THE ROLE OF SIMILARITY IN PHONOLOGY: EXPLAINING OCP-PLACE

Stefan Frisch, Michael Broe, and Janet Pierrehumbert
Northwestern University, Evanston, IL, USA

ABSTRACT

This paper introduces an improved similarity model to account for cooccurrence restrictions in the verbal roots of Arabic, extending the results of [1]. A new quantitative measure of OCP effects and a new similarity metric are presented based on information theory. Similarity is computed by applying an entropic formulation of the cognitive "basic level" to a hierarchical representation of natural classes as in [2].

INTRODUCTION

The canonical Arabic verb form is a sequence of three consonants. Vowels provided by other morphemes are interleaved with the consonants to produce surface forms. There are strong, gradient cooccurrence restrictions among the consonants in the root. In particular, roots containing more than one consonant from one of the following classes are highly underrepresented [3].

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

There are also gradient effects within the major classes. For example, /l/ and /r/ have a stronger restriction than /l/ and /n/.

The traditional account of the OCP effects relies on categorical cooccurrence rules. The traditional model encounters a number of problems. It cannot properly account for the gradience of OCP phenomena, both among adjacent consonants and over distance [1]. In addition, it is a negative constraint, and thus does not account for the patterns of overrepresentation presented below.

THE SIMILARITY ACCOUNT

According to [1], the degree to which the OCP is violated by two homorganic consonants in a root is a function of the perceived similarity of those two consonants. In addition, intervening consonants are interference and thus reduce the perceived similarity of more distant consonants. In this way, the gradient nature of the OCP is captured. The OCP is strongest in the case of adjacent identical consonants, which have a high degree of perceived similarity. It is weaker for identical consonants at a distance and for non-identical consonants. It is weakest for non-identical consonants which are non-adjacent. Given that the similarity account can capture the gradience of the OCP effects, it is a more empirically adequate account.

The challenge for the similarity account is to determine a function that supplies the best fit between the similarity gradient over the consonant inventory and the observed cooccurrence restrictions. In [1], similarity was computed for each consonant pair by the ratio of shared to shared plus nonshared features. Contrastive underspecification was used to capture gradient effects across classes. For small classes, like the labials and coronal sonorants, very few features are needed to differentiate sounds, which increases the value of the similarity function. For large classes like the coronals, there are many features needed to differentiate them, which reduces the value of the similarity function between members within the class.

However, the original similarity account did not capture all of the OCP effects [1]. The model failed to capture the strength of the restriction between /χ/ and /ʁ/ and the other dorsals. Also, the division within the coronal sonorants, where /l/ and /r/ form a subclass in contrast to /n/ was not captured. In addition, the use of contrastive underspecification is undesirable. It is responsible for the failure to differentiate the subclasses within the coronal sonorants. Recent work shows that the phenomena which were originally taken to support contrastive underspecification can be given a more satisfactory reanalysis in terms of privative features and licensing [5]. Finally, it is undesirable on formal grounds: contrastive underspecification is inherently derivational and logically intractable [2].

In the remainder of the paper, we first present additional evidence that the cooccurrence restrictions in Arabic are based on similarity, and not on categorical rules. We then present a new approach to the similarity function which more adequately models the data.

COMPUTING OCP EFFECTS

We have studied the Arabic root cooccurrence constraints using notions from information theory. Any consonant can be characterized by examining the quantitative extent of its cooccurrence with each of the other consonants in the system. This set of values can be represented as a cooccurrence vector, the elements of which are normalized to indicate overrepresentation or underrepresentation. The vectors of two consonants can then be compared, revealing the degree to which their cooccurrence profiles match across the entire consonant inventory. The match is quantified using information theoretic interdependence. Two consonants will have a high degree of interdependence if the cooccurrence restrictions of one consonant are predictable from the cooccurrence restrictions of the other. This can occur in two distinct ways: either the two consonants can have identical restrictions, or they can have complementary restrictions. If the restrictions are identical, the consonants pattern the same way with respect to OCP effects. If the restrictions are complementary, the consonants have opposite patterns with respect to OCP effects. When the interdependence of two consonants is low, there is no relation between the distributions of the consonants.

