The Interplay of Phonetics and Grammar in Determining V-to-'V Phonotactics

Eleonora C. Albano
Phonetics and Psycholinguistics Laboratory (LAFAPE), CP 6045,
State University of Campinas (UNICAMP), Campinas, SP 13084-971, Brazil
E-mail: albano@unicamp.br

ABSTRACT

This is a progress report on an ongoing probabilistic study of Brazilian Portuguese (henceforth BP) V-to-V phonotactics. The current focus is on the relationship between prestressed and stressed lexical vowels. The corpus consists of 21,208 words selected from an electronic dictionary. Prestressed vs. stressed vowel pair relative frequencies are computed to estimate vowel co-occurrence – i.e., pair occurrence – probabilities, in subsets divided into phonic, grammatical, and gross etymological classes. Biases are evaluated through observed to expected ratios (henceforth O/E) where E is estimated from the independent probabilities of the two vowels. Statistical analysis of the encountered biases indicates that not only syllable structure and grammatical class but also non-European etymology may affect pair probability. For example, close back vowel iteration is favored in non-European nouns; opening harmony is favored in verbs, whereas opening contour is favored in adjectives.

1. INTRODUCTION

Vowel-to-vowel relations are known to play an important role in the lexicons of so-called vowel harmony languages (e.g., Turkish, Finnish, etc. [1]). Semitic languages, on the other hand, exhibit plenty of consonant-to-consonant relations which may play a morphological role in their lexicons [2, 3].

Some speech production theories, in turn, argue, congenially, that consonants and vowels belong to separate channels in the speech “plan”, but must interact through implementation in the same vocal tract [4].

This paper attempts to contribute to the discussion of the complex issues involved in pinpointing the very basic phonic units of language by looking at vowel-to-vowel relations in the lexicon of Brazilian Portuguese, a language that has been claimed to exhibit lexical umlaut and ablaut [5], as well as phonological vowel harmony [6].

As the processes at issue, however amenable to rule-based description, are very much exception ridden, the approach that appeared scientifically the most promising was probabilistic phonotactics, along the lines set forth by [3, 7]. Since, nevertheless, such works focus on consonant to consonant relations – in addition to dealing with smaller *corpora* (and a smaller set of possible combinations) –, the methodology adopted here is not quite the same as theirs.

2. BACKGROUND

Brazilian Portuguese has 5 prestressed [i, e, a, o, u] and 7 stressed vowels [i, e, ɛ, a, ɔ, u], which yields 35 possible combinations, i.e., the following pairs, as reported in [8]: [i'i, e'i, a'i, o'i, u'i; i'e, e'e, a'e, o'e, u'e; i'ɛ, e'ɛ, a'ɛ, o'ɛ, u'ɛ; i'a, e'a, a'a, o'a, u'a; i'ɔ, e'ɔ, a'ɔ, o'ɔ, u'ɔ; i'o, e'o, a'o, o'o, u'o; i'u, e'u, a'u, o'u, u'u].

Word shapes range from monosyllables, which are rather rare, to eight-syllable polysyllables; the most frequent case being the trisyllable (34% of the total corpus described in section 3).

Stress may fall on the final, the penult, or the antepenult; penultimate stress being by far the most frequent (53% of the total corpus). Stress position is largely predictable from morphophonological context, but all rules are exception-ridden [9].

The subset of the Brazilian Portuguese lexicon under study thus comprises from finally stressed disyllables to polysyllables that run across the entire stress pattern – the grand total of which is 21,208 words.

3. METHODOLOGY

The corpus was extracted from an electronic version of Ferreira's dictionary [10], written at LAFAPE, which contains 27,074 words. The relevant word shapes were selected with the help of the same program, called *Listas.*

Listas has orthographic and phonetic script entries and allows for search by syllable structure and grammatical class.

Two experiments were run with the same dependent variable, i.e., vowel pair occurrence probability, as estimated by relative frequency in the *corpus*.

Since the data is linguistic, i.e., lexical statistical – and not behavioral (production or perception) –, simple vowel co-occurrence or pair probability was chosen over dependent probability, in order to avoid the direction issue (left to right or right to left), as direction could, obviously,

ISBN 1-876346-48-5 © 2003 UAB

be neither manipulated nor controlled for.

The independent variables for each of the experiments were: (1) statistically treatable syllable structure – with varying number and syllable affiliation of the intervening consonants –, which yields three categories: V.'CV, VC.'CV, and V.'CCV; and (2) statistically treatable grammatical class, which also yields three categories: noun, verb, and adjective. Other existing syllable structures and word classes were discarded on account of low frequency.

