ABSTRACT PHONETIC CATEGORIES ARE PREDICTABLE FROM QUANTITATIVE PHONOTACTICS

Eleonora Cavalcante Albano

LAFAPE-IEL, State University of Campinas (UNICAMP), Brazil
albano@unicamp.br

ABSTRACT

This paper presents a new and intriguing finding: phonetic categorization is largely predictable from quantitative phonotactics in 3 Brazilian Portuguese word databases. The predictors are log frequencies of 'VC, 'CV and V'CV phone sequences in word types. Vowel categories emerge from discriminant analysis of '_C and 'C_data. Consonant categories emerge from discriminant analysis of V'_V data. Results suggest that lexical phonotactic biases can encode abstract phonetic categories.

Keywords: categorization, lexicon, phonotactics, phonology.

1.INTRODUCTION

Phonotactics has long been known to convey general phonetic information, such as syllable structure and sonority [1]. Yet its role in signaling specific phonetic categories has not been much explored to date. The suggestion that phonotactic probabilities may be phonetically informative is present in the literature on psycholinguistics [2], among other fields. Maddieson [3] has, in addition, shown that at least some phonotactic biases reflect linguistic phonetic trends.

This paper takes this lead to derive phonetic categorization from the frequencies of 'VC, 'CV and V'CV phone sequences in 3 phonetically transcribed databases of Brazilian Portuguese (henceforth BP). Co-occurrence biases with the preceding or following consonant yield correct categorization of stressed vowels. Co-occurrence biases with adjacent vowels, both stressed and pre-stressed, yield correct categorization of onset consonants.

Results throw light on the relationship between phonotactics and phonetics.

2.METHODOLOGY

Since knowledge of BP phonotactics is scant, more than one corpus was investigated.

Issues about corpus size and representativeness (e. g., written vs. oral language) were thus settled empirically.

2.1.Materials and Treatment

2.1.1.Database makeup

The materials consist of three word databases. One, called *Mini-Aurélio,* was directly drawn from the 27,074 entries of an abridged dictionary [4].

The other two were compiled from running text. The first, called *CETEN*, has 223,193 words, derives from a South East Brazil newspaper, and is available from [5]. The second, called *NURC-SE*, has 45,579 words and derives from orthographic transcriptions of a set of lectures, dialogues and interviews recorded in the same region [6].

2.1.2.Orthography to phone conversion

Orthography to phone conversion was performed with the software described in [7]. The resulting broad transcription follows the allophonic rules of Southeastern BP. IPA script is used.

2.1.3.Frequency counts

The analysis units revolve around stressed syllables in word medial position – where the maximal vowel and consonant inventories occur. There are seven stressed vowels: /i, e, ɛ, a, ɔ, o, u/; and nineteen medial consonants: / p, b, f, v, m, t, d, s, z, n, l, r, r, ʃ, ʒ, ʎ, ɲ, k, g/. The 'VC and 'CV combinations yield, each, 133 data points. These are quite sufficient for vowel discrimination.

Consonant discrimination, however, requires more information. Accordingly, a V'CV unit was formed by resorting to the five pre-stressed vowels: /i, e, a, o, u/. This yields a total of 665 data points.

Type and token frequencies were computed from the text databases, i. e., *CETEN* and *NURC-SE*. All raw values were converted to logarithms.

2.1.4.Linguistic and statistical representativeness

In spite of the obvious desirability of an oral corpus, *CETEN* was the only sample to meet homoscedasticity and multivariate normality, which are assumed in discriminant analysis. In addition, Table 1 shows how its mean log frequencies correlate to those of the other corpora.

Table 1: Pearson's correlation coefficients for mean log frequencies in the 3 corpora.

R values for Mean Log Frequencies (p<0.5)		*Mini-Aurélio*	*NURC-SE*
V_V in V'CV	*CETEN*	0.96	0.98
C in V'CV	*CETEN*	0.96	0.99

These high correlations justify considering *CETEN* representative enough to serve as the base for deriving the classification functions to be used in further research.

