherwise. An example is given in Table 1.

**Table 1:** Translating the lexicon into attested and unattested triplets (Aymara)

Lexicon	CV-Template	Triplet	Attested
/tʃaxta/	CVCCV	/tʃxt/	1
/tʃaxwa/	CVCCV	/tʃxw/	1
/tʃaxwi/	CVCCV		
		/tʃjk/	0

Template attestation is based on de Lucca (1987) for Aymara, Baayen *et al.* (1993) for Dutch and Uhlenbeck (1978) for Javanese. Table 2 shows the attested-to-unattested ratio for the three corpora.

**Table 2:** Corpora of logically possible triplets employed in our studies

Language	Template	# of consonants	Possible triplets	% Attested
Aymara	CVCCV	26	17,576	3%
Dutch	CVCVC	23	12,167	11%
Javanese	CVCVC	21	9,261	20%

### 3 Methods

In order to investigate whether the same set of grammatical constraints predicts triplet attestation in Aymara, Dutch and Javanese, we utilize Maximum Likelihood fitted logistic regression. This approach has a variety of advantages over the current gold-standard in phonological research, observed-over-expected ratios (in terms of simple rank comparisons or log-linear models, Coetzee & Pater 2008; or  $\chi^2$ -tests, Mester 1986, Pierrehumbert 1993). Logistic regression is a generalized linear model predicting the probability of a binary variable having one value over another (in our case attestation over non-attestation) from a set of predictors. The model finds the best weight (coefficient) for each predictor, such that the likelihood of the observed data given the model (data likelihood) is maximized.

Compared to the use of observed-over-expected ratios (O/E) in previous work (Coetzee & Pater 2008, Frisch *et al.* 2004), logistic regression has a number of advantages that we exploit in our analysis. First, logistic regression allows us to control for different non-similarity related effects in triplet attestation. The “expected” probability of a consonant may be affected by several variables, such as positional occurrence restrictions and identity effects (see section 4). Unlike the O/E approach used in previous work, logistic regression not only controls for multiple effects, but also provides measures of each of the controls’ effect on the predicted outcome (attestation of a logically possible triplet). Second, logistic regression provides a statistically sound way of comparing the success of different models by systematically weighing data coverage against model complexity. For large data sets like ours, differences in the data likelihood of nested models are approximately  $\chi^2$ -distributed (Agresti 2002:86). This is convenient, as it facilitates systematic comparison of nested models using  $\chi^2$ -tests over differences in the models’ data likelihoods given differences in their complexity (differences in the number of coefficients to be fitted). Here we take advantage of model comparison to compare different theories of how similarity avoidance affects the phonological lexicon. Third, as we illustrate below, logistic regression facilitates effect size comparison across data sets and makes it possible to compare the strength of OCP effects across languages.

## 4 Controls

### 4.1 Occurrence restrictions

The attestation of a particular set of triplets is not solely driven by co-occurrence restrictions on similar sounds. Simple occurrence restrictions, such as gradient and categorical phonotactics, also apply. For example, triplets beginning with a frequent consonant are more likely to be attested than triplets beginning with a rare one (otherwise the consonant would not be frequent in that position). Getting a good grasp

of occurrence restrictions is vital for determining the effect of co-occurrence restrictions. This is because the non- or under-attestation of certain forms may otherwise be wrongly attributed to co-occurrence restrictions.

Consider the example of word-final /w/ in Javanese. The template /tVmVw/ is unattested in Javanese. Without knowing that Javanese lacks any word ending in /w/, it is possible to wrongly attribute the non-attestation to an OCP effect. We encode occurrence restrictions as predictors of positional frequency to capture both gradient and categorical generalizations about the phonotactics of different positions. These predictors are defined as follows.

- (1)  $\text{Occurrence}_{C_x \text{ in } C_x}$   
Specified for the number of attested triplets that have the  $x^{\text{th}}$  consonant in position  $x$ .

This means that any triplet—attested or not—that starts with e.g. /t/ is specified for the total number of attested triplets with /t/ in the initial position. In the case of categorical absence of a consonant in a certain position, such as Javanese /w/ in  $C_3$ , the relevant occurrence factor is specified zero. By including these predictors in the logistic regression, we can assess the partial effect of OCP constraints independent of occurrence restrictions. In Aymara, we face an additional complication. Since our triplets contain consonant clusters (CVCCV), it is insufficient to encode frequency of  $C_2$  and  $C_3$  separately. The frequency of a consonant is likely to be crucially dependent on its neighboring consonant, since occurrence of consonants in clusters is much more rigidly governed cross-linguistically. We therefore include an additional control predictor in Aymara only, modeling the frequency of the  $C_2C_3$  cluster.

