DOI: 10.20396/joss.v11i00.16537

# ON THE EMERGENCE OF [n] IN THE DERIVATION OF NASAL-FINAL WORDS IN BRAZILIAN PORTUGUESE

**SCHWINDT, Luiz Carlos**[1*]
**ABAURRE, Maria Bernadete**[2]

[1] Postgraduate Program in Language Studies – Federal University of Rio Grande do Sul – ORCID: https://orcid.org/0000-0003-0533-589X
[2] Postgraduate Program in Linguistics – State University of Campinas – ORCID: https://orcid.org/0000-0002-5467-468X

**Abstract:** *This article deals with the correspondence between words closed by [n] followed by a vowel-initial suffix (e.g., ca[n]al 'channel', tupi[n]ismo 'Tupinism') and their respective bases in Brazilian Portuguese (e.g., cano 'pipe', tupi 'Tupi'). The data sources for the analysis are the Corpus Brasileiro, representing the lexicon in use, and a pseudoword test, representing the potential lexicon. Two representational hypotheses are contrasted in the analysis of this correspondence: the abstract approach, in which the bases are assumed to be closed by a VN structure, and the concrete approach, according to which [n] is part of the base or a product of epenthesis. The selection of the pattern VN was assumed as the response variable in the statistical analysis. According to the logistic regression test, final-stress bases and mid and high vowels preceding [n]Vsuffix are predictors of the VN pattern for the two samples. In the potential lexicon, the interaction between high frequency lexical strings and the suffixes -icV and -ismV as well as the random variables 'participant' and 'pseudoword' also contribute to the selection of the base pattern VN. The results confirm the plausibility of the abstract approach in the analysis of [n]Vsuffix forms in Brazilian Portuguese.*

**Keywords:** Nasal diphthongs; Consonant epenthesis; Suffix derivation; Irregular plural; Morphophonology.



[*] Corresponding author: schwindt@ufrgs.br

## 1 Introduction

The representation of nasal vowels/diphthongs in word-final position, in particular the diphthong [ɐ̃w̃], most often stressed, is challenging for studies of Portuguese morphophonology. The main phonological problem concerns the definition of the underlying form (UR) of these sounds: whether originally a nasal vowel/diphthong or a vowel followed by a nasal consonant. In this article, we call the first approach concrete (CA), as opposed to the second, which we call abstract (AA).

CA is mainly based on the principle of parsimony (or Ockham's Razor), according to which, when faced with two ways of explaining a given phenomenon, the simplest one is always the most adequate. This principle can also be defined by the maxim borrowed from computer science *what you see is what you get*. In the relationship between phonetics and phonology, this means opposing the distinction between a concrete and a more abstract level of analysis. On the other hand, the object of this study is among the arguments of AA: in words closed by a nasal vowel or by [ɐ̃w̃], an alveolar nasal consonant, [n], which in principle is not present on the phonetic surface of simple forms (e.g., c[ɐ̃w̃] 'dog'), emerges categorically between base and suffix (e.g., ca[n]ino 'canine'). This argument is often questioned by CA advocates, especially as this consonant can occasionally also emerge in words not closed by a nasal segment (e.g., tupi → tupi[n]ismo 'Tupinism').

In this article, we explore the productivity of this phenomenon in Brazilian Portuguese (BP). The objective is mainly descriptive, that is, to map the bases related to the words formed by [n]suffix considering data from the lexicon in use and from the potential lexicon. The sources are samples from Corpus Brasileiro (CBras)[1] and from a Test involving pseudowords, respectively.

The text is organized as follows. In section 2, we present a non-exhaustive review of the phenomenon of nasality in coda position in Portuguese, emphasizing the problem addressed in the article, which concerns the emergence of a nasal consonant in morphologically derived forms related to simple forms closed by nasal vowels or nasal diphthongs. In section 3, we detail the methodological procedures adopted in the data collection and analysis of the two investigated samples. In section 4, we present and discuss the main results of the study. Finally, in section 5, we summarize the main findings of the research and its limits.

## 2 Word-final nasals in Portuguese

The phenomenon discussed in this article is articulated with several other phenomena in Portuguese, including the debate about the phonemic character of nasal vowels, the discussion about the opposition between underlying and derived diphthongs and the controversy about allomorphy in the formation of plurals of words closed by [ɐ̃w̃]. Regarding BP, these topics were studied, in different perspectives, by Camara Jr. (1–3), Lemle (4), Leite (5), Cagliari (6), Abaurre-Gnerre (7), Wetzels (8–10), Bisol (11, 12), Huback (13, 14), Cristófaro-Silva (15), Guimarães & Nevins (16), Becker et al. (17), Rizzato (18), Gomes, Prado & Amaral (19), Schwindt, Gaggiola & Petry (20), among other authors.

In BP, five of the seven phonemic vowels can be opposed in terms of nasality in the stressed position, as observed in the minimal pairs listed in (1).

| | | | | | |
|---|---|---|---|---|---|
| (1) | l[i]do | ‘read$_{\text{PastPart}}$’ | ≠ | l[ĩ]do | ‘beautiful’ |
| | m[u]do | ‘mute’ | ≠ | m[ũ]do | ‘world’ |
| | t[e]ta | ‘teat’ | ≠ | t[ẽ]ta | ‘try$_{\text{3rdSingPres}}$’ |
| | r[o]do | ‘squeegee’ | ≠ | r[õ]do | ‘prowl$_{\text{1stSingPres}}$’ |
| | t[a]pa | ‘slap’ | ≠ | t[ɐ̃]pa | ‘cover’ |

There are basically two hypotheses in the literature to account for this phenomenon, which differ

[1] http://corpusbrasileiro.pucsp.br/cb/Inicial.html

from each other in terms of abstraction. From the CA perspective, nasal vowels are treated as phonemic or monosegmental, a point of view supported by Head (21), Leite (5), Matta Machado (22), among others. From the AA perspective, they are treated as disegmental, with the nasality of the resulting vowel being interpreted as a product of assimilation of the consonant that follows an oral vowel in the underlying structure, as argued by Camara Jr. (1,3), Lemle (4), Cagliari (26), Bisol (11), among others.

This controversy reaches to some extent diphthongs. Although Camara Jr. (2) argues that decreasing diphthongs are *true diphthongs* (or underlying diphthongs, in generative terms), this does not apply to nasal diphthongs. In this case, the author also assumes the disegmental analysis: to the final portion of the root a theme vowel, -e, is affixed, which emerges in the most frequent plural form in the language, [õj̃s]. This theme vowel must be deleted to ensure well-formedness of the singular form, as illustrated with the simplified derivation in (2), adapted from Abaurre-Gnerre (7) and Wetzels (10).

| (2) | /padɾon+e/ | /padɾon+e+s/ |
|---|---|---|
| stress | pa'dɾone | pa'dɾones |
| regressive nasalization | pa'dɾõne | pa'dɾõnes |
| n deletion | pa'dɾõe | pa'dɾões |
| apocope (theme vowel deletion) | pa'dɾõ | NA |
| glide insertion | pa'dɾõw | NA |
| õ → ɐ̃ | pa'dɾɐ̃w̃ | NA |
| e → j | NA | pa'dɾõjs |
| progressive nasalization | pa'dɾɐ̃w̃ | pa'dɾõj̃s |
| | [pa'dɾɐ̃w̃] | [pa'dɾõj̃s] |
| | 'pattern' | 'patterns' |

This derivation can be problematized in at least two aspects. First, the full specification of the nasal consonant in coda is debatable. Camara Jr. (3) treated it as an archiphoneme, as it is subject to neutralization in this position (tending to agree in place of articulation with the following segment or, in some cases, with the preceding one). Underspecifying the nasal in the underlying form would require adding a step to (2) to account for the filling of place features). Second, the emergence of the less frequent plural alternants [ãj̃s] (e.g., capit[ɐ̃j̃s] 'capitains') and [ɐ̃w̃s] (e.g., irm[ɐ̃w̃s] 'siblings') implies, in both cases, forgoing the conversion from /o/ into [ɐ̃] predicted in (2). Also, the emergence of [ɐ̃w̃s] in particular (in principle the most regular alternant) demands the assumption that the theme vowel is -o instead of -e. Consequently, there is no apocope in this case: the theme vowel /o/ becomes [w] (as /e/ becomes [j] in the case of the other two plural alternants).

