SEGMENT FREQUENCY AND WORD STRUCTURE IN BRAZILIAN PORTUGUESE

Eleonora C. Albano, Agnaldo A. Moreira, Patrícia A. Aquino, Adelaide H. P. Silva and Regis K. Kakinohana
Laboratório de Fonética Acústica & Psicolingüística Experimental
Universidade Estadual de Campinas, Campinas, S. P., Brazil

ABSTRACT

Segment frequencies in two representative *corpora* of Brazilian Portuguese are interpreted as evidence for a lexically-based phonetic pattern. The pattern consists in underexploiting the vowel inventory, allowing for much coarticulation to the right of the vowel and favoring auditorily salient segments to the left of the vowel.

INTRODUCTION

Maddieson [1] convincingly argues that cross-linguistic frequency studies of segments and segment sequences may help clarify the interplay of articulatory economy and acoustic/auditory distinctiveness in shaping sound patterns. At the same time, he regrets the insufficiency of the current statistical knowledge about the sound units of the world's languages. The study of a particular language may thus contribute directly to fill this empirical gap, shedding at least some indirect light on the theoretical issues.

There are nevertheless some more direct ways to gain theoretical insight from the statistical study of a single language. If the database is large, rich and carefully coded, segment frequencies may be taken as an estimate of the language's phonetic preferences, whatever their ultimate articulatory or acoustic interpretation may be. The question then arises whether such preferences are consistent with what is independently known about the language's sound pattern. For example, do preferred and avoided segments form phonetically consistent classes? Are such classes related to allophony and sound change trends? Do they agree with other aspects of the pattern such as syllable structure? If so, can segment preferences throw further light on the pattern itself? In particular, can they contribute to the phonetic interpretation of the segmental notation?

This paper is a progress report on an investigation of such questions in Brazilian Portuguese. Segment preferences were found to be stable in the language by comparing a dictionary to a large oral language database whose unity had been established through the comparison of several samples. The local differences between the dictionary and the oral language database, rather than obscuring their correlation, were useful in interpreting the segment preference pattern that emerged.

This pattern can be summarized as follows: Brazilian Portuguese underexploits its vowel contrasts and balances its consonant preferences between those which can be loosely coarticulated with the vowel (cf. [2]) and those which rank high in acoustic/auditory distinctiveness (cf. [3]). Such an interpretation is consistent not only with the historical and synchronic trends of the language's phonology but also with a deductive phonetic stance on the problem of sound organization and selection, as currently advocated by some leading phoneticians [4, 5, 6].

METHOD

The dictionary sample consists of the entire set of orthographic entries (27,074) from Ferreira [7]. The spoken language sample consists of 57 orthographically transcribed tapes from a public database originally collected for a nationwide oral language survey (*Projeto Norma Urbana Culta*, henceforth NURC). The NURC team had recorded a representative number of highly educated adults from five major Brazilian state capitals in various lecture and conversation settings. We have used all the available electronic data, which amounts to 72 hours of recording on 58 males and 43 females.

Both *corpora* were converted from official orthography to an abstract phonological code with the aid of a computer program which determines syllabification and lexical stress and resolves orthographic ambiguities by means of rules (including some grammatical ones) and exception lists. The initial version of this code (to be further discussed below) distinguishes a total of seven vowels [i, e, ɛ, a, ɔ, o, u] and twenty-four consonants [p, b, m, f, v, t, d, s, z, n, ɹ, l, ʎ, ɲ, j, j̃, ʃ, ʒ, k, g, w, w̃, ʁ, ɴ]. Consonants are further classified as to whether they occur to the left (L) or to the right (R) of the vowel. All vowels are considered oral, since nasal vowels and diphthongs are dealt with by allowing three nasals (unspecified ɴ, labiovelar w̃, and palatal j̃) in R position. For morphophonological coherence, orthographic e and o are always assigned to /e/ and /o/, though phonetic [i] and [u] tend to surface in most post-stressed and some pre-stressed contexts.

To further inquire into the differences between the lexicon and connected speech, two subset files were derived from the NURC *corpus*: GRAMNURC, containing all the occurrences of grammatical words, and CONTNURC, containing the remaining (content) words. A computer program was then elaborated to make the segment counts in the dictionary (henceforth DIC) and in the NURC files.

The lexicographers' convention of listing infinitives, which are rather infrequent relative to other verb forms in the language, introduced a bias for /ɹ/ in the DIC file, which was corrected by simply ignoring infinitive /ɹ/'s in the counts.