2.2.Statistical Analysis

Best subset discriminant analysis (henceforth BSDA) was performed on type frequencies. Token frequencies were excluded because they violated most of the assumptions of the model. In addition, cluster analysis was performed to help interpret the findings. Both were run with Statistica 6.0.

3.RESULTS

Analysis based on the appropriate phonotactic contexts led to 100% correct discrimination within all phonetic categories. In addition, all pairwise comparisons, which are not reported here for lack of space, are significant at the 0.05 level.

3.1.Vowel Classes

'C_ analysis yielded the best results for tongue position and rounding. '_C analysis yielded the best results for tongue root position and height. Analysis evaluation considers Wilks' Lambda, which is the equivalent of ANOVA's F in discriminant analysis, and tolerance (1 - R^2), which expresses the extent to which each variable is uncorrelated with the others. Tolerance below 0.01 is to be avoided.

3.1.1.Tongue position

A three consonant subset – /r, ʒ, g/ – was found to be sufficient to classify vowels into front and back. Assuming that /r/ is dental or alveolar, all lingual places of articulation are implied in this choice.

Table 2: BSDA selection for tongue position from co-occurrence frequencies in 'C_ .

N= 7	**Discriminant Function Analysis** No. of variables in model: 3 Grouping: **Position** (2) Wilks' Lambda: 0.01145 approx. F (3,3)=86.307 p< 0.0021			
	Wilks' λ	p	Tolerance	R^2
r	0.0464	0.0565	0.0469	0.9531
ʒ	0.0594	0.0383	0.0778	0.9222
g	0.4396	0.0018	0.0201	0.9799

3.1.2.Rounding

The best subset for distinguishing round from unround vowels is also as small as 3 consonants.

Table 3: BSDA selection for rounding from co-occurrence frequencies in 'C_ .

N= 7	**Discriminant Analysis** No. of variables in model: 3 Grouping: **Rounding** (2) Wilks' Lambda: 0.01928 approx. F (3,3)=50.864 p< 0.0045			
	Wilks' λ	p	Tolerance	R^2
v	0.7488	0.0018	0.0271	0.9729
d	0.1818	0.0152	0.0483	0.9517
z	0.1951	0.0136	0.0683	0.9317

Note that all of the above are anterior.

3.1.3.Tongue height

BP height is a three-way distinction: high, mid, and low. Recall that 'C_ analysis performed poorly with it. By contrast, '_C analysis achieved 100% correct discrimination with a subset of only 3 consonants.

Table 4: BSDA selection for tongue height from co-occurrence in '_C.

N= 7	**Discriminant Analysis** No. of variables in model: 3 Grouping: **Tongue Height** (3) Wilks' Lambda: 0.00034 approx. F (6,4)=35.742 p< 0.0020			
	Wilks' λ	p	Tolerance	R^2
b	0.0398	0.0084	0.0726	0.9274
ɾ	0.0279	0.012	0.0132	0.9868
m	0.0152	0.0221	0.028	0.972

Again, all are anterior: two labials plus the tap.

3.1.4.Tongue root position

The relevant phonotactic context for distinguishing between /e, o/ and /ɛ, ɔ/ is also '_C.

Table 5: BSDA selection for tongue root position from co-occurrence frequencies in '_C.

N= 7	**Discriminant Analysis** No. of variables in model: 3 Grouping: **Tongue Root Position** (2) Wilks' Lambda: 0.01017 approx. F (4,2)=48.684 p< 0.0202			
	Wilks' λ	p	Tolerance	R^2
ɾ	0.3971	0.0129	0.0216	0.9784
t	0.0459	0.1178	0.0375	0.9625
s	0.0147	0.4456	0.0898	0.9102
m	0.4489	0.0114	0.0211	0.9789

Note, again, that all of the above are anterior.

3.2.Consonant Classes

While both '_C and 'C_ performed poorly in classifying consonants, V'_V gave 100% correct results with subsets of 4 to 6 vowel pairs.

3.2.1.Place of articulation

Five pairs are needed to classify the 19 consonants into 4 places of articulation, namely: labial, dental/alveolar, palatal, and velar.

Table 6: BSDA selection for place of articulation from co-occurrence frequencies in V'_V.