This complex picture led Abaurre-Gnerre (7), Wetzels (9, 10), Rizzato (18), among others, to assume CA in the analysis of these forms. According to this approach, the underlying representation of word-final stressed diphthongs is ṼG̃, as opposed to VN or VGN.

Nonetheless, one of the main arguments in defense of AA is the fact that an alveolar nasal consonant fully emerges when the bases under discussion are followed by vowel-initial suffixes. This is exemplified in (3) for the three plural patterns.

| | | | | |
|---|---|---|---|---|
| (3) a. vulcão / vulcões | 'volcano/es' | → | vulcâ[n]ico | 'volcanic' |
| b. mão / mãos | 'hand/s' | → | ma[n]ual | 'manual' |
| c. alemão / alemães | 'German/s' | → | alemâ[n]ico | 'Germanic' |

This thesis is criticized in studies such as those by Abaurre-Gnerre (7), Wetzels (9) and Guimarães & Nevins (16) for alleged lack of synchronic (or psychological) evidence. Formations such as *tupi[n]ismo 'Tupinism'* or *faraô[n]ico 'pharaonic'*, which are not derived, respectively, from tupĩ(n) or faraõ(n), appear as examples to suggest that the nasal consonant has no affiliation at the base. However, the fact that examples like these are rare — when contrasted with the high incidence of correspondence

between base/V(G)N/ and base[n]Vsuffix — points to the need of a more accurate analysis.

There are at least three alternatives for analyzing this correspondence between base and derived forms outside the scope of AA. The first alternative is to assume that forms like -nico and -nual, for example, are allomorphs listed together with -ico and -al. The second alternative is to consider that, in addition to morphemes and isolated words, sets of words or structures, such as *vulcão 'volcano'* ↔ *vulcânico 'volcanic'*[2], are paradigmatically related in our mental lexicon. Finally, a purely phonological alternative is to analyze the nasal consonant of the derived forms as epenthetic, which emerges as a repair strategy for malformed structures, such as hiatus avoidance, in the case of V+Vsuffix.

None of these explanations is exempt from criticism. The suffix allomorphy hypothesis demands that formations such as *ladrão 'thief'* → *ladroagem* / **ladro[n]agem 'thievery'* are accounted for. Listing word pairs, on the other hand, demands additional memory storage for a process that is quite regular (or frequent) from a morphological point of view. Finally, the epenthesis hypothesis requires dealing with the fact that base[ɐ̃w̃]/[n]Vsuffix correspondence is much more recurrent than baseV/base[n]Vsuffix. Furthermore, it is necessary to justify the choice of [n] as the epenthetic consonant in this case.

In this text, we do not intend to choose the best among these hypotheses. The discussion about the representations and constraints involved in the emergence of the correspondence base[ɐ̃w̃] → base[n]suffix, although fundamental, is not an immediate objective of this work. Our purpose is to report an empirical exercise that responds to a demand signaled in Abaurre-Gnerre (7), namely, that the best formalization of this phenomenon involves mapping the elements that account for their effective use and learnability[3].

We assume that morphophonological knowledge can be learned by examining both the lexicon in use and the potential lexicon. We call lexicon in use the forms effectively employed in communication, which can be accessed in corpora with large amounts of data. By potential lexicon we understand the latent formations of a language: words that are phonologically and morphologically well-formed, but not yet instantiated in the language by any type of blocking, including mere lexical inertia. Among other methodological strategies, the potential lexicon can be accessed by tests involving pseudowords or logatomes — semantically contextualized forms that obey general well-formedness properties of the language despite not being listed in the speakers' mental dictionary.

## 3 Method

Broadly, the question of this work is the following: which variables play a role in the selection of bases corresponding to [n]Vsuffix words?

The analysis we propose is unidirectional, that is, our view departs from the derived word to the base and not the other way around. We did not test the emergence of [n] from a certain type of base, because this would possibly lead to a quasi-categorical response in the case of bases closed by [ɐ̃w̃] and to a probable proliferation of structures in the case of other bases, including other intervening consonants. This can be seen in Figure 1 below, which presents percentages for potential consonants (not necessarily epenthetic) that can precede vowel-initial suffixes. Among these suffixes, we chose three highly recurrent ones in BP, -al, -ico and -ismo, based on the general, non-lemmatized sample of CBras. Although [n] appears among one of the most frequent consonants before these suffixes, it competes with other consonants, especially other alveolars, the most common segment class in the language in cases of epenthesis (26, 27).

[2] Abaurre-Gnerre (7) considers, in line with Natural Phonology, that this correspondence, historically inherited, is a phenomenon of a lexical nature, and suggests it be formalized by an expedient known as *via-rule*. (23, 24).

[3] Schwindt (25), analyzing the correspondence between bases closed by [w] and derived forms with the structure base[l]Vsuffix (e.g., papel 'paper' → papel[l]eiro 'papermaker'), proposes an analysis within the Optimality Theory framework for the learning of underlying representations, starting from an empirical exercise similar to the one presented in this article.

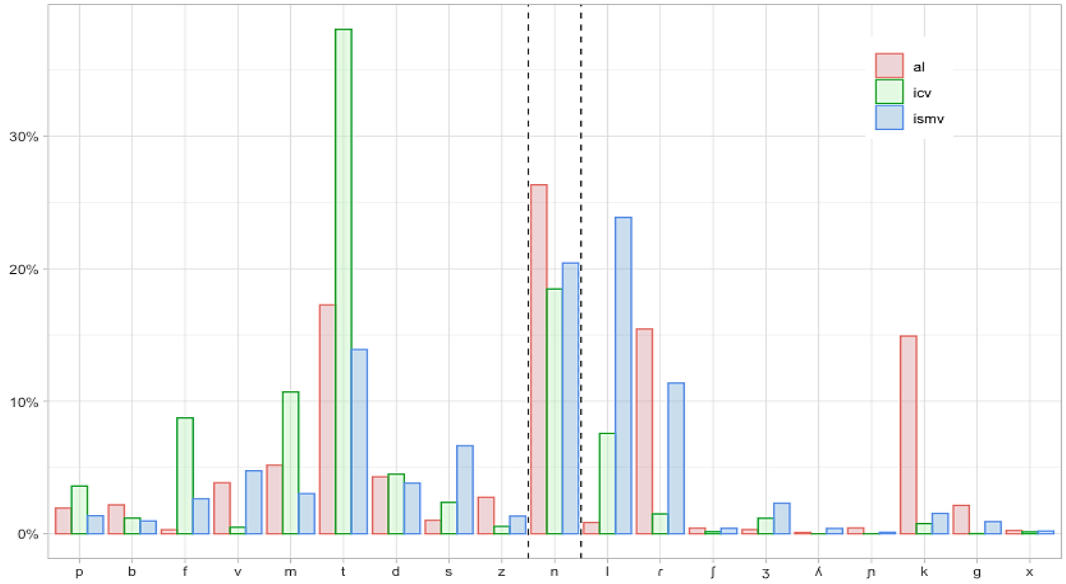


**Figure 1:** Consonants preceding vowel-initial suffixes – CBras

The two samples analyzed in this study are detailed below.

### 3.1 Lexicon in use

The lexicon items in use come from CBras, an online corpus that gathers BP speech and writing data produced between 2008 and 2010. The database contains 155,842 types and 691,758,151 tokens available for download. All the words that presented <n> preceded by a vowel and directly followed by the suffixes -al, -ico/a/s and -ismo/s in word-final position were extracted from these data to constitute the sample of this research. Different filters were applied to conform the data to the research objectives: (i) masculine, feminine and plural forms were combined; (ii) prefixes, as well as the first part of compounds, were deleted whenever the right base could be recognized as a possible base (isolated or part of a recurrent paradigm); (iii) identical items or items with minor spelling discrepancies were combined, and their individual lexical frequencies were recalculated after that. The final sample resulted in 1,574 lemmas and 5,456,903 tokens.