- (2)  $\text{Occurrence}_{C_2C_3 \text{ in } C_2C_3}$   
Specified for the number of attested triplets that have the cluster  $C_2C_3$  in positions  $C_2$  and  $C_3$ .

## 4.2 Identity

Several languages that show strong OCP effects allow for total identity between consonants co-occurring within words (MacEachern 1999). While e.g. /k'ap'i/ and other words with two ejectives might be prohibited, /k'ak'i/ is allowed as the two ejective stops are completely identical. MacEachern (1999) terms this the identity effect. Generally, identity between segments in close proximity is quite common cross-linguistically and might in fact be preferred if one considers cases such as those identified by Zuraw (2002) as aggressive reduplication.

As with occurrence restrictions, we need to ensure that a potential preference for identity can be captured by our model. Take, for example, the following four groups of templates that start with labial stops. Table 3 shows their respective frequencies in Javanese.

In spite of the intermediate frequency of /bVpVC/, the generalization seems clear. In Javanese, labial plosives are much more likely to co-occur in  $C_1$  and  $C_2$  if they are identical. If we do not give the model the opportunity to fit a probabilistic effect of identity, we run the risk of predicting more variation in the co-occurrence

**Table 3:** Illustration of a gradient identity effect in Javanese

Template	# Attested templates	Identity
/bVpVC/	5	N
/bVbVC/	10	Y
/pVbVC/	1	N
/pVpVC/	9	Y

of labials than necessary. We define the following identity predictors to include as controls in the model.

- (3) Identity $C_xC_y$   
Specified 1 if  $C_x$  and  $C_y$  are identical.

## 5 Comparing accounts of OCP effects in the phonological lexicon

### 5.1 Establishing OCP effects

In this section, we establish that similarity between consonants within a triplet significantly improves the coverage of our models. We start with a simple definition of co-occurrence restrictions and propose a constraint that simply counts the feature matches between each pair of adjacent or V-adjacent (i.e., only separated by a vowel) consonants. It is important to note here that we only count matching positive featural specifications as feature matches. The feature systems we use are based on MacEachern (1999), Booij (1995) and Adisasmito-Smith (2004) for Aymara, Dutch, and Javanese, respectively (see Appendix, section 8).

To test whether feature matches affect triplet attestation, we define a single predictor that sums the feature matches between each pair of consonants that are adjacent or V-adjacent. This predictor encodes how similar the consonants in each theoretical word are to each other. The computation is strictly local, as feature matches between  $C_1$  and  $C_3$  are not counted. Note further that this predictor does not distinguish among individual features, in that a feature match counts the same whether it is due to, e.g., a [+labial] or a [+alveolar] specification. We call this predictor *Local Matches*.

- (4) Local Matches  
The sum of feature matches between  $C_1/C_2$  and  $C_2/C_3$ .

In the resulting models, the *Local Matches* predictor is always fitted with a negative coefficient, as shown in table 4.

**Table 4:** *Local Matches* predictors in the three languages

Language	Coefficient (logged odds)	Standard Error	p
Aymara	-0.063	0.039	< 0.12
Dutch	-0.178	0.036	< 0.0001
Javanese	-0.079	0.028	< 0.006

Every additional feature match decreases the predicted probability of a triplet being attested in the language and suggests that similarity has the predicted effect on the shape of the phonological lexica of the three languages; namely, words with more feature matches are less likely to occur. The *Local Matches* predictor significantly improves the data likelihood in Dutch ( $\chi^2(1) = 24.84, p < .0001$ ) and Javanese ( $\chi^2(1) = 7.69, p < .006$ ) and is non-significant only in Aymara ( $\chi^2(1) = 2.58, p < .11$ ).

We conclude that even this very crude notion of similarity improves data likelihood significantly in a model predicting the selection of certain consonant triplets over others in two of the three languages. In all three languages the effect of local feature matches goes in the predicted direction: predicted probability of attestation is decreased.

We may now ask whether a single predictor counting the number of feature matches within a word is sufficient to account for similarity effects or whether the grammar incorporates a more fine-grained distinction among the individual similarity-related properties of triplets. Logistic regression allows us to ask this question relative to the complexity introduced by more fine grained models of similarity. In the next subsection, we ask whether there is evidence that the extent of similarity avoidance differs for different phonological features.

## 5.2 Feature specificity and the OCP

Recent work on the OCP suggests that features may differ in the extent to which they influence the phonological similarity metric. Coon & Gallagher (2007) find that features such as [+strident] and [+ejective] trigger greater OCP effects than others. It is also known that non-coronal place features (Pierrehumbert 1993, Coetzee & Pater 2008) as well as marked laryngeal features (MacEachern 1999, Gallagher 2009) tend to trigger OCP effects of varying strengths (i.e., the extent to which OCP violating structures are under-attested) in a variety of languages.