N= 19	**Discriminant Analysis** No. of variables in model: 5 Grouping: **Place of Articulation** (4) Wilks' Lambda: 0.01059 approx. F (15,30)=8.6035 p< 0.0000			
	Wilks' λ	P	Tolerance	R^2
o_e	0.0555	0.0003	0.1562	0.8438
u_e	0.0371	0.0025	0.1544	0.8456
o_ɛ	0.0887	0.0000	0.145	0.855
e_u	0.0323	0.0052	0.2143	0.7857
o_u	0.048	0.0006	0.1413	0.8587

Most differ as to rounding and tongue position.

3.2.2.Manner of articulation

Five pairs are needed to discriminate among the 4 manners of articulation, namely: stops, fricatives, liquids, and nasals.

Table 7: BSDA selection for manner of articulation from co-occurrence frequencies in V'_V.

N= 19	**Discriminant Function Analysis** No. of variables in model: 5 Grouping: **Manner of Articulation** (4) Wilks' Lambda: 0.01382 approx. F (15,30)=7.6224 p< 0.0000			
	Wilks' λ	p	Tolerance	R^2
o_i	0.0757	0.0002	0.0808	0.9192
a_e	0.0754	0.0002	0.0751	0.9249
u_a	0.0728	0.0003	0.1848	0.8152
a_ɔ	0.0528	0.0016	0.2363	0.7637
a_u	0.0409	0.0061	0.1471	0.8529

The predominance of back vowels and the presence of /a/ in all but one pair are worth noting.

3.2.3.Obstruence

Only four pairs are needed to classify the 19 consonants into sonorants and obstruents.

Table 8: BSDA selction for obstruence from co-occurrence frequencies in V'_V.

N= 19	**Discriminant Analysis** No. of variables in model: 4 Grouping: **Obstruence** (2) Wilks' Lambda: 0.22919 approx. F (4,14)=11.771 p< 0.0002			
	Wilks' λ	P	Tolerance	R^2
a_e	0.378	0.0093	0.3251	0.6749
a_ɛ	0.3808	0.0088	0.0944	0.9056
e_u	0.5068	0.001	0.1312	0.8688
a_u	0.7543	0.0001	0.0381	0.9619

Note recurring /a_e/ and /a_u/, in addition to the presence of front/back and/or tongue root distinctions in most pairs.

3.2.4.Voicing

Although voicing is binary and partly redundant with obstruence, this was the distinction that required the most pairs, namely, six.

Table 9: BSDA selection for voicing from co-occurrence frequencies in V'_V.

N= 19	**Discriminant Function Analysis** No. of variables in model: 6 Grouping: **Voicing** (2) Wilks' Lambda: 0.15268 approx. F (6,12)=11.099 p< 0.0003			
	Wilks' λ	P	Tolerance	R^2
u_e	0.3711	0.0014	0.2151	0.7849
a_a	0.5247	0.0002	0.0849	0.9151
o_a	0.8669	0.0000	0.017	0.983
i_ɔ	0.6166	0.0001	0.0791	0.9209
a_o	0.3079	0.0044	0.1655	0.8345
o_o	0.7141	0.0000	0.0329	0.9671

Note that two pairs have identical vowels and four differ as to tongue position and rounding.

4.DISCUSSION

The BSDA selections seem to capitalize on two underlying aspects of the data: vowel and consonant similarities or differences; and vowel pair contrast or redundancy.

Table 2 stands alone in that non-anterior consonants predominate – as expected for *Tongue Position*. By contrast, anterior consonants are ubiquitous in Tables 3 through 5. The latter, in

particular, exhibits 3 dental/alveolars. As this place of articulation tends to favor pharynx expansion, interaction of the lips and tongue apex with the tongue body and root may well underlie the phonotactics of vowel opening.

As to Tables 6 through 9, tongue position distinctions predominate in the V'_V pairs classifying place of articulation, whereas manner distinctions are conveyed mainly by opening and tongue root contrasts.

For a summary of the phonetic relationships implied by quantitative phonotactics in BP, let us now look at two cluster analyses: vowels in 'C_ and consonants in V'_V. The distance metric is 1 minus Pearson's R. SAMPA script is used.