### 3.2 Potential lexicon

A pseudoword test prepared with the Google Forms tool and shared for 72 hours on the social network Facebook, between May 21 and 23, 2020, was answered by 210 subjects. Two non-native BP subjects and one subject who did not agree to answer it were excluded, resulting in 207 participants.

Given the nature of the phenomenon under analysis, these participants were not previously socially stratified, although their main social characteristics, informed in the introduction of the Test, were computed, as shown in Table 1.

**Table 1:** Social characterization of participants – Test

| Categories | (N = 207 participants) | % |
|---|---|---|
| Education | Elementary | 1.93 |
| | Secondary | 15.94 |
| | Higher | 82.13 |
| Degree course (concluded or not) | Languages / Literature | 46.86 |
| | Other (detailed in the questionnaire) | 35.75 |
| | NA | 17.39 |
| Gender | female | 65.70 |
| | male | 34.30 |
| Age | 15–81 years old (mean = 33.82) | |

The distribution of these social variables in relation to the response variable is presented in the next section.

The Test consisted of 20 logatomes created from typical BP syllable and stress patterns, 5 of which were distractor items. The pseudowords were presented in the derived form, always with a mono or disyllabic base preceding the structure [n]Vsuffix (-al/-icV/-ismV) in sentences containing a blank space to be filled in by the respective plural form. The presentation of the alternatives in the plural form, in addition to providing complimentary information for the research, is another distractor strategy to obtain more reliable bases in the singular form, the object we focused on in this Test. As it is a written test, participants were initially notified that all alternatives were graphically accented in their stressed syllable, despite the official stress rules of BP, with acute (′), circumflex (∧) and tilde (∼) distinguishing open, close, and nasalized vowels, respectively.

The pseudowords were created from 5 base patterns involving the height of the vowel that precedes [n] in the derived form: pattern I corresponds to forms for which a low vowel emerges (e.g., anal, ânico, anismo); pattern II, a mid-high vowel (eg onal, ônico, onismo); pattern III, a mid-high vowel in hiatus (aonal, aônico, aonismo); pattern IV, a high vowel (unal, único, unismo); the V pattern, finally, a rising diphthong or a hiatus (ianal, iânico, ianismo). Due to the extension limitation, and to ensure that all vowel heights were covered in the Test, we restricted the alternatives to items formed by back vowels preceding the vowel-initial suffix. Furthermore, all bases have between 2 and 4 syllables. In Table 2 below, we list the pseudowords used in the Test accompanied by examples from the lexicon on which they were based.

**Table 2:** Pseudoword groups – Test

| Groups | Pseudowords | | | Distractors | Lexicon examples |
|---|---|---|---|---|---|
| I | BEGANAL | LUDÂNICA | GOJANISMO | ZIGATISMO | cabanismo 'hut culture' |
| II | DOBONAL | ZADÔNICO | JALONISMO | GODATISMO | sônico 'sonic' |
| III | TAZAONAL | JALAÔNICO | ZARAONISMO | MEVAÓRICO | faraônico 'pharaonic' |
| IV | FAGUNAL | DELÚNICO | CHODUNISMO | BEGUZAL | descomunal 'monumental' |
| V | BAJIANAL | ZIVIÂNICO | BAVIANISMO | NABIÁZICO | asianismo 'Asianism' |

A test question is exemplified in (4).

(4) Marina aderiu, por fim, ao JALONISMO. Ressente-se, porém, de ter demorado tanto a ouvir os sábios _____.
'Marina finally joined the JALONISMO. However, she regrets having taken so long to listen to the wise _____.'
(A) jalões
(B) jalãos
(C) jalães
(D) jálos
(E) jalôs
(F) jalônes

Disregarding the distractor questions, the test sample resulted in 3,105 tokens.

### 3.3 Response variable

In view of the question that guides this research, we assumed the base pattern as the response variable for both samples. We started from an analysis of 7 plural patterns related to the derived forms for the CBras sample and 6 patterns for the Test sample, with the distribution shown in Table 3 below.

**Table 3:** Base Pattern – CBras and Test

| | CBras (N=1,574) | | | Test (N=3,105) | |
|---|---|---|---|---|---|
| Base | % | Example | | % | Example |
| õj̃s | 31.44 | adições | 'addition$_{Pl}$' | 38.29 | dobões |
| | | adicional | 'additional' | | dobonal |
| ãw̃s | 0.50 | órgãos | 'organ$_{Pl}$' | 7.02 | dobãos |
| | | orgânico | 'organic' | | dobonal |
| ãj̃s | 0.57 | catalães | 'catalan$_{Pl}$' | 11.98 | dobães |
| | | catalanismo | 'catalanism' | | dobonal |
| vns | 8.63 | comuns | 'common$_{Pl}$' | 11.67 | dobõns |
| | | comunismo | 'communism' | | dobonal |
| nvs | 51.48 | demônios | 'devil$_{Pl}$' | 14.56 | dobônes |
| | | demonismo | 'demonism' | | dobonal |
| deriv_n | 4.28 | mecânico/s | 'mechanical$_{Sing/Pl}$' | | |
| no_n | 3.09 | faraós | 'Pharaoh$_{Pl}$' | 16.55 | dôbos, dobôs |
| | | faraônico | 'pharaonic' | | dobonal |

The patterns in Table 3 have been reconfigured. Firstly, the distinction between the alternants [õj̃s], [ẽw̃s] and [ẽj̃s] is not relevant for this specific study, as they converge to a single singular base, [ẽw̃]. Second, it is also irrelevant to distinguish such forms from the alternant *vns*, which surface in the language as a nasal vowel (e.g., l[ẽ]s 'wool$_{pl}$', f[ẽ]s 'fans') or as a nasal diphthong (e.g., t[õw̃]s 'tones', com[ũw̃]s 'common$_{pl}$'). For this reason, these bases have been grouped into one category, VN, which reflects AA. On the other hand, we group together in a single category forms whose base contains an alveolar nasal consonant in the onset position, either *nvs* or *deriv_n*, as well as the specific contexts for epenthesis of [n], *no_n*. We call this category no_VN, which reflects CA. The motivation for grouping these variables is distributional and representational. In terms of distribution, there is a shortage of bases in the pattern *derivative_n* and *no_n* categories, whereas between bases closed by diphthong or nasal vowel and bases *nvs* there is less imbalance. From a representational point of view, VN bases seem more difficult to be deduced by the language learner from plural forms, while no_VN bases are more easily deducible, either because they already contain [n] in their structure, in the cases of nV, or because the emergent [n] is one of the segments supposedly reserved in the language to perform epenthesis. The response variable thus resulted in the binary variable exemplified in Table 4, in which a clear inversion

of preference can be observed, with no_VN being the predominant base pattern in CBras and VN in the Test.

**Table 4:** Response variable

| | CBras (N=1,574) | Test (N=3,105) | | |
|---|---|---|---|---|
| Base | % | % | Corresponding bases | Hypothesis |
| VN | 41.49 | 68.89 | ões, ãos, ães, vns | AA |
| no_VN | 58.51 | 31.11 | nvs, deriv_n, no_n | CA |

The social characteristics of the Test participants, not controlled in this study, present a balanced distribution in relation to the response variable.

Regarding education and degree course, the result shown in Figure 2 allows us to discard the hypothesis that more educated informants or those who attended courses related to languages or linguistics might prefer VN (the one that would be associated with greater metalinguistic reflection).