In order to test whether OCP effects of different place, laryngeal, and manner features are weighted differently, we fit a feature-specific model by adding a separate OCP predictor for each feature. This allows for the effect of each type of feature match to differ, but also adds a considerable amount of complexity to the model. Instead of encoding similarity in terms of a single *Local Matches* predictor, we now need to fit one predictor for every feature. The question is whether the increase in data coverage justifies feature-specific OCP constraints as part of the grammar.

Certain features that form part of the standard distinctive feature systems utilized in phonology are inherently correlated. The feature [strident], for example, is only contrastive in coronal fricatives. If we know a sound is [+strident] we also know the sound is [+coronal]. This introduces a problem for the logistic regression model. Correlated (i.e., non-orthogonal or partially redundant) predictors in a model can result in collinearity. Collinearity can lead to inflated standard error estimates and hence loss of power. Collinearity also makes it harder to reliably interpret individual effect sizes and effect directions.

Indeed, a full feature system would lead to serious multicollinearity in the feature-specific model (several variance inflation factors, VIFs  $> 2.5$  in Aymara and Javanese,  $> 5.5$  in Dutch, where  $VIF > 4$  is a common cut-off point (see Fox 1991;

see also Hepworth *et al.* 2007 for further detail).

For now, we make the, admittedly conservative, choice of excluding all manner features whose absolute correlation with any other feature is  $> 0.2$ , in order to reliably compare effect sizes across languages. Furthermore, we split [coronal] place into [alveolar] and [post-alveolar] to avoid having to distinguish among coronals with highly correlated features such as [anterior]. The resulting feature system includes all major place features, marked laryngeal features ([+breathy] in Javanese, [+voiced] in Dutch, and [+longVOT]<sup>1</sup> in Aymara) and non-collinear manner features such as , [+rhotic], [+lateral], [+nasal] and [+continuant]. See Table 6 - 8 in Appendix section 8 for the complete feature systems. These feature systems successfully removed all multicollinearity (all VIFs  $< 1.55$  ). We view these feature systems and the procedure we used to derive them as an acceptable first step to assess the partial OCP effects of different phonological features. However, it is possible that, for example, the feature [strident] would have improved the models' overall prediction. Future work will investigate a wide range of possible feature systems to test the extent to which the effects we report below depend on the particular feature systems we have chosen here.

Based on these feature systems, the number of similarity predictors is thus 11 (Dutch) or 12 (Aymara and Javanese; see Appendix section 8). We define the following set of feature-specific OCP predictors.

- (5) OCP-[+feature]  
 +1 for each pair of V-adjacent or adjacent consonants that match for this feature.

The predictor Local OCP-[+labial] is specified 1 in the case of a match between  $C_1$  and  $C_2$  or  $C_2$  and  $C_3$  for the feature [+labial]. If all three consonants match for [+labial] the predictor is specified 2 as there are 2 local matches in the word. If no sounds or only  $C_1$  and  $C_3$  are specified [+labial] the predictor is specified 0 (=unsatisfied). This is illustrated in Table 5.

**Table 5:** Example of Local OCP-[+labial] predictor specification

Template	OCP-[+labial]
/pVbVt/	1
/tVbVm/	1
/pVbVm/	2
/pVdVm/	0
/tVdVn/	0

In order to evaluate whether the increase in complexity is justified relative to the increase in data likelihood, we once again use model comparison. Recall that only differences in data likelihood between *nested* models are guaranteed to be approximately  $\chi^2$ -distributed. The feature-specific model described above is, however, not a superset model to the Local Matches model described in the previous section,

<sup>1</sup>See Gallagher (2009) for reasons why this is an appropriate way to encode laryngeal specifications in Aymara.

because the *Local Matches* predictor is *replaced* by the set of feature-specific predictors.

It is possible to indirectly compare two non-nested models by comparing each of them against their joint superset model (resulting in two nested model comparisons). First we fit a superset model containing both the new feature-based OCP predictors and the *Local Matches* predictor. To see whether the additional feature-specific predictors improve the model more than 10-12 additional predictors would be expected to improve a model under chance, we remove the feature-specific OCP predictors in bulk and take the difference in data likelihood between the remaining model (Local Matches plus controls) and the superset model. In all three languages there is considerable improvement obtained from feature specificity. The data likelihood  $\chi^2$ -test reveals that the improvements are highly significant in spite of the substantial increase in complexity (Aymara:  $\chi^2(12) = 35.98, p < .0005$ ; Dutch:  $\chi^2(11) = 128.18, p < .0001$ ; Javanese:  $\chi^2(12) = 85.35, p < .0001$ ). Furthermore, the addition of the feature-specific OCP predictors renders the *Local Matches* predictor superfluous (Aymara:  $\chi^2(1) = 0.02, p = .88$ ; Dutch:  $\chi^2(1) = 0.92, p = .34$ ; Javanese  $\chi^2(1) = 0.05, p = .83$ ). We can thus confidently conclude that the strength of OCP effects is feature-dependent. That is, across the three languages, words with consonants matching for certain features are avoided more than words with consonants matching for other features. This result supports the notion, that the grammar makes a separate OCP-constraint available for each feature.