The nested design type was avoided because, since probabilistic data do not meet the normality assumption of parametric tests, the use of non-parametric tests is mandatory. And – at least the ones available in statistical analysis programs – only allow for relatively simple matrices.

O/E ratios were calculated as follows. O equals vowel pair probability, estimated as stated above, i.e.: pair frequency divided by total number of the words in the *corpus*. E equals the conditional independent probability of two vowels in the pair to co-occur, estimated as follows: total for prestressed vowel divided by grand total multiplied by total for stressed vowel divided by grand total.

Equiprobability of the 35 pairs was deemed inadequate to estimate their expected probability because it produces O/E ratios that enhance the frequency effects of single vowels. Thus, pairs such as [a'a] tend to be overrepresented and pairs such as [u'u] tend to be underrepresented, due to the high frequency of single [a] and the low frequency of single [u], as reported in [11].

The O/E ratios thus obtained were subjected to statistical analysis through the Kruskal-Wallis test; the 35 pairs being the cases, and the above mentioned syllable structure and grammatical class triads being the independent variables. In addition, the following procedures were applied: (1) pairs were ranked by O/E ratio for each independent variable; and (2) Spearman’s rank order correlation coefficients were calculated for all combinations of independent variables. Results are summarized in sections 4.1 and 4.2 below.

At the same time, a rationale was created to permit a subsegmental analysis of the trends encountered without adhering to a specific feature or gesture framework.

This consisted of analyzing pairs into the possible combinations of two complementary constraints, harmony (sameness) and contour (difference), taken to act upon three phonetic dimensions: place, opening, and rounding. Such a maneuver yields six possible V-to-'V relations: place harmony (PH), place contour (PC); opening harmony (OH), opening contour (OC); and rounding harmony (RH), rounding contour (RC). Place and rounding were considered binary while opening was considered ternary for phonological and morphophonological reasons discussed at length in [11].

Table 1 below shows the constraint tabulation by pair. For convenience, pairs are ordered as if the vowel trapeze were always run through from [i] to [u].

Constraint/ Pair	Pair	Place	Opening	Rounding
1	i'i	PH	OH	RH
2	e'i	PH	OC	RH
3	a'i	PC	OC	RH
4	o'i	PC	OC	RC
5	u'i	PC	OH	RC
6	i'e	PH	OC	RH
7	e'e	PH	OH	RH
8	a'e	PC	OC	RH
9	o'e	PC	OH	RC
10	u'e	PC	OC	RC
11	i'ɛ	PH	OC	RH
12	e'ɛ	PH	OC	RH
13	a'ɛ	PC	OH	RH
14	o'ɛ	PC	OC	RC
15	u'ɛ	PC	OC	RC
16	i'a	PC	OC	RH
17	e'a	PC	OC	RH
18	a'a	PH	OH	RH
19	o'a	PH	OC	RC
20	u'a	PH	OC	RC
21	i'ɔ	PC	OC	RC
22	e'ɔ	PC	OC	RC
23	a'ɔ	PH	OH	RC
24	o'ɔ	PH	OC	RH
25	u'ɔ	PH	OC	RH
26	i'o	PC	OC	RC
27	e'o	PC	OH	RC
28	a'o	PH	OC	RC
29	o'o	PH	OH	RH
30	u'o	PH	OC	RH
31	i'u	PC	OH	RC
32	e'u	PC	OC	RC
33	a'u	PH	OC	RC
34	o'u	PH	OC	RH
35	u'u	PH	OH	RH

Table 1: Pairs tabulated as to H or C, by phonetic dimension.

Preferred and rejected pairs were selected by setting the criterion for preference at O/E>=1.1, and the criterion for rejection at O/E=< .9. Fisher's exact test was used to inquire into association of preferred or rejected pairs with specific harmony and contour types.

This procedure ended up calling attention to the bias favoring [u'u] in nouns, and led to a post-hoc investigation of its possible etymological origin, which is reported below.

4. RESULTS

4.1 Syllable structure

The Kruskal-Wallis test on syllable structure yielded the following results: for H=180.2091, with df = 34, p=0.0000. This indicates a highly significant effect of syllable structure on pair ranking by O/E ratio. Thus, intervening consonants probably affect vowel choice in the BP lexicon.

Table 2 summarizes the Spearman rank correlation coefficient results:

Syl.vs Syl.	%corpus	V ." CV	V."CCV	VC."CV	V(C)."(C)CV
V ." CV	58.148%	1	0.256863	0.108683	0.960644
V."CCV	2.404%		1	0.067227	0.289496
VC."CV	18.046%			1	0.267647
V(C)."(C)CV	78.599%				1

Table 2: Syllable structure relations, as indicated by Spearman's rank correlation coefficients.