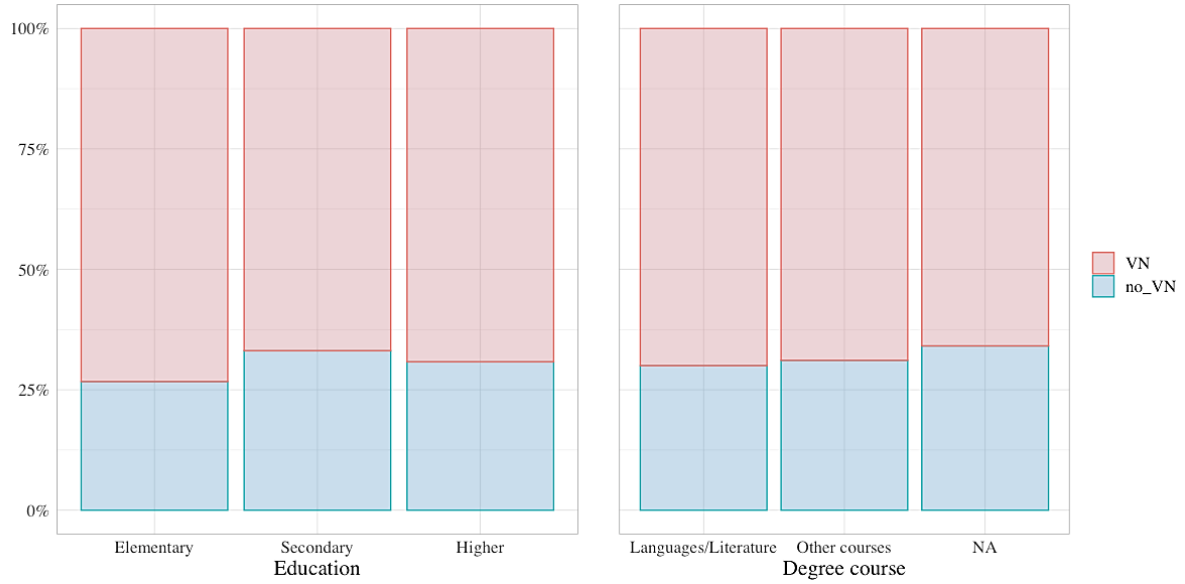


**Figure 2:** Education and Degree Course – Test

Gender and age also present a fairly uniform distribution, as seen in Figures 3 and 4 below, confirming the irrelevance of extralinguistic properties for this particular analysis.

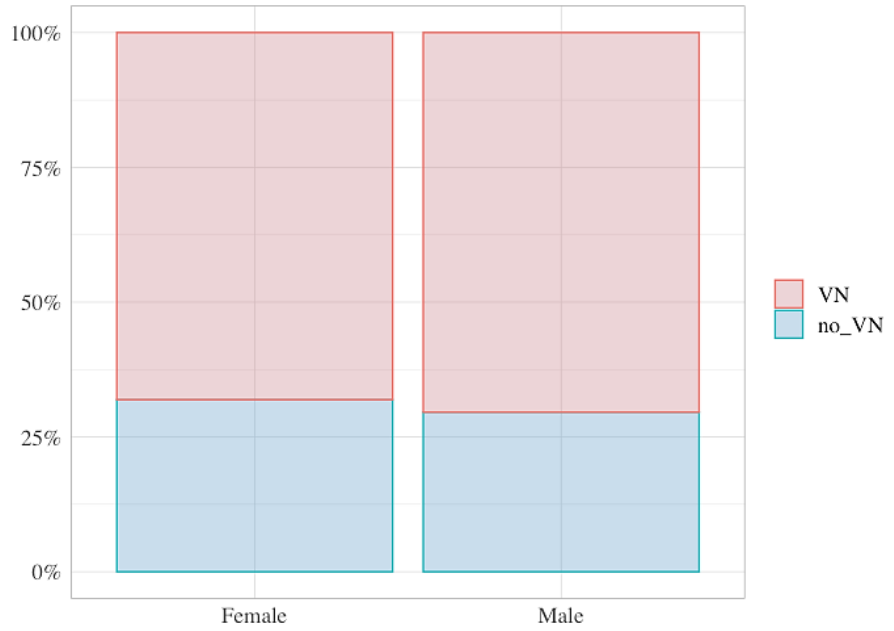


**Figure 3:** Gender – Test

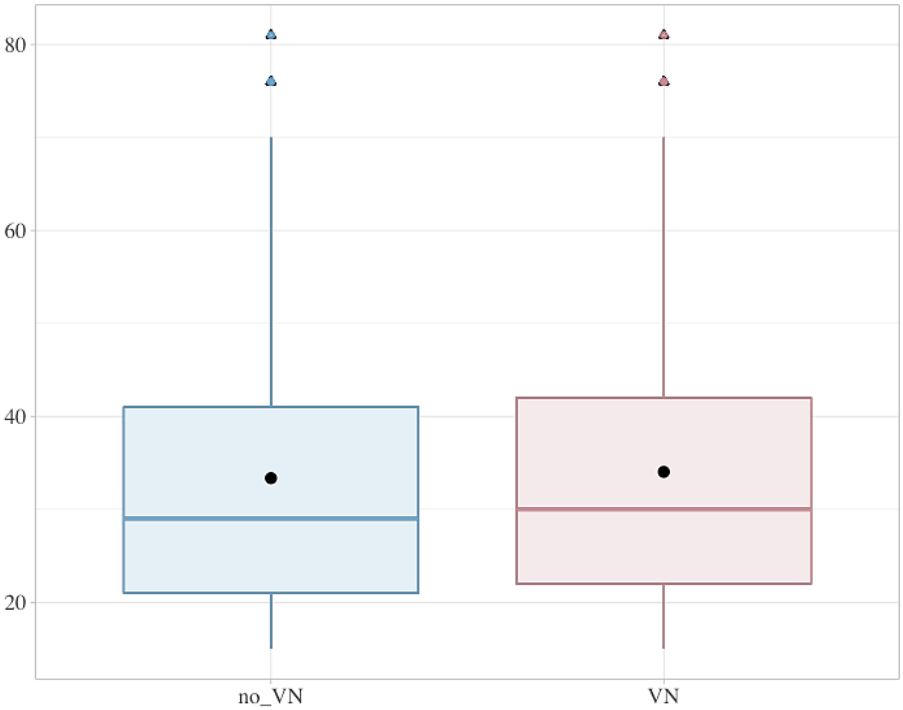


**Figure 4:** Age – Test

## 3.4 Predictor variables

In this article we propose to describe the behavior of the following variables, which can directly or indirectly explain the relationship between the bases under study and their respective forms derived by [n]Vsuffix: the stress of the base word, the number of syllables of the derived word, the vowel preceding the nasal consonant, the suffix involved in the derivation and lexical frequency.

### 3.4.1 Base stress

The derived forms under analysis present an almost perfect distribution in terms of stress: those closed

by -al are oxytone (e.g., admissional 'hiring$_{adj}$'), those closed by -ico are predominantly proparoxytone (e.g., canônico 'canonical') and those closed by -ismV are paroxytone (e.g., alpinismo 'mountaineering'). Although the two base patterns investigated accept derivation with the three suffixes, there is some restriction regarding stress, mainly concerning the emergence of proparoxytones with VN pattern (e.g., ômicron 'omicron'). For this reason, and to preserve uniformity between the CBras and Test samples — the latter which does not present proparoxytones among its base alternatives — this variable was analyzed from two subcategories, namely, *nonfinal stress* (e.g., cânone 'canon', estamina 'stamina') and *final stress* (e.g., lã 'wool', alemão 'German'). The question that arises with this variable concerns the role of stress in the emergence of the bases related to the derived forms under analysis.