### 5.3 Comparing feature-specific OCP effects across languages

Model comparison allows us to test whether a set of predictors contributes significantly to a model, but model comparison does not test the *direction* of the effect. Conveniently, standard logistic regression output includes tests that are based on the standard error estimate of the maximum likelihood fitted coefficients. If similarity avoidance shapes the mental lexicon, we would expect negative coefficients for significant OCP effects. That is, OCP effects should *decrease* the predicted probability of a word. This is indeed observed for all but one significant OCP effect in the feature-specific model. Figure 1 shows all significant OCP effects in the three languages, sorted by their effect sizes (Gelman 2008).

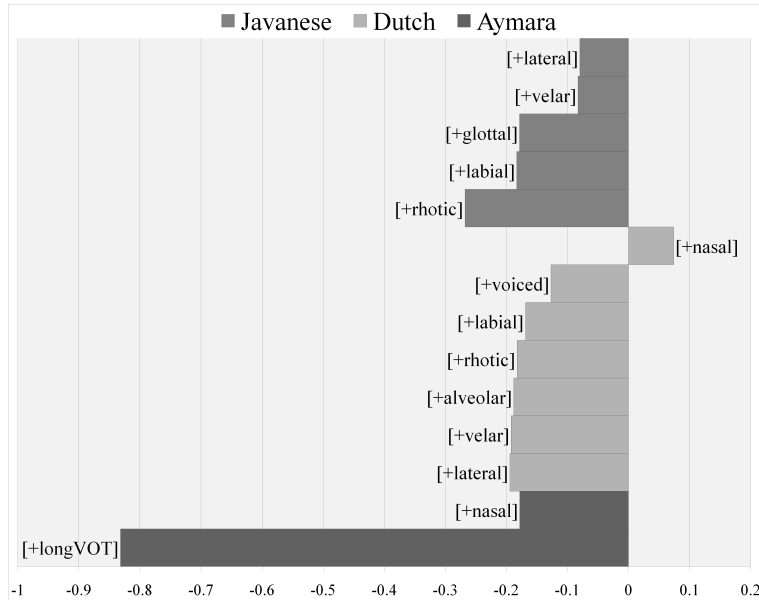
The only OCP factor fitted with a positive coefficient is OCP-[nasal] in Dutch. It should, however, be noted that this positive effect is the smallest of all significant effects. For now, we conclude that there is considerable evidence for feature specificity of OCP effects in the grammars of all three languages.

### 5.4 Locality and the OCP

The final question we intend to address in this paper is whether similarity avoidance is limited to local consonant pairs. Pierrehumbert (1993), for example, finds that similar consonants separated by another consonant are also underrepresented but to a lesser extent than similar adjacent or V-adjacent consonants in Arabic<sup>2</sup>. The question then is whether the added complexity of locality-sensitive OCP constraints

---

<sup>2</sup>See Steriade 1987, Odden 1994, Suzuki 1998 and Nevins 2004 among others for further discussion on the locality issue in assimilation and dissimilation.



**Figure 1:** Feature-specific OCP Effects ordered by effect size (coefficient of standardized predictors)

in the grammar is justified by data coverage. To test this, we add one non-local OCP predictor for each feature, tracking feature matches between the first and third consonant of the triplet.

- (6) Non-local OCP-[+feature]  
Specified 1 if  $C_1$  and  $C_3$  match for this feature.

Adding non-local feature matches again adds a substantial amount of complexity to the model. These 11-12 predictors need to substantially improve data likelihood to lower the probability of such improvement being obtained from chance.

We again fit a superset model containing both our initial feature-specific OCP predictors and the new non-local OCP predictors and then remove non-local predictors in bulk to calculate the added data likelihood. The data likelihood  $\chi^2$ -square test shows that the improvement in all three languages is highly significant. In Aymara ( $\chi^2(12) = 30.99, p < .002$ ), Dutch ( $\chi^2(11) = 48.46, p < .0001$ ) and Javanese ( $\chi^2(12) = 102.64, p < .0001$ ) non-local predictors significantly improve data coverage, thus replicating the findings of Pierrehumbert (1993) with respect to locality of similarity effects in Arabic. The structural descriptions in the final set of OCP constraints for Aymara, Dutch and Javanese are given below.