The only significant correlation (p=0.0000) is that of V."CV (majority set) with the total set. This reinforces the above interpretation about the role of syllable structure and intervening consonants.

4.2 Grammatical class

The Kruskal-Wallis test on grammatical class yielded the following results: for H=169.9071, with df = 34, p=0.0000.

This indicates a highly significant effect of grammatical class on pair ranking by O/E ratio.

Table 3 summarizes the Spearman rank correlation coefficient results:

Gr. & Gr.	%corpus	Noun	Adjective	Verb	N+A+V
Noun	45.346%	1	0.241176	0.185714	0.241176
Adjective	18.394%		1	0.209524	0.42507
Verb	15.117%			1	0.169888
N+A+V	78.857%				1

Table 3: Grammatical class relations, as indicated by Spearman's rank correlation coefficients.

All correlations are low and non-significant. This reinforces the inference that grammatical class affects vowel pair choice. Together, these results point to a possible role of derivational morphology on vowel combination.

4.3 Preference and rejection

Associations were found among preference and/or rejection and given constraints, as laid out in Table 1. The number of preferred or rejected pairs supporting a given constraint was subjected to Fisher's exact test for each grammatical class. There are significant associations between: (1) OH and preference in verbs (p=.0307); (2) OC and rejection in verbs (p=.0047); and (3) OC and preference in adjectives (p=.0008).

The same tests were performed with the syllable structure results, and no association was found.

This suggests that H and C are randomly distributed in the BP lexicon, but get stratified locally, for reasons that may have to do more with the grammar than with the phonetics.

4.4 Etymology

Further support for the above hypothesis comes from the post-hoc study on etymology.

A gross etymological categorization of nouns in [u'u] was inserted in the *corpus* manually to further inquire into the bias of nouns for this pair (O/E=1.9206).

Words of Amerindian and African origin were labeled non-European.

When these are removed from the *corpus*, O/E falls dramatically (O/E =.7736).

Also, a marginal association between preference and OC emerges in nouns when such non-European words are discarded (p =.0867).

5. CONCLUSIONS

These results support the conclusion that the phonetics and the grammar interact in determining V-to-'V phonotactics, and call for further investigation of the complex issues involved.

They also suggest that opening contour may be a lexical index of prosodic strength [12], a hypothesis which poses challenging questions concerning the relationship of speech to language.

Acknowledgements

Research funded by CNPq (grant no. 52.4110/96-4 LL) and FAPESP (grant no. 01/00136-2). Plínio Barbosa, Orlando B. Coelho, Edson Françozo, Ana Luiza Navas, and Leonardo Couto Franco de Oliveira have contributed helpful comments. All errors are the responsibility of the author.

REFERENCES

[1] M. Kraemer, “Vowel harmony and correspondence theory”. Unpublished Heinrich-Heine University, Düsseldorf, Germany, 2001.

[2] J. McCarthy, *Formal problems in Semitic morphology and phonology*. Bloomington, In: Indiana University Linguistics Club, 1982.

[3] J. Pierrehumbert, “Dissimilarity in the Arabic verbal roots.” *Proceedings of the North East Linguistics Society*, 23: 367-38, 1993.

[4] O. Fujimura, “Phonology and phonetics: a syllable-based model of articulatory organization”, *Journal of the Acoustical Society of Japan*, vol. 13 (E), pp. 39-48, 1992.

[5] A. A. Cavacas, *A língua portuguesa e sua metafonia*. Coimbra: Imprensa da Universidade de Coimbra, 1920.

[6] J. M. Câmara Jr., *História e estrutura da língua portuguesa*. Rio de Janeiro: Padrão, 1976.

[7] S. Frisch, “Temporally organized lexical representations as phonological units”, in *Papers in Laboratory Phonology V: Acquisition and the Lexicon*, M. Broe & J. Pierrehumbert, Eds, pp. 283-298. Cambridge: Cambridge University Press, 2000.

[8] E. C. Albano, “V-to-'V phonotactics, syllable structure, and morphological productivity,” Poster presented at the 8th Conference on Laboratory Phonology, New Haven, Yale University, on June 28th, 2002. Download from http://www.lafape.iel.unicamp.br/.

[9] E. Maia (née C. Albano), “Phonological and lexical processes in a generative grammar of Portuguese.” Unpublished Brown University doctoral dissertation, 1981.

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

[11] E. C. Albano, *O gesto e suas bordas: esboço de Fonologia Acústico-Articulatória do português brasileiro.* Campinas: Mercado de Letras, 2001.

[12] C. Shi, personal communication. ATR, Nara, Japan, September 6th 2002.