# Eu nunca falei nada pra ninguém não, né! A variationist corpus study of negative concord in Brazilian Portuguese
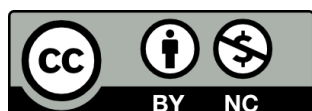
J.A. (Jac) Brown

# Eu nunca falei nada pra ninguém não, né!

# A variationist corpus study of negative concord in Brazilian Portuguese

*J.A. (Jac) Brown*
*Master Linguistics, specialization General Linguistics*
*Radboud University, Nijmegen*

**Abstract**: Portuguese is a negative concord (NC) language, specifically a non-strict NC language (i.e. a language that requires NC only in certain constructions). However, colloquial Brazilian Portuguese (BP) alternatively allows the "negative quantification" construction (as described by Agostini & Schwenter, 2015; in this paper, the term "lack of NC" is used). The present study compares three corpora of written and spoken BP in an attempt to uncover the reasons behind variation between NC and lack of NC. The findings by Agostini & Schwenter are largely replicated. Although lack of NC is generally a colloquial phenomenon, there are multiple factors at work, both language-internal (the various NC items have their own syntactic properties) and language-external (properties of the speakers, formality, type of conversation). It is found that the two written language corpora, despite being relatively informal, differ significantly from the spoken language corpus, as well as from each other.

**Keywords**: Brazilian Portuguese, negative concord, negative indefinite, variationist sociolinguistics, corpus linguistics

## 1.     Introduction

### 1.1 Theoretical background

In Brazilian Portuguese, a sentence with a postverbal negative indefinite (hereafter: NI) can be constructed in one of two ways. The first, common to all varieties of Portuguese, makes use of negative concord, that is to say, a negative licensor *não* 'not' or *nunca* 'never' is inserted preverbally and an NI, which may also be *nunca*, but also *ninguém* 'no one, nobody', *nenhum(a)* 'no, none [an adjective in Portuguese]' or *nada* 'nothing' is inserted postverbally, giving a single negative meaning to the sentence as a whole, as in the following examples

| (1) | O João | nunca | veja | ninguém | na | rua |
|-----|--------|-------|------|---------|-----|-----|
|     | John   | never | see.3SG.PRES | nobody | on.the | street |
|     | "John never sees anybody on the street" | | | | | |

| (2) | A Maria | não | beijou | nenhum | homem | ontem |
|-----|---------|-----|--------|--------|-------|-------|
|     | Mary    | not | kiss.3SG.PF | no | man | yesterday |
|     | "Mary hasn't kissed any man yesterday" | | | | | |

(3) | Nós | nunca | vandalizaremos | | nada
--- | --- | --- | --- | --- | ---
| 1PL | never | vandalize.1PL.FUT | | nothing

"We will never vandalize anything"

This is the default construction for all the above expressions. However, *nunca* as a postverbal NI behaves differently from the other three, as the following construction is possible but marked

(4) | O Flamengo | não | caiu | nunca | na | Segunda
--- | --- | --- | --- | --- | --- | ---
| Flamengo | not | fall.3SG.PF | never | in.the | Second

"Flamengo (a football team) haven't ever been relegated to the second level' "

The unmarked construction in this case would be *O Flamengo nunca caiu na Segunda* 'Flamengo have never been relegated to the second level', i.e. a construction without negative concord (hereafter: NC). The reason this is so is beyond the subject of this paper, but one can imagine it is the same or similar to the reason why *not ever* is marked in English compared to *never*.

However, another construction is possible, where the preverbal negative licensor (and thus the NC) disappears. This construction, although still somewhat marked, is common in colloquial (but not formal) Brazilian Portuguese (hereafter: BP). Consider the following examples:

(1a) | O João | veja | ninguém | na | rua
--- | --- | --- | --- | --- | ---
| John | see.3SG.PRES | nobody | on.the | street

"John doesn't see anybody on the street"

(2a) | A Maria | beijou | nenhum | homem | ontem
--- | --- | --- | --- | --- | ---
| Mary | kiss.3SG.PF | no | man | yesterday

"Mary didn't kiss any man yesterday"

(3a) | Nós | vandalizaremos | nada
--- | --- | --- | ---
| 1PL | vandalize.1PL.FUT | nothing

"We won't vandalize anything"

(The meaning of 1a and 3a has changed with respect to 1 and 3, but this is irrelevant here because it is due to the semantics of *nunca*. 1a and 3a would be equivalent in meaning to 1 and 3 if 1 and 3 had *não* in the place of nunca.)