Both figures are self-evident and show enough phonetic structure. Vowels split correctly into front and back, at least at lower levels. Similarly, consonants cluster reasonably well into place or manner categories.

Figure 1: Cluster analysis for V in 'C_ in CETEN.

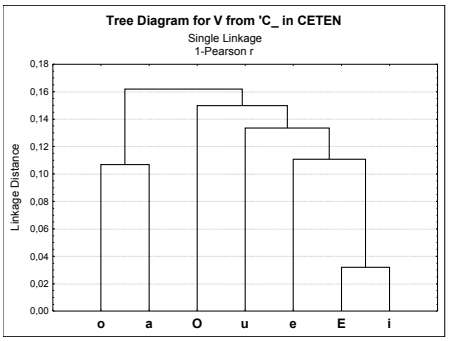
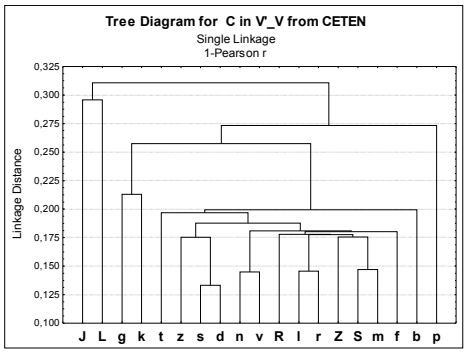


Figure 2: Cluster analysis for C in V'_V in CETEN.



A striking outcome of the analyses is the generality and abstractness of the resulting categories. While rhotics, especially the so-called strong /r/, are extremely variable phonetically – ranging from trill to approximant or fricative; and from dental to uvular or glottal – their BSDA and cluster analysis results are highly conservative. In all 3 corpora, they come out as dental/alveolar liquids. This can only be because co-occurrence biases are consistent within such categories, in spite of surface phonetic variability.

Preliminary tests of the discriminating context variables extracted from *CETEN* on *Mini-Aurélio* and *NURC-SE* have, on the average, yielded 75% correct classification, with statistically significant category distances for all contrasts. This is within the range generally considered successful for a test of discriminant analysis functions as classifiers.

The robustness of such results must, however, be tested against other, more representative corpora. Such materials are currently in preparation.

5.CONCLUSION

Taken together, these results give some important clues to the nature of natural segment taxonomies. First, they suggest that categories need not be pre-specified, for they can be inferred from context. Second, they indicate that classification might be based on whatever vowels and consonants have in common, for co-occurrence biases do imply "attraction" or "repulsion". Third, they point to the importance of two vocal tract regions which are not traditionally considered relevant to the commonalities of vowels and consonants: the anterior region, which is traditionally associated with consonants only, and the tongue root region, which is traditionally associated with vowels only.

Acknowledgements: This research was supported by *FAPESP,* the São Paulo State research agency (grant no. 01/00136-2), and *CNPq,* the Brazilian federal research agency (grant no. 304621/2003-0). Three anonymous ICPhS reviewers have provided helpful comments.

6.REFERENCES

[1] Clements, G. 1990. The role of the sonority cycle in core syllabification. In: Beckman, M., Kingston, J. (eds.), *Papers in Laboratory Phonology I.* Cambridge: Cambridge University Press, 283-333.

[2] Vitevitch, M., Luce, P. 1999. Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language* 40, 374-408.

[3] Maddieson, I. 1993. The structure of segment sequences. *UCLA Working Papers in Linguistics* 83, 1-7.

[4] Ferreira, A. 1977. *Minidicionário Aurélio*. Rio de Janeiro: Nova Fronteira.

[5] Linguateca. http://www.linguateca.pt/, visited June 5-06.

[6] Albano, E., Moreira, A., Aquino, P., Silva, A., Kakinohana, R. 1995. Segment frequency and word structure in Brazilian Portuguese. *Proc. ICPhS'95*, 3, 346-349.

[7] Albano, E., Moreira, A. 1996. Archisegment-based letter-to-phone conversion for concatenative synthesis in Portuguese. *Proc. ICSLP'96*, 3, 1708-1711.