- (7) Local OCP-[+feature]  
\* $[+F]V_0[+F]$
- (8) Non-local OCP-[+feature]  
\* $[+F]V_0CV_0[+F]$

In our current implementation, local and non-local predictors for each feature are allowed to vary independently. In future work, we plan to investigate how this

particular formulation of locality-sensitive OCP constraints compares to other conceivable ones, such as formulations that encode local feature matches separately from the sum of local and non-local feature matches. Furthermore, we plan to investigate whether the complexity of the model can be reduced by tying local and non-local constraint violations for a given feature more closely together.

In sum, we have arrived at a model of similarity that captures feature-specific differences in OCP strength and differences between local and non-local OCP effects in three genetically unrelated languages. This model is relatively complex (i.e., it contains a relatively large number of parameters that are fitted to the data). While the model comparisons reported above suggest that this complexity is warranted by the data coverage of the model, it is a good idea to compare our model against alternative models. Next, we compare the predictions of the current state-of-the-art model of phonological similarity avoidance (Frisch *et al.* 2004) against our feature-specific locality-sensitive model.

## 6 Comparison with Frisch *et al.* 2004

Frisch *et al.* (2004) have shown that the observed-over-expected ratios (O/E) of consonant pairs co-occurring within Arabic roots are strongly correlated with their *similarity metric of natural classes*. Their formula for phonological similarity is given below.

$$(9) \quad \textit{similarity} = \frac{\textit{shared natural classes}}{\textit{shared+unshared natural classes}}$$

The Frisch *et al.* (2004) model is far less complex than our feature-specific locality-sensitive model in terms of the number of free parameters. Yet it provides a good fit against the Arabic data. Hence, it is worth testing to what extent the data from Aymara, Dutch, and Javanese can be accounted for by the Frisch *et al.* (2004) model.

Unfortunately, it is unclear how to extend Frisch *et al.* (2004) to other languages. Frisch *et al.* (2004) use a three-way feature system, where sounds can be specified “+”, “-” or remain unspecified for given a feature. Only unspecified features are assumed *not* to define natural classes. For their Arabic data, Frisch and colleagues make several informed, but nevertheless not entirely principled, decisions as to which sounds are specified as “-” vs. which ones remain unspecified.

The basis of the Frisch *et al.* (2004) similarity metric are the natural classes defined by a feature system through *Structured Specification* (Broe 1993). This means that feature specifications only matter as far as they pick out natural classes. The specification of the feature [acute] in Arabic, for example, only matters for coronals, as it only subdivides the class of segments specified as [+coronal]. In their particular feature system for Arabic, however, the feature [-nasal] is specified for labial, coronal and dorsal, non-continuants even though there are no dorsal nasals in Arabic. Since not all dorsals are specified [-nasal], this specification does matter, as it renders dorsal oral stops more similar than their specification for [-continuant] alone would. However, there is no phonological evidence for a featural specification of segments for orality (Steriade 1995) and Frisch *et al.* (1997) do in fact not specify dorsal oral stops for non-nasality.

Additionally, Frisch *et al.* (2004) set all similarity values for pairs of sounds that do not have a place feature in common to zero, even though they may share natural classes (e.g. /b/ and /d/ have similarity 0 even though they are both [+voice]). It is not clear how or whether this decision should be implemented in languages like Aymara that have laryngeal co-occurrence restrictions.

In short, there is more than one way to extend the assumptions made in Frisch *et al.* (2004) to the languages in our data set. One way to compare our feature-specific locality-sensitive model against the Frisch *et al.* (2004) would be to explore a variety of models that base similarity on feature systems compatible with the general assumptions spelled out in Frisch *et al.* (2004). In ongoing work, we are pursuing this approach. Here, however, we take a simpler approach as a first step. We instead test the pivotal prediction made by Frisch *et al.* (2004): the size of the class picked out by a particular feature should be inversely correlated with the size of the OCP effect. Frisch *et al.* (2004) state this with reference to the organization of the Arabic sound system.