**3.4.2 Number of syllables of the derived word**

One of the suffixes under consideration is monosyllabic, -al, and the other two are disyllabic, -icV and -ismV. The suffix -al, when attached to the base, either does not make the word longer (eg ca.no 'pipe' → ca.nal 'channel') or extends it by just one syllable (eg tom. 'tone' → to.nal 'tonal'). The suffixes -icV and -ismV can add one (ex. fo.ne 'phone' → fô.ni.co 'phonic', pleno. 'full'→ ple.nis.mo 'fulness') or two syllables to the base (ex. sa.xão 'Saxon' → sa.xô .ni.co 'saxonic'; ba.lão 'balloon' → ba.lo.nis.mo 'ballooning'). These last two suffixes can be resyllabified with the base, regardless of whether [n] is part of it (ex. í.co.[n]e 'icon' → i.cô.[n]+i.co 'iconic'; pa.gão 'pagan' → pa.ga.[n]is.mo 'paganism') or emerges by epenthesis (ex. cri.me 'crime' → cri.m+i.[n]al 'criminal') — although cases of preference for hiatus in this context are also attested in the language (ex. cipó 'liana' → cipoal / * cipo[n]al 'woods of lianas'). Taken together for the two samples, the derived forms for the three suffixes are categorized in this work as *up to 3 syllables* and *4 or more syllables*. With this variable, we intend to verify if the extension of [n]Vsuffix-words influences the selection of base patterns.

**3.4.3 Preceding vowel**

In all derived words considered in this study, in both samples, [n] is preceded by a vowel. In this context, the height of mid vowels is neutralized in Portuguese, and mid-low vowels rarely emerge. In the CBras sample, five different vowels were attested (eg albug[i]nico 'albuginic', lac[u]nal 'lacunar', c[e]nico 'scenic', adici[o]nal 'additional', americ[ɐ̃]nismo 'Americanism'). In the Test, as justified in 3.2, the three heights were considered, albeit restricted to back vowels (e.g., del[u]nico, jal[o]nismo, baji[a]nal). The variable preceding vowel was then categorized in the two samples into *high*, *mid*, and *low*. The question that arises is whether this vowel, which is often different from the final base vowel, plays a role in the selection of the base pattern.

**3.4.4 Suffix**

The suffix allomorphy hypothesis referred to earlier, according to which some V-initial suffixes could have a lexically listed variant already containing the alveolar nasal consonant [n] requires checking whether this structure would be specific to some suffixes. This is what is indirectly measured with this variable: whether forms in -al, -icV and -ismV behave differently with regard to the base patterns under study.

**3.4.5 Lexical frequency**

Some phonological phenomena can be explained by their frequency of types and tokens. However, when these phenomena interact with morphology, such an explanation is more complex. We can suggest that more frequent items demand less reflection from the speaker about their internal structure, but, on the other hand, we can also assume that very infrequent items may be the product of analogy rather than the application of morphophonological processes. For this reason, it is important to know whether the frequency of items derived by nVsuffix in our study can explain the emergence of more and less abstract

base patterns (in the sense of phonetic proximity between base and derived forms).

In the analysis of the lexicon in use, we consider the frequency rates provided by CBras, converted into a logarithmic scale. We used the median to divide this scale into two frequency levels: *low*, 0–2,833; *high*, 2,834–13,866. In the analysis of the potential lexicon, as it consists of pseudowords, we imported the frequency of CBras. The frequency rates of words closed by each of the strings Ca/Co/Cu/Va/Vo+nal/nico/nismo (the same that close the pseudowords) were summed and also converted into logarithms. The same procedure adopted for CBras resulted in the following scale: *low*, 4,718–10,613; *high*, 10,614–12,771.

### 3.4.6 Random variables

Variables understood as random, due to their characteristic peculiar to each sample, can have an effect on the predictive character of variables with fixed effects in different phenomena. This examination is particularly important when the research involves voluntary participants and pseudowords, as is the case with our experiment. The analysis of mixed effects of our Test thus includes *participant* and *pseudoword* as random variables.

The predictor variables analyzed in the two samples are summarized in Table 5.

**Table 5:** Predictor variables – CBras and Test

| Variables | | Examples | | |
|---|---|---|---|---|
| | | CBras | | Test |
| Base stress | nonfinal | ti**ra**no | 'tyrant' | ja**lô**ne |
| | final | conven**ção** | 'convention' | ja**lão** |
| Number of syllables | ≤ 3 | to.nal | 'tonal' | fa.gu.nal |
| | ≥ 4 | ca.nô.ni.co | 'canonical' | de.lú.ni.co |
| Preceding vowel | high | tup**i**nismo | 'Tupinism' | chod**u**nismo |
| | mid | ecum**ê**nico | 'ecumenical' | zad**ô**nico |
| | low | cr**â**nico | 'cranial' | baji**a**nal |
| Suffix | -al | artesan**al** | 'artisanal' | began**al** |
| | -icV | platôn**ico** | 'Platonic' | ludân**ica** |
| | -ismV | catalan**ismo** | 'catalanism' | gojan**ismo** |
| Lexical frequency | low | angolanismo$_{log0}$ | 'angolanism' | Vanal$_{log4.72}$ |
| | high | nacionalismo$_{log13.87}$ | 'nationalism' | Conico$_{log12.77}$ |
| Random variables (only Test) | participant | | | |
| | pseudoword | | | |

Statistical analysis was conducted in R (28), with the glm and glmer functions to perform binary logistic regression models with fixed and mixed effects, respectively. The figures were produced with ggplot2 package (29).

## 4 Results and discussion

In this section, the logistic regression results applied to the two samples are presented and discussed, considering the predictive potential of the analyzed variables. First, the results obtained for the CBras data are presented and, subsequently, for the Test data. Finally, a comparison between the results obtained for the two samples is proposed, focusing on the assumptions of AA and CA, as well as the limits of what is defined as lexicon in use and potential lexicon.

The models presented were selected after several statistical rounds, including interactions between predictor variables considered linguistically coherent. Once the variables and interactions were defined, the models chosen were those that presented the lowest AIC and highest R2Tjur index. The confidence interval used in the analysis is 95%.

### 4.1 CBras

Table 6 presents the results of the fixed effects logistic regression test applied to the CBras data.

**Table 6:** Predictors for VN – CBras

| Predictors | | Estimate | Std. Error | z-value | *p*-value | N | Total | % |
|---|---|---|---|---|---|---|---|---|
| Intercept | | -3.89 | 0.60 | -6.53 | **<0.001***** | | | |
| Base stress | nonfinal | (reference) | | | | 76 | 992 | 7.66 |
| | final | 6.95 | 0.49 | 14.31 | **<0.001***** | 577 | 582 | 99.14 |
| Number of syllables | ≥ 4 | (reference) | | | | 618 | 1391 | 44.43 |
| | ≤ 3 | -0.68 | 0.42 | -1.60 | 0.109 | 35 | 183 | 19.13 |
| Preceding vowel | low | (reference) | | | | 50 | 335 | 14.93 |
| | mid | 1.63 | 0.38 | 4.34 | **<0.001***** | 553 | 909 | 60.84 |
| | high | 1.00 | 0.41 | 2.43 | **0.015*** | 50 | 330 | 15.15 |
| Suffix | -al | (reference) | | | | 305 | 478 | 63.81 |
| | -icV | -0.07 | 0.56 | -0.13 | 0.895 | 104 | 544 | 19.12 |
| | -ismV | 0.57 | 0.54 | 1.07 | 0.284 | 244 | 552 | 44.20 |
| Lexical frequency | low | (reference) | | | | 332 | 775 | 42.83 |
| | high | 0.91 | 0.55 | 1.65 | 0.098 | 321 | 799 | 40.20 |
| **Interactions** | | | | | | | | |
| Suffix & Lexical freq | -icV&high | -0.53 | 0.66 | -0.81 | 0.420 | 75 | 358 | 20.95 |
| | -ismV&high | -0.98 | 0.70 | -1.40 | 0.161 | 72 | 175 | 41.14 |
| Observations | 1574 | | | | | | | |
| $R^2$ Tjur | 0.808 | | | | | | | |
| AIC | 583.481 | | | | | | | |

model_CBras = base.pattern ~ base.stress + number.syllables + prec.vowel + suffix * lex.freq

According to the logistic regression test presented in Table 6, final-stress bases and derived forms with medium and high vowels preceding [n]Vsuffix favor the VN pattern. The variables *number of syllables of the derived word, suffix, lexical frequency* and the interaction between *suffix* and *lexical frequency* were not significant in this model.