The computation of interdependence is based on the entropy of the pattern of cooccurrence restrictions in the system. Entropy is a measure of uncertainty of the outcome of an event. Entropy is

$$H(x) = H(p_1, \ldots, p_n) = -\sum_i p_i \log_2(p_i)$$

where p_i is the probability of x having outcome i. If all outcomes are equiprobable, then there is high uncertainty and the entropy is large. Less equiprobable outcomes result in lower entropy, as the outcome is relatively more predictable.

For a single consonant, we are interested in the uncertainty between two possible outcomes: overrepresentation or underrepresentation with respect to other consonants. For a pair of consonants there are four possible outcomes. They may be: both underrepresented, one underrepresented and the other not (for each), or they may both be overrepresented with respect to other consonants. As the correlation between cooccurrence vectors increases, the uncertainty of the joint outcome goes down. So interdependence can be expressed as:

$$J(x,y) = H(x) + H(y) - H(x,y)$$

Interdependence quantifies the degree to which entropy is shared by both consonants, an entropic measure of correlation of information. Table 1 is a sample calculation on a simplified data set, employing discrete over and underrepresentation.

Table 1: Computing interdependence of t and s. All outcomes equiprobable.

	b	d	z	ʃ
t	Over	Under	Under	Over
s	Over	Under	Under	Under
H("t") = H(0.5, 0.5) = 1				
H("s") = H(0.25, 0.75) = 0.81				
H("t,s") = H(0.25, 0.5, 0.25) = 1.5				
J("t,s") = 1 + 0.81 - 1.5 = 0.31				

Table 2 shows the interdependence computed over the entire Arabic system, based on the degree of over and underrepresentation between consonant pairs. Interdependence is normally unsigned, but a sign has been added for clarity. Positive values indicate shared cooccurrence restrictions, negative values indicate complementary restrictions.

Gray shading in table 2 indicates interdependence of at least 0.03. All of the major classes with cooccurrence restrictions have interdependence at or above this level. In addition, a cooccurrence restriction between /w/ and the labials is revealed.

The interdependence measure also reveals a pattern of overrepresentation, indicated by boxes in table 2. Labials are consistently overrepresented with the coronal obstruents; the glides /w/ and /y/ are overrepresented with the coronal obstruents; and the coronal sonorants are overrepresented with the dorsals and gutturals.

We claim these patterns of overrepresentation are another reflex of similarity. The coronal sonorants share place of ar-

Table 2: Interdependence of cooccurrence restrictions among the Arabic consonants.

	bfm	td	TD	θð	sz	SZ	ʃ	kg	q	χʁ	ħʕ	hʔ	lr	n	wy
bfm	0.60	-0.09	-0.10	-0.03	-0.11	-0.05	-0.04	-0.01	-0.02	-0.01	0.00	-0.03	-0.04	0.00	0.13
td	-0.09	0.14	0.12	0.07	0.15	0.08	0.04	-0.02	0.00	-0.02	-0.03	0.00	-0.02	-0.01	-0.06
TD	-0.10	0.12	0.31	0.08	0.14	0.15	0.06	0.01	-0.02	0.00	-0.03	0.01	-0.04	-0.03	-0.08
θð	-0.03	0.07	0.08	0.09	0.12	0.05	0.03	0.00	0.00	-0.01	-0.03	0.00	-0.01	0.01	-0.05
sz	-0.11	0.15	0.14	0.12	0.37	0.09	0.09	-0.01	0.01	-0.04	-0.07	0.00	0.00	0.00	-0.09
SZ	-0.05	0.08	0.15	0.05	0.09	0.15	0.05	0.00	-0.01	-0.01	-0.01	0.01	-0.02	-0.01	-0.05
ʃ	-0.04	0.04	0.06	0.03	0.09	0.05	0.06	0.00	0.00	-0.02	-0.02	0.01	0.00	0.00	-0.04
kg	-0.01	-0.02	0.01	0.00	-0.01	0.00	0.00	0.48	0.08	0.05	-0.01	-0.08	-0.26	-0.17	-0.01
q	-0.02	0.00	-0.02	0.00	0.01	-0.01	0.00	0.08	0.16	0.12	-0.07	0.03	-0.13	-0.09	-0.02
χʁ	-0.01	-0.02	0.00	-0.01	-0.04	-0.01	-0.02	0.05	0.12	0.37	0.20	0.13	-0.10	-0.12	-0.01
ħʕ	0.00	-0.03	-0.03	-0.03	-0.07	-0.01	-0.02	-0.01	-0.07	0.20	0.38	0.16	-0.07	-0.04	0.00
hʔ	-0.03	0.00	0.01	0.00	0.00	0.01	0.01	-0.08	0.03	0.13	0.16	0.25	-0.06	-0.05	-0.03
lr	-0.04	-0.02	-0.04	-0.01	0.00	-0.02	0.00	-0.26	-0.13	-0.10	-0.07	-0.06	1.00	0.36	0.01
n	0.00	-0.01	-0.03	0.01	0.00	-0.01	0.00	-0.17	-0.09	-0.12	-0.04	-0.05	0.36	0.30	-0.01
wy	0.13	-0.06	-0.08	-0.05	-0.09	-0.05	-0.04	-0.01	-0.02	-0.01	0.00	-0.03	0.01	-0.01	0.18