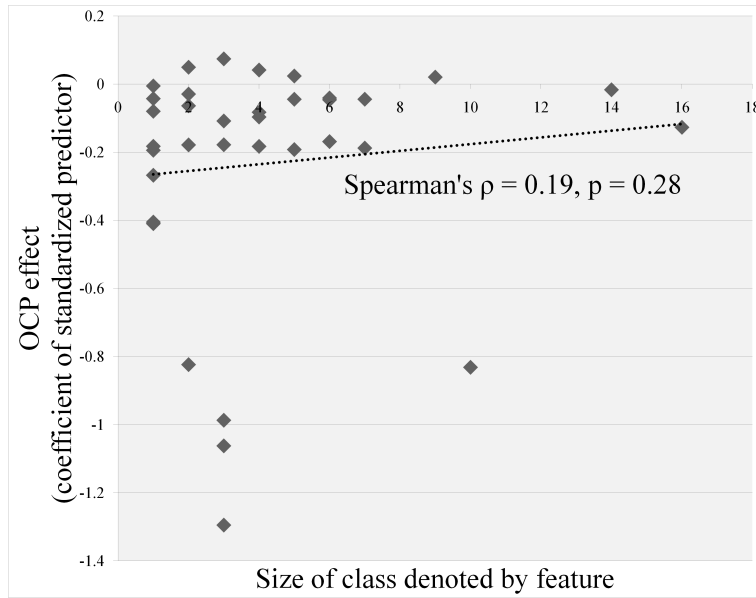
We claim that similarity for coronal pairs like /s, n/ is less than the similarity of /f, m/ due to the larger space of contrasts in the coronals. As a result, the co-occurrence restriction for /f, m/ is stronger than the co-occurrence restriction for /s, n/...The larger number of natural classes among the coronals, and to some extent the dorsals and pharyngeals, decreases the similarity between consonant pairs in those classes that share few features. In a small class, like the Arabic labials, these non-contrastive features do not contribute to dissimilarity.

Assuming *Structured Specification* (Broe 1993) the *similarity metric of natural classes* achieves that, everything else being equal, matching for place features that pick out smaller sets of sounds renders two sounds more similar than matching for place features that pick out larger sets. This is supported by data from Arabic, where, e.g., coronals (11 consonants) co-occur more readily within the same root than labials (3 consonants)<sup>3</sup>.

This prediction can be tested using a simple correlation test. We computed the correlation between the size of a class denoted by a given feature and the strength of the OCP effect as expressed by our effect size measure. As apparent in Figure 2, we found no evidence for the correlation predicted by Frisch *et al.* (2004). The Spearman rank correlation reveals that the prediction of Frisch *et al.* (2004) is not supported by any of the three languages included in our study (Aymara:  $S = 162.6, p = 0.16, \rho = 0.43$ ; Dutch:  $S = 210.8, p = 0.9, \rho = 0.04$ ; Javanese:  $S = 192.3, p = 0.3, \rho = 0.33$ ). The correlation is also not significant if we only consider place classes and in fact goes in the opposite direction ( $S = 1535.4, p = 0.52, \rho = -0.15$ ). A similar lack of this correlation has previously been identified for Muna and Rotuman (Coetzee & Pater 2008).

---

<sup>3</sup>As Adam Albright points out to us, it is theoretically possible to conceive feature systems that increase the number of natural classes at less populated places of articulation while decreasing the number of natural classes at more populated ones. In such systems it is possible for this prediction not to hold. Such feature systems, however, seem to be highly unlikely systems for natural languages (Clements 2003).



**Figure 2:** Correlation between OCP effect of class and size of class denoted by a given feature (all languages)

If confirmed by a more rigorous test of the Frisch *et al.* (2004) model, the lack of the predicted correlation raises questions about the results Frisch and colleagues obtained for Arabic. Since Frisch *et al.* (2004) did not compare their model against a model with feature-specific OCP effects, it is possible that the correlation between OCP effect and class size observed in Arabic is an artifact of the properties of the features that denote small classes in Arabic (e.g. their psychoacoustic properties). For example, the feature [+labial], which shows strong OCP effects in Arabic, also triggers strong OCP effects in Dutch and Javanese, even though the relative size of the class of labials in all three languages differs. Javanese, for example, has an equal number of labials and velars (4 each). Nonetheless OCP effects are twice as strong for labials (-.18 vs. -.08; coefficients of standardized predictors).

## 7 Conclusion

We have presented evidence in favor of a single formulation of the OCP constraints, which manages to best account for consonant triplet attestation in three genetically unrelated languages. The most successful model of the OCP, in terms of data likelihood, is feature-specific and locality-sensitive. The complexity of this model is justified by the data coverage it achieves, as evidenced by model comparison. The observed feature specificity of OCP effects raises an intriguing possibility. It is possible that psychoacoustic or articulatory properties of phonological features, rather than their language specific contribution to the contrast system of a given language, determine the strength of associated OCP effects. In future work, we plan to address this question experimentally. Finally, we have *not* found evidence that OCP effects are inversely correlated with natural class size—contrary to a prediction made by

Frisch *et al.* (2004).

We close by relating our findings to work on probabilistic grammars (e.g. Boersma & Hayes 2001, Hayes & Wilson 2008; also see Pierrehumbert 2001 and references therein). Although the data presented in this paper were analyzed using logistic regression, our findings are generally compatible with an interpretation in terms of probabilistic grammar frameworks (for more on the relation between logistic regression and such frameworks, see Manning 2003). Among other things, our findings can be taken as support for the hypothesis that grammar needs to include OCP constraints that are sensitive to specific features (e.g., Coon & Gallagher 2007) and to the distance between similar consonants (e.g., Pierrehumbert 1993).