RESULTS

Figure 1 displays the segment percentages in the DIC and NURC data, with syllable positions shown in stacks. Spearman's ρ is 9.4 ($p<0.001$), which indicates a high correspondence between the two rankings. At this point, double stacks apply only to [s, ɹ, l], since the functional distinction between [i, u] and [j, w] is built into the segment code. It is already clear, in any case, that vowel selection does not conform to a "maximal contrast" [8] or "quantal" [2] principle, as the high vowel and semivowel bars generally add to less than the mid vowel bars.

Equally intriguing is the question of consonant selection. The ability to occupy R position clearly pushes [s, ɹ] into higher ranks, but the same cannot be said of [l], at least in the NURC data. As [ɴ], in turn, ranks very high in both *corpora* without ever occurring in L position, a possible explanation for the high ranks of R consonants might be freedom to coarticulate with the vowel. This interpretation cannot, however, be pushed too far, since it leaves unexplained the lower rank of [l] as well as the general pattern of L consonants, where coronals rank higher than labials and velars.

On the other hand, the presence of a certain drive for distinctiveness is very clear in these data. Seven out of the ten most frequent consonants in both samples correspond to the "salient" segments which Stevens and Keyser [3] deduce from the combination of their "primary" features (sonorant, continuant, and coronal) with their "enhancing" features (voice, strident, nasal, etc.). This set consists of four voiceless obstruents [p, t, s, k], two consonantal nasals [m, n], and one lateral [l]. As should be expected, ranks are slightly higher in the DIC than in the NURC *corpus*, though the total percentages are about the same (26.4% *vs* 27.4%, respectively).

Let us now look at two rearrangements of the same data.

Figure 2 is a remake of Figure 1 with high vowels and semivowels collapsed into a single category (i.e., [j→i] and [w→u]). Note that the linearity deliberately adopted in the initial notation has been abandoned here: nasal semivowels are reassigned simultaneously to high vowels and [ɴ] (i.e., [j̃,w̃→ j+ɴ, w+ɴ]).

Figure 3 displays the count made on the GRAMNURC and CONTNURC files, with the reduced (less linear) inventory.

It is now clear that high vowels pattern with R/L consonants in ranking between L consonants and [a, e, o]. This tendency is evident in the NURC data, and somewhat more dispersed in the DIC data, due to the high rates of [t, d]. In addition, the conformity of L consonants to Stevens & Keyser's predictions is much clearer in these graphs. Note, for example, that the set of less optimal L consonants [b, f, v, z, ʎ, ɲ, ʃ, ʒ, ʁ] - which amount to 10.5 and 7.5 % in the DIC and NURC data, respectively - is very small in the GRAMNURC file (1.17%), where prosodic weakness increases the need for efficient encoding of selected distinctions.

The role of the other not so "salient" L segments [d] and [ɹ] is also clarified by comparing the GRAMNURC with the CONTNURC data. Their near complementarity in these graphs suggests that the crowded coronal region of the language's consonant space is sensitive to lexical strata, with grammatical words accounting for most of the [d]'s and content words, in turn, largely accounting for the rather conspicuous auditory contrast between [t] and [ɹ].

It follows from all of the above that the reason why Brazilian Portuguese does not fully exploit the basic vowel triangle in lexical distinctions is that acoustic distinctiveness and articulatory economy divide the spoils within the syllable: onsets tend to be auditorily salient while rhymes tend to be free from any restrictions on coarticulation. Non-high vowels are preferred underlyingly because they coarticulate easily with most R segments including high vowels while leaving enough room for surface raising, which does in fact occur very frequently.

CONCLUSION

A simple conceptual and graphic analysis of segment preferences in Brazilian Portuguese has shown that both an ideal lexicon (DIC) and the lexicon actually used in connected speech (NURC) exhibit trends which are consistent with processes observed in the language's synchronic and diachronic phonology. These trends can be summarized as a preference for loosely coordinated gestures to the right of the vowel, counterbalanced by a preference for auditorily salient gestures to the left of the vowel.

The tension generated by this asymmetry has been responsible for the historical loss of stops in syllable final position and for the emergence of syllable initial approximants in word medial position. It is also currently responsible for a number of phonological processes taking place in various dialects such as rhotacization of laterals in syllable initial clusters and loss of lateral, sibilant and rhotic gestures in some medial and final environments.

However preliminary, the results of this study support experimental approaches to phonology and point to the need for further quantitative studies of Brazilian Portuguese. In addition, they indirectly support theories which translate segments into articulatory gestures, particularly those which allow for the interaction of auditory and motor factors in their selection and distribution [9].

ACKNOWLEDGEMENTS

Support from CNPq (grant no.50.0400/90) and FAPESP (grant no.93/0565-2) is gratefully acknowledged.