ticulation with the coronal obstruents, and thus are more similar to them than to the dorsals and gutturals. The glides /w/ and /y/ involve a dorsal articulation, and thus are more similar to the dorsals and gutturals than the coronals. Finally, the glide /w/ has a labial articulation, which gives some OCP effect between the glides and the labials.

The patterns of overrepresentation are especially significant, showing that the cooccurrence data cannot be accounted for solely in terms of a negative constraint. If the OCP effects were only based on a cooccurrence restriction, the unrestricted cases should be uniformly overrepresented. Instead, pairs of similar consonants are restricted, and highly dissimilar ones are more likely to cooccur. Thus, the patterns of overrepresentation provide additional evidence for the similarity account, extending its application from underrepresentation to overrepresentation. We now turn to the new similarity model, which provides an improved fit for the data.

A NEW APPROACH

The problems with the previous similarity account [1] can be remedied by adopting a heirarchical approach to natural classes and feature specification [2]. In the approach developed there, the system of contrasts in a language is a structural relation among the natural classes of the language. The natural classes, which form a partially ordered set, can be represented as a "tangled" hierarchy, or lattice.

In this model, the classification of the phonemes of a language into natural classes is very similar to other cognitive classification systems. The natural kinds of mammal, dog, and German Shepherd are related to one another in the same way that the classes coronal, coronal stop, and /t/ are. Research in cognitive science has shown that within such hierarchies, mid-level categories, like dog, are privileged with respect to superordinate or subordinate ones [6]. For example, these so called "basic level" categories are the first to be acquired by children and are more readily accessible (reflected in faster reaction times).

We claim that there is also a cognitive basic level in phonological systems, and this basic level is the most important one in determining OCP effects. We propose that the correct feature specification of the Arabic consonant system has the categories in (1) as basic level categories.

COMPUTING SIMILARITY

The function we use to compute similarity differs in two ways from [1]. First, rather than computing similarity based on individual features, we propose to compute similarity based on natural classes. Second, we use a weighting scheme to capture the primacy of basic level categories in perceived similarity. Natural classes at the basic level are weighted

Table 3: Computation of similarity of the Arabic consonants.