## References

- Adisasmito-Smith, Niken, 2004. *Phonetic and Phonological Influences of Javanese on Indonesian*. Ithaca, New York: Cornell University dissertation.
- Agresti, Alan. 2002. *Categorical Data Analysis (2nd ed.)*. New York, NY: John Wiley & Sons.
- Baayen, R. Harald, Richard Piepenbrock, & Hedderik van Rijn. 1993. The CELEX lexical database (CDROM). Philadelphia: Linguistic Data Consortium, University of Pennsylvania.
- Boersma, Paul, & Bruce Hayes. 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32.45–86.
- Boll-Avetisyan, Natalie, & René Kager. 2008. Identity avoidance between non-adjacent consonants in artificial language segmentation. In *Laboratory Phonology 11 - Book of Abstracts*, ed. by Paul Warren, 15–16.
- Booij, Geert. 1995. *The Phonology of Dutch*. Oxford: Clarendon Press.
- Broe, Michael, 1993. *Specification Theory: the Treatment of Redundancy in Generative Phonology*. Edinburgh, UK: University of Edinburgh dissertation.
- Clements, George. 2003. Feature economy in sound systems. *Phonology* 20.287–333.
- Coetzee, Andries, & Joe Pater. 2008. Weighted constraints and gradient restrictions on place co-occurrence in Muna and Arabic. *Natural Language and Linguistic Theory* 26.289–337.
- Coon, Jessica, & Gillian Gallagher. 2007. Similarity and correspondence in Chol roots. In *North East Linguistics Society* 38, 167–180.
- de Lucca, Manuel. 1987. *Diccionario practico Aymara-Castellano Castellano-Aymara*. Cochabamba: Los Amigos del Libro.
- Fox, John. 1991. *Regression Diagnostics: Quantitative Applications in the Social Sciences*. Thousand Oaks, CA: Sage.
- Frisch, Stefan, Michael Broe, & Janet Pierrehumbert. 1997. Similarity and phonotactics in Arabic. Ms. (Available as ROA-223 on <http://roa.rutgers.edu>).
- , Janet Pierrehumbert, & Michael Broe. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22.179–228.
- Gallagher, Gillian. 2009. Perceptual distinctness and laryngeal (dis)harmony. Submitted.
- Gelman, Andrew. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine* 27.2865–2873.

- Hayes, Bruce, & Colin Wilson. 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry* 39.379–440.
- Hepworth, Graham, Ian Gordon, & Michael McCullough. 2007. Accounting for dependence in similarity data from DNA fingerprinting. *Statistical Applications in Genetics and Molecular Biology* 6.1–13.
- Leben, William, 1973. *Suprasegmental Phonology*. Cambridge, MA: MIT dissertation.
- MacEachern, Margaret. 1999. *Laryngeal Co-occurrence Restrictions*. New York: Garland.
- Manning, Christopher. 2003. Probabilistic syntax. In *Probabilistic Linguistics*, ed. by Rens Bod, Jennifer Hay, & Stefanie Jannedy, 289–341.
- McCarthy, John. 1986. OCP effects: Gemination and anti-gemination. *Linguistic Inquiry* 17.207–263.
- Mester, R. Armin, 1986. *Studies in Tier Structure*. University of Massachusetts, Amherst dissertation.
- Nevins, Andrew, 2004. *Conditions on (Dis)Harmony*. Cambridge, MA: MIT dissertation.
- Odden, David. 1994. Adjacency parameters in phonology. *Language* 70.289–330.
- Pierrehumbert, Janet. 1993. Dissimilarity in the Arabic verbal roots. In *North East Linguistics Society* 23, 367–381.
- . 2001. Stochastic phonology. *Glott International* 5.1–13.
- Prince, Alan, & Paul Smolensky. 2004. *Optimality Theory: Constraint Interaction in Generative Grammar*. Malden, MA, and Oxford, UK: Blackwell. Published version of Prince and Smolensky (1993), ROA-537.
- Rose, Sharon, & Rachel Walker. 2004. A typology of consonant agreement as correspondence. *Language* 80.475–531.
- Steriade, Donca. 1987. Locality conditions and feature geometry. In *North East Linguistics Society* 17, 595–617.
- . 1995. Underspecification and markedness. In *The Handbook of Phonological Theory*, ed. by John Goldsmith, 114–175. Blackwell.
- Suzuki, Keiichiro, 1998. *A Typological Investigation of Dissimilation*. Tucson, AZ: University of Arizona dissertation.
- Uhlenbeck, Eugenius. 1978. *Studies in Javanese Morphology*. The Hague: Martinus Nijhoff.
- Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19.395–439.