### 4.2 Test

Table 7 below presents the results of the mixed effects logistic regression model for the Test data.

**Table 7:** Predictors for VN – Test

| Predictors | | Estimate | Std. Error | z-value | *p*-value | N | Total | % |
|---|---|---|---|---|---|---|---|---|
| Intercept | | -1.87 | 0.34 | -5.51 | **<0.001***** | | | |
| Base stress | nonfinal | (reference) | | | | 51 | 709 | 7.19 |
| | final | 5.28 | 0.21 | 25.04 | **<0.001***** | 2088 | 2396 | 87.15 |
| Number of syllables | ≥ 4 | (reference) | | | | 1574 | 2277 | 69.13 |
| | ≤ 3 | 0.36 | 0.28 | 1.31 | 0.190 | 565 | 828 | 68.24 |
| Preceding vowel | low | (reference) | | | | 968 | 1242 | 77.94 |
| | mid | -1.11 | 0.22 | -5.15 | **<0.001***** | 862 | 1242 | 69.40 |
| | high | -1.36 | 0.29 | -4.71 | **<0.001***** | 309 | 621 | 49.76 |
| Suffix | -al | (reference) | | | | 689 | 1035 | 66.57 |
| | -icV | -0.35 | 0.41 | -0.85 | 0.396 | 715 | 1035 | 69.08 |
| | -ismV | -0.64 | 0.36 | -1.76 | 0.078 | 735 | 1035 | 71.01 |
| Lexical frequency | low | (reference) | | | | 995 | 1449 | 68.67 |
| | high | -1.11 | 0.37 | -3.02 | **0.003**** | 1144 | 1656 | 69.08 |
| **Interactions** | | | | | | | | |
| Suffix & Lexical freq | -al&low | (reference) | | | | 270 | 414 | 65.22 |
| | -icV&high | 1.56 | 0.53 | 2.93 | **0.003**** | 405 | 621 | 65.22 |
| | -ismV&high | 1.22 | 0.48 | 2.56 | **0.011*** | 320 | 414 | 77.30 |
| **Random Effects** | | | | | | | | |
| $\sigma^2$ | 3.29 | | | | | | | |
| $\tau_{00}$ | 0.85 $_{participant}$ | | | | | | | |
| | 0.06 $_{pseudoword}$ | | | | | | | |
| ICC | 0.22 | | | | | | | |
| N | 207 $_{participant}$ | | | | | | | |
| N | 15 $_{pseudoword}$ | | | | | | | |
| Observations | 3105 | | | | | | | |
| $R^2$ Tjur | 0.564/0.659 | | | | | | | |
| AIC | 2050.358 | | | | | | | |

model_Test = base.pattern ~ base.stress + number.syllables + prec.vowel + suffix * lex.freq + (1|participant) + (1|pseudoword)

Final-stress bases and interactions between forms derived by the suffix -icV or -ismV and high frequent lexical strings are pointed out as favoring the pattern VN in the logistic regression model shown in Table 7. On the other hand, forms with mid or high vowels preceding [n]Vsuffix as well as high frequent lexical strings (in isolation) disfavor this base pattern. The variables *number of syllables of the derived word* and *suffix* are not significant in the Test data.

### 4.3 Discussion

In this section, we comparatively discuss the results of the logistic regression models for the CBras and the Test data, focusing on the properties of the lexicon in use and the potential lexicon and on the contrast between AA and CA.

Figure 5 presents the combined results of the two samples.

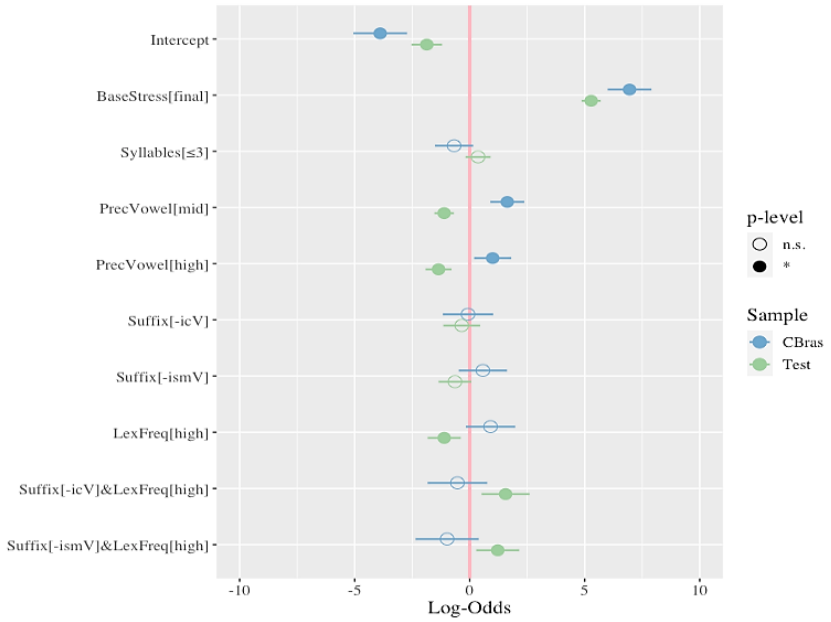


**Figure 5:** Predictors for VN – CBras and Test

Most of the investigated variables behave similarly in the two samples. The negative intercept, significant in both samples, is due to the low occurrence of VN when the predictor variables are not considered. Comparatively, it can be said that AA is generally less productive in the lexicon in use than in the potential lexicon.

As for base stress, the vast majority of VN pattern forms in both samples are stressed on the last (sometimes the only) syllable (e.g., ladrão 'thief', tom 'tone'; dobão, delúm), with few exceptions (e.g., órgão 'organ', pólen 'pollen'; délum). In contrast, the intercept pattern, no_VN, presents non-final stress preferentially (e.g., arquiteto 'architect', tirano 'tyrant', ícone 'icon'; dobône, delúnes). Even though words unknown to the speakers are at stake, the data from the potential lexicon, with slightly lower rates, confirm the almost complementary distribution observed for this variable in the lexicon in use.

Although stress is the condition par excellence for the selection of the bases under analysis, the inclusion of other variables represented an increase in the coefficient of determination of the models, indicated by the R2Tjur index, mainly in the potential lexicon.

Regarding the number of syllables, a non-significant variable in both samples, it is worth noting that the lexicon in use shows a considerably higher number of long words (with 4 or more syllables) than short words (up to 3 syllables). In the set of longer words, a relatively balanced distribution of these two patterns is observed (e.g., in.ten.cio.nal 'intentional' < intenção$_{VN}$ 'intention'; ca.nô.ni.co 'canonical' < cânone$_{no_VN}$ 'canon'). In the narrower set of short words, there is a distributional preference for the pattern no_VN (e.g., ca.nal 'channel' < ca.no 'pipe'). In the potential lexicon, a balanced distribution of long and short derived words in the two patterns is attested, with some advantage for VN (e.g., cho.du.nis.mo < chodún ~ chodúne; be.ga.nal < begão ~ begâne), the most prevalent pattern overall.

As for the height of the vowel preceding the nasal consonant, which is a significant variable in both models, there is an apparent contradiction in the results, with mid and high vowels presenting positive and negative logodds for lexicon in use and potential lexicon models, respectively. However, if we add the coefficients obtained for these vowels to the negative intercepts of each model, we will reach negative coefficients for these variables in both samples. This suggests parsimony in interpreting these results in a stricter sense. On closer examination, it is noted that the advantage in the coefficient obtained for mid vowels in the lexicon in use is driven by [o] (e.g., pulmônico 'pulmonic'), more than by [e] (e.g., cênico 'scenic'), which is consistent with the hypothesis that the speaker pairs such forms with bases

closed by [õj̃s], the most productive plural marker of words closed by nasal diphthongs (e.g., pulmões 'lungs'). Also in the lexicon in use, the coefficient obtained for high vowels (e.g., final 'final', comunismo 'communism') is significant as it contrasts with the result of mid vowels, although it is close to the result of the low vowel, taken as the reference variable (e.g., catalanism 'Catalanism'). In the potential lexicon, the result is inverted, not because of the distribution of mid and high vowels per se, but because of the contrast with the relevance that the low vowel assumes in this sample, which is distinct from the lexicon in use. Based on the same argument that relates [o] to the most productive plural form, we may suggest that the preference for VN in the case of [a] is explained by its relationship with [ãw̃], which is the most productive singular form in this case. It is important to emphasize that no peculiar lexical behavior that could explain these results was observed in the context of this variable.