	bfm	td	TD	θð	sz	SZ	ʃ	kg	q	χʁ	ħʕ	hʔ	lr	n	wy
bfm	0.49	0.04	0.04	0.03	0.03	0.04	0.04	0.04	0.03	0.05	0.05	0.05	0.07	0.10	[illegible]
td	0.04	0.68	0.33	0.28	0.27	0.13	0.28	0.34	0.35	0.06	0.01	0.01	0.05	0.05	0.03
TD	0.04	0.33	0.69	0.13	0.12	0.30	0.12	0.17	0.17	0.01	0.08	0.08	0.07	0.07	0.04
θð	0.03	0.28	0.13	0.68	0.62	0.25	0.59	0.14	0.14	0.17	0.05	0.05	0.05	0.05	0.03
sz	0.03	0.27	0.12	0.62	0.69	0.24	0.63	0.13	0.14	0.16	0.05	0.05	0.05	0.05	0.03
SZ	0.04	0.13	0.30	0.25	0.24	0.68	0.24	0.05	0.06	0.06	0.22	0.22	0.07	0.07	0.04
ʃ	0.04	0.28	0.12	0.59	0.63	0.24	1.00	0.14	0.22	0.17	0.05	0.05	0.02	0.02	0.01
kg	0.04	0.34	0.17	0.14	0.13	0.05	0.14	0.68	0.66	0.18	0.06	0.06	0.01	0.01	0.07
q	0.03	0.35	0.17	0.14	0.14	0.06	0.22	0.66	1.00	0.19	0.06	0.06	0.00	0.00	0.03
χʁ	0.05	0.06	0.01	0.17	0.16	0.06	0.17	0.18	0.19	0.67	0.23	0.23	0.01	0.01	0.09
ħʕ	0.05	0.01	0.08	0.05	0.05	0.22	0.05	0.06	0.06	0.23	0.68	0.62	0.01	0.01	0.09
hʔ	0.05	0.01	0.08	0.05	0.05	0.22	0.05	0.06	0.06	0.23	0.62	0.68	0.01	0.01	0.09
lr	0.07	0.05	0.07	0.05	0.05	0.07	0.02	0.01	0.00	0.01	0.01	0.01	1.00	0.76	0.34
n	0.10	0.05	0.07	0.05	0.05	0.07	0.02	0.01	0.00	0.01	0.01	0.01	0.76	1.00	0.35
wy	0.19	0.03	0.04	0.03	0.03	0.04	0.01	0.07	0.03	0.09	0.09	0.09	0.34	0.35	0.70

higher than those which are above or below it. The weighting function is also entropic and based on the optimization of information balance inside and outside of the category [7].

Table 3 shows the results of one similarity computation. Shading indicates homorganic consonant pairs with similarity greater than 0.1. All of the major classes are modeled at this level of similarity. The correct patterning of /χ/ and /ʁ/ with both the gutturals and the velars is captured, and the subclassification of the coronal sonorants is also obtained. In addition, the significant overrepresentation shown in table 2 is accounted for by pairs with very low similarity. The boxed regions in table 3 show similarity below 0.05.

Weighting replaces the use of contrastive underspecification; the higher weighting of basic level categories compensates for the additional noncontrastive features that might increase the perceived differences within categories. We are still exploring the proper combination of feature assignments and weighting functions in order to find the best fit to the data. The empirical advantage of this approach obtains regardless of the particular function used.

CONCLUSION

OCP-Place is a phonological reflex of a cognitive universal: similarity. The pattern of cooccurrence restrictions across the lexicon of Arabic reflects both cooccurrence restrictions between similar consonants and an overrepresentation of highly dissimilar consonants. Perceived similarity between two consonants is a function of the natural classes in which those consonants are found which are weighted based on their proximity to the basic level.

ACKNOWLEDGMENT

This work was supported in part by NSF Grant No. BNS-9022484.

REFERENCES

[1] Pierrehumbert, J. (1992), "Dissimilarity in the Arabic verbal roots", *Proceedings of NELS 23*, Amherst: GSLA.

[2] Broe, M. (1993), *Specification theory: the treatment of redundancy in generative phonology*, Ph.D. dissertation, Edinburgh.

[3] McCarthy, J.J. (1994), "Guttural phonology", in *Papers in laboratory phonology III*. Cambridge: Cambridge University Press.

[5] Steriade, Donca. (1994), "Underspecification and markedness", in *Handbook of Phonology*, J. Goldsmith, ed. Oxford: Basil Blackwell.

[6] Rosch, E. et al. (1976), "Basic objects in natural categories", *Cognitive Psychology* 8: 382-439.

[7] Broe, M. (1995), "An entropic measure of category utility", Ms., Northwestern University.