## 8 Appendix

Tables 6 to 8 summarize the feature systems assumed for Aymara (based on MacEachern 1999), Dutch (based on Booij 1995), and Javanese (based on Adisasmito-Smith 2004), respectively. Shaded features were omitted from the feature-specific model to control for collinearity (see section 5.2).

**Table 6: Aymara Feature system**

	p'	tʃ'	t'	k'	q'	p <sup>h</sup>	tʃ <sup>h</sup>	t <sup>h</sup>	k <sup>h</sup>	q <sup>h</sup>	p	tʃ	t
labial	+	-	-	-	-	+	-	-	-	-	+	-	-
alveolar	-	-	+	-	-	-	-	+	-	-	-	-	+
post-alveolar	-	+	-	-	-	-	+	-	-	-	-	+	-
palatal	-	-	-	-	-	-	-	-	-	-	-	-	-
velar	-	-	-	+	-	-	-	-	+	-	-	-	-
uvular	-	-	-	-	+	-	-	-	-	+	-	-	-
glottal	-	-	-	-	-	-	-	-	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	-	-	-
continuant	-	-	-	-	-	-	-	-	-	-	-	-	-
sonorant	-	-	-	-	-	-	-	-	-	-	-	-	-
approximant	-	-	-	-	-	-	-	-	-	-	-	-	-
rhotic	-	-	-	-	-	-	-	-	-	-	-	-	-
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-
aspirated	-	-	-	-	-	+	+	+	+	+	-	-	-
ejective	+	+	+	+	+	-	-	-	-	-	-	-	-
strident	-	+	-	-	-	-	+	-	-	-	-	+	-
longVOT	+	+	+	+	+	+	+	+	+	+	-	-	-
	k	q	s	m	n	ɲ	l	ʎ	r	j	w	x	h
labial	-	-	-	+	-	-	-	-	-	-	+	-	-
alveolar	-	-	-	-	+	-	+	-	+	-	-	-	-
post-alveolar	-	-	-	-	-	-	-	-	-	-	-	-	-
palatal	-	-	-	-	-	+	-	+	-	+	-	-	-
velar	+	-	-	-	-	-	-	-	-	-	+	+	-
uvular	-	+	-	-	-	-	-	-	-	-	-	-	-
glottal	-	-	-	-	-	-	-	-	-	-	-	-	+
nasal	-	-	-	+	+	+	+	+	+	+	+	+	+
continuant	-	-	+	-	-	+	+	+	+	+	+	+	+
sonorant	-	-	-	+	+	+	+	+	+	+	+	+	-
approximant	-	-	-	-	-	-	+	+	+	+	+	-	-
rhotic	-	-	-	-	-	-	-	-	+	-	-	-	-
lateral	-	-	-	-	-	-	+	+	-	-	-	-	-
ejective	-	-	-	-	-	-	-	-	-	-	-	-	-
strident	-	-	+	-	-	-	-	-	-	-	-	-	-
longVOT	-	-	-	-	-	-	-	-	-	-	-	-	-

**Table 7: Dutch Feature system**

	p	b	t	d	k	g	ŋ	m	n	l	r	f	v	s	z	ʃ	ʒ	j	x	ɣ	ɦ	w	dʒ
labial	+	+	-	-	-	-	-	+	-	-	-	+	+	-	-	-	-	-	-	-	-	+	-
alveolar	-	-	+	+	-	-	-	-	+	+	+	-	-	+	+	-	-	-	-	-	-	-	-
post-alveolar	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	+
palatal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
velar	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-
glottal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-
nasal	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
continuant	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-
sonorant	-	-	-	-	-	+	+	+	+	+	-	-	-	-	-	-	-	+	-	-	-	+	-
approximant	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	+	-
rhotic	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
lateral	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-
voiced	-	+	-	+	-	+	+	+	+	+	-	+	-	+	-	+	+	+	+	+	+	+	+
strident	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	+

**Table 8: Javanese Feature system**

	b <sup>h</sup>	d <sup>h</sup>	dʒ <sup>h</sup>	q <sup>h</sup>	g	p	t	tʃ	ʈ	k	m	n	ŋ	ŋ	ʔ	h	r	l	s	w	j	
labial	+	-	-	-	-	+	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-	-
alveolar	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	+	-	+	-	-
retroflex	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-
post-alveolar	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-
palatal	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	+
velar	-	-	-	-	+	-	-	-	+	-	-	-	-	+	-	-	-	-	-	-	+	-
glottal	-	-	-	-	-	-	-	-	-	-	-	+	+	+	+	+	+	-	-	-	-	-
nasal	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	-	-	-	-	-	-
continuant	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	+	+	+	+
sonorant	-	-	-	-	-	-	-	-	-	-	+	+	+	+	-	-	+	+	-	-	+	+
approximant	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	+	+
rhotic	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
breathy	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
strident	-	-	+	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	+	-	-