The suffix involved in the derivation was not significant in isolation in the logistic regression models for any of the two samples. In terms of distribution, in the lexicon in use, there is a prevalence of -al in the VN pattern, while -icV and -ismV predominate in the no_VN pattern. In the potential lexicon, the three suffixes are more recurrent in the VN pattern, with very approximate rates. There is, therefore, nothing to say in terms of the morphological role of the affix itself in the prediction of these bases.

Lexical frequency was not considered significant in the analysis of the lexicon in use. In terms of distribution, there is a great balance between low and high frequency items in both the VN and no_VN patterns, the latter generally concentrating most of the examples. In Figure 6 below we present, by way of illustration, the 100 most frequent words in the sample.


**Figure 6:** 100 most frequent words – CBras

In opposition, in the potential lexicon, high frequency items disfavor VN bases. The general frequency distribution is not able to explain this result, which seems to be driven by the atypical behavior of suffixes when related to certain lexical strings.

The analysis of the interaction between the variables *suffix* and *lexical frequency*, with no role in the lexicon in use, is relevant in the analysis of the potential lexicon, revealing a favoring of VN bases in the context of -icV and -ismV and high frequency strings. This is possibly due to the predominance of the string *unico* and the lack of strings closed by -al (the suffix taken as a reference variable) among the high frequency items. Figure 7 below shows the distribution of strings analyzed in the Test in relation

to the response variable in ascending order, from left to right, according to the lexical frequency imported from CBras.

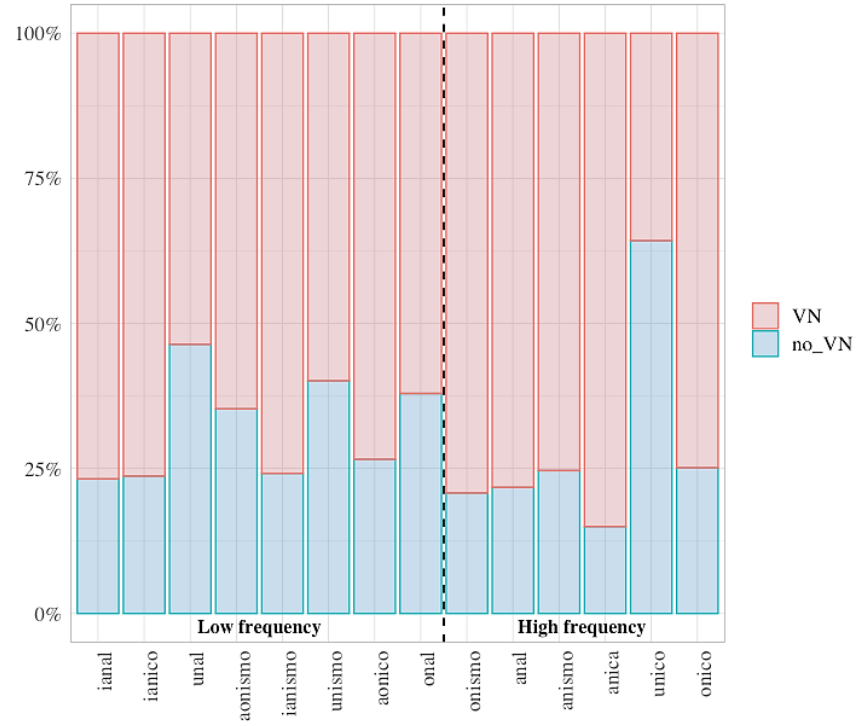


**Figure 7:** Lexical strings (from least to most frequent in CBras) – Test

Regarding the role of random variables in the analysis of the potential lexicon, it is observed that their inclusion represents an increase of about 10% in the coefficient of determination of the model, indicated by the R2Tjur index, when compared to the fixed effects model. The main contribution comes from the variable *participant* and, secondarily, from the variable *pseudoword*, as shown in Figure 8 below.

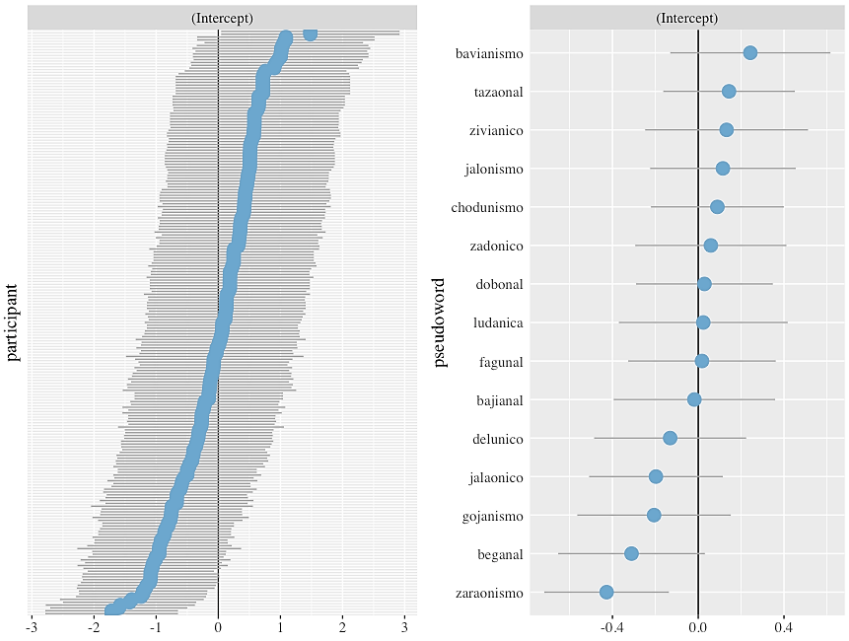


**Figure 8:** – Random variables – Test

## 5 Final remarks

In this study we descriptively examine the productivity of bases that relate to words ended in [n] followed by the vowel-initial suffixes -al, -icV and -ismV (e.g, adicional 'additional'; longitudinal 'longitudinal'; tônico 'tonic'; arquitetônico 'architectural'; conexionismo 'connectionism'; tupinismo 'Tupinism'). Data representing the lexicon in use, from Corpus Brasileiro, and the potential lexicon, from a pseudoword test, are considered. The objective is to analyze variables capable of predicting the selection of two opposed patterns of bases, VN (e.g., adição 'addition'; tom 'tone'; conexão 'connection') and no_VN (e.g., longitude 'longitude'; arquiteto 'architect'; tupi 'Tupi'). VN reflects an abstract approach (AA) to word-final nasality in BP, whereas no_VN is representative of a concrete approach (CA) to the phenomenon.

In terms of distribution, the first point to highlight is the preference for no_VN in the data from the lexicon in use as opposed to the preference for VN in the data from the potential lexicon. The motivation may be historical, if we take into account that forms closed by nasal diphthongs have their origin in forms closed by VnV, as is still the case in other Romance languages (e.g., irmão, Portuguese ~ hermano, Spanish < ermano < germano, Latin 'brother'). This would explain the preservation of forms with no_VN pattern in items already consolidated in the lexicon, in contrast to the productivity of AA in new items.