REFERENCES

[1] Maddieson, I. (1993) "The structure of segment sequences", *UCLA Working Papers in Phonetics*, vol. 83, pp. 1-7.
[2] Sproat, R. & O. Fujimura (1993) "Allophonic variation in English /l/ and its implications for phonetic implementation", *Journal of Phonetics*, vol. 21, pp. 291-311.
[3] Stevens, K. & J. Keyser (1989) "Primary features and their enhancement in consonants", *Language*, vol. 65, pp. 81-106.
[4] Stevens, K. (1989) "On the quantal nature of speech", *Journal of Phonetics*, vol. 17, pp. 3-45.
[5] Ohala, J. (1990) "There is no interface between phonology and phonetics: a personal view", *Journal of Phonetics*, vol. 18, pp. 153-171.
[6] Lindblom, B. (1990a) "Phonetic content in phonology" *PERILUS*, vol. XI, pp. 101-118.
[7] Ferreira, A. (1977) *Minidicionário Aurélio*. Rio de Janeiro: Nova Fronteira.
[8] Jakobson, R. (1968 [1941]) *Child language, aphasia, and phonological universals*. The Hague: Mouton.
[9] Lindblom, B. (1990b) "Models of phonetic variation and selection", *PERILUS*, vol. XI, pp. 65-100.

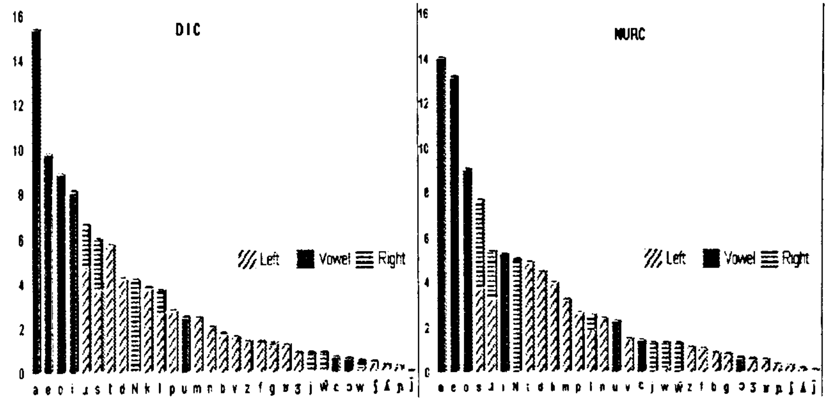

Fig. 1- Segment percentages in the DIC and the NURC samples (original inventory).

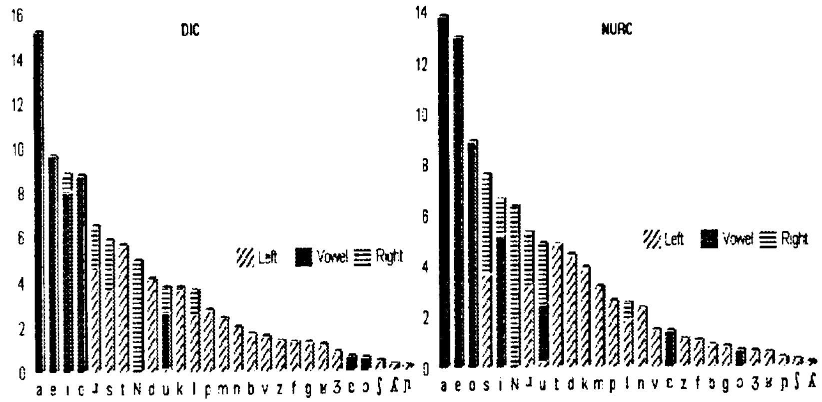

Fig. 2 - Segment percentages in the DIC and the NURC samples (reduced inventory).

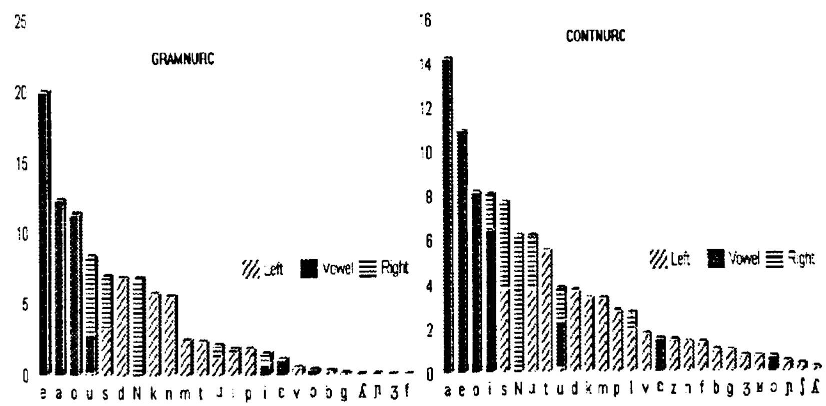

Fig. 3 - Segment percentages in the GRAMNURC and CONTNURC samples.