Final stress is a clear predictor of VN pattern bases in both samples. The logistic regression test, however, points to the predictive role of variables other than stress. In the lexicon in use, mid and high vowels preceding [n]suffix favor VN. On the other hand, in the potential lexicon, these same vowels disfavor VN. Nevertheless, this apparent contradiction is relativized when the estimate and intercept values of the two samples are contrasted. In the potential lexicon, the high frequency of certain lexical strings present in pseudowords generally disfavors the selection of the VN pattern. This effect, however, is resized when the interaction between this variable and the variable *suffix* is analyzed. In this case -icV and -ismV in the context of high frequency strings favor VN, due to the predominance in this context of the string *unico* and the lack of strings ending in -al. Finally, concerning the Test analysis, when the random variables *participant* and *pseudoword* are added to the logistic regression model, an increase of around 10% in the coefficient of determination index is observed.

This study contributes to the hypothesis that forms derived by [n]Vsuffix are to some extent related to structures containing a nasal structure (segment or equivalent) in final position. The fact that this correspondence is more prevalent in the potential lexicon may be indicative of the explanatory plausibility of AA. However, this does not rule out the productivity of CA. The same phonological constraints that condition the selection of abstract bases in the lexicon in use operate to some extent in the potential lexicon. In addition to phonological constraints, the high frequency of certain lexical chains related to certain suffixes can be another conditioning factor for this selection.

In this analysis, we chose to verify specifically the contrast between more and less abstract bases and their correspondence with morphologically derived forms in two types of samples of what we define as internalized lexicon types. Other categorizations of the response variable may, however, contribute to broadening the understanding of how these representations are learned by BP speakers.[4]

[4] As an alternative approach to the response variable, we can assume, for example, that bases of pattern VN and no_VN are phonologically similar at the root level (e.g., $caN_{root}+o_{theme}$ > c[ãw̃] 'dog' / ca[n]ino 'canine'; $can_{root}+o_{theme}$ > can[ʊ] 'pipe' / canal 'channel'). This hypothesis is defended by Schwindt (25) for words closed by /l/.

## Acknowledgments

This article is part of the production of the first author under the supervision of the second one, regarding his collaboration as a visiting scholar at Institute for Language Studies/Campinas State University, from August 2019 to July 2020. We gratefully thank the National Council for Scientific and Technological Development (CNPq) for the support to this research (grants PQ-310921/2018-0 and PQ-312620/2020-9). We also thank the students Isabela Petry, Nathan Barcellos, and Pedro Gaggiola, for contributing in different stages to this research, and the anonymous reviewers for their comments and useful advice to improve this work. All remaining errors are our own responsibility.

### REFERENCES

1. Câmara Jr. JM. *Para o estudo da fonêmica portuguesa*. Rio de Janeiro: Organização Simões; 1953.

2. Câmara Jr. JM. *Problemas de linguística descritiva*. Petrópolis: Editora Vozes; 1969.

3. Câmara Jr. JM. *Estrutura da língua portuguesa*. 35. ed. Rio de Janeiro: Editora Vozes; 1970.

4. Lemle M. *Phonemic system of the Portuguese of Rio de Janeiro*. [Dissertation – Masters]. University of Pennsylvania; 1965.

5. Leite Y. *Portuguese stress and related rules*. [PhD]. University of Texas; 1974.

6. Cagliari LC. *An experimental study of nasality with particular reference to Brazilian Portuguese*. [PhD]. University of Edinburgh, Edimburgo; 1977.

7. Abaurre-Gnerre MBM. *Alguns casos de formação de plural em português: uma abordagem natural*. Cad. Est. Ling., Campinas. 1983;5:127-156.

8. Wetzels L. *Contrastive and allophonic properties of Brazilian Portuguese vowels*. In: Wanner D, Kibbee DA, editors. New analyses in Romance linguistics. Amsterdam: J. Benjamins, 1991.

9. Wetzels L. *The lexical representation of nasality in Brazilian Portuguese*. Probus. 1997; 9:203-232. http://dx.doi.org/10.1515/prbs.1997.9.2.203.

10. Wetzels L. *Comentários sobre a estrutura fonológica dos ditongos nasais no Português do Brasil*. Revista de Letras, Fortaleza. 2000;1(22):25-30.

11. Bisol L. *A nasalidade, um velho tema*. DELTA, São Paulo. 1998;14(nº especial):27-46. http://dx.doi.org/10.1590/S0102-44501998000300004.

12. Bisol L. *A nasalidade fonológica no português e suas restrições*. Diadorim, Rio de Janeiro. 2016;18:116-126. http://dx.doi.org/10.35520/diadorim.2016.v18n0a4050.

13. Huback AP. *Plurais irregulares do português brasileiro: efeitos de frequência*. Revista da Abralin, Curitiba. 2010a;9(1):11-40. http://dx.doi.org/10.5380/rabl.v9i1.52337.

14. Huback AP. *Plurais em -ão do português brasileiro: efeitos de frequência*. Revista Linguística, Rio de Janeiro. 2010b;6(1):9-26. http://dx.doi.org/10.31513/linguistica.2010.v6n1a4436.

15. Cristófaro-Silva T. *Organização fonológica de marcas de plural no português brasileiro: uma abordagem multirrepresentacional*. Revista da Abralin, Curitiba. 2012;11:273-305. http://dx.doi.org/10.5380/rabl.v11i1.32468.

16. Guimarães M, Nevins A. *Probing the representation of nasal vowel in Brazilian Portuguese with language games*. ORGANON, Porto Alegre. 2013;28(54):155-178. http://dx.doi.org/10.22456/2238-8915.38298.

17. Becker M, Nevins A, Sandalo F, Rizzato É. *The acquisition path of [w]-final plurals in Brazilian Portuguese*. J. Port Linguist, Lisboa. 2018;17(4):1-17. http://dx.doi.org/10.5334/jpl.189.

18. Rizzato É. *Interação do plural de -ão e do aumentativo -zão na formação de compostos no português brasileiro*. [Dissertation – Masters]. Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas; 2018.

19. Gomes CA, Prado LO, Amaral TLA. *Aspectos cognitivos e sociais da variação linguística na alternância de formas de plural de nomes do PB*. In: Marins J, Orsini M, Cavalcante SR, editors. Contribuições à descrição e ao ensino do português brasileiro: da fonética ao discurso, com parada obrigatória na sintaxe - uma homenagem a Maria Eugênia Lammoglia Duarte. Rio de Janeiro: EDUFRJ, 2021.

20. Schwindt LC, Gaggiola PE, Petry IP. *Frequência e distribuição de plurais irregulares no Corpus Brasileiro*. Rev. Estud. Ling. 2021;29(2):1289-1324. http://dx.doi.org/10.17851/2237-2083.29.2.1289-1324.

21. Head BA. *Comparison of the segmental phonology of Lisbon and Rio de Janeiro*. [PhD]. University of Texas at Austin; 1965.

22. Matta Machado MT. *Étude articulatoire et acoustique des voyelles nasales du portugais de Rio de Janeiro: analyses radiocinematographique, sonagraphique et oscillographique*. [PhD]. Université de Sciences Humaines de Strasbourg; 1981

23. Vennemann T. *Rule inversion*. Lingua. 1972;29:209-242.

24. Possenti S. *Via-rules: um problema metodológico para a fonologia gerativa*. In: Atas do III Encontro Nacional de Linguística. Rio de Janeiro, Departamento de Letras/PUC. 1979.

25. Schwindt LC. *Underlying representation of [w]-final words in Brazilian Portuguese: evidence from morphological derivation*. Acta Linguistica Academica. 2021;68(1-2):139-157. http://dx.doi.org/10.1556/2062.2021.00482.

26. Cagliari LC, Massini-Cagliari G. *A epêntese consonantal em português e sua interpretação na Teoria da Otimalidade*. RELIN. 2000;9(1):163–192. http://dx.doi.org/10.17851/2237-2083.9.1.109-162.

27. Canfield SS. *Breve descrição da epêntese consonantal em palavras derivadas por sufixação no português brasileiro*. Cadernos do IL. 2018;56:57–69. http://dx.doi.org/10.22456/2236-6385.83495.

28. R Development Core Team. *R: A Language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. 2021. Available from: http://www.R-project.org.

29. Wickham H. *Ggplot2: Elegant graphics for data analysis*. New York: Springer; 2009.