Agostini & Schwenter (2015) investigated the variability between constructions with and without NC in BP (what they call "variable NC"), by means of an online survey into the acceptability of lack of NC (pp. 5–7). Although they state that the corpus data they had access to was not suitable for researching variable NC in BP (p. 5), I thought it worthwhile to investigate whether the cor-

pora I found (see below) could be used to replicate their findings and answer the question "Which factors determine the presence or absence of NC in BP?".

## 1.2 Hypotheses

Agostini & Schwenter draw several conclusions about the factors influencing the acceptability and frequency of lack of NC in BP. Strikingly, they found that acceptability ratings for the various NIs differ in a way that is consistent with their frequency: *nada*, the most common NI, is also judged to be most acceptable, whereas *nunca*, the least common one, is judged to be least acceptable. The acceptability ratings of *ninguém* and *nenhum(a)*, which lie in between, are the reverse of their frequencies, but the difference is so small that this could easily be an artifact of the samples used (frequency data came from the *Corpus do Português*; cf. pp. 5–9, in particular the frequency graph on p. 7 and the acceptability graphs on pp. 8 and 9). They also found a correlation between acceptability ratings and the token-to-type ratio of the verbs with which the four different NIs occur: *nada* occurs with the highest token-to-type ratio of verbs, *nunca* with the lowest, and the other two are in the middle (pp. 15–16). However, because the overall frequencies and the token-to-type ratios are so similar, it is difficult to say which one of them is more relevant for the acceptability of the four NIs.

Furthermore, they found that the respondents on their survey did not give significantly different acceptability judgments depending on their age, leading them to suppose that variable NC in BP is a case of stable variation, rather than a sign that BP is in the process of moving from NC toward lack of NC diachronically (pp. 12–13); nor did the respondents give significantly different acceptability judgments depending on their gender or education level (p. 13).

Native speakers commented that the construction without NC was more "emphatic" than the construction with NC (p. 13). This is in accordance with my own, anecdotal, observations from conversations in colloquial BP. Finally, their respondents were divided into three geographical areas (see Fig. 1): Rio Grande do Sul (the southernmost state of Brazil), São Paulo (in the southeast, more northernly but still relatively close to Rio Grande do Sul), and the Nordeste (northeast) region; the respondents from Rio Grande do Sul rated lack of NC as less acceptable overall (p. 14).



**Fig. 1.** *Regiões e estados do Brasil* [Regions and states of Brazil]. Note that São Paulo state is in the Sudeste region (red) and Rio Grande do Sul in the Sul region (purple). (Source: Polon, n.d.)

These regions have different demographics. For example, inhabitants of Rio Grande do Sul are called *gaúchos* and many of them own cattle ranches; they are stereotyped as being in majority white, rich and politically conservative. On the other hand, the Nordeste region is known for its large number of inhabitants of African descent and is stereotyped as being poor and left-wing (cf. stereotype maps of Brazil at https://www.reddit.com/r/brasil/comments/9xvs76/12_jeitos_de_dividir_o_brasil/). Thus, it stands to reason that these demographic differences contribute to the differing acceptability rates for lack of NC.

Based on both Agostini & Schwenter's findings and my own anecdotal observations, then, we can hypothesize a number of outcomes that will be tested in the present study, as follows:

Hypothesis 1: We expect to replicate Agostini & Schwenter's findings for the frequency of the various NIs in postverbal position. It will be interesting to see here if we can find a significant difference between *ninguém* and *nenhum(a)*.
 Hypothesis 2: Because lack of NC is highly colloquial, we expect to find it in higher frequencies the more informal our corpus is (see below for more data on the corpora used). In more formal language, we should expect more NC, whereas especially in spoken language, we should expect a large amount of variation, where lack of NC is used in particular for the purposes of emphasis.
Hypothesis 3: We expect any regional variation found in our corpora to, broadly speaking, match that found by Agostini & Schwenter.
At this point, then, it will be useful to divide our research question into two sub-questions, namely "Which NIs occur most often in non-NC constructions?" and "In which corpora do non-NC constructions most often occur?", in order to distinguish between the language-internal (frequency, syntax) and language-external (demographic) factors under investigation.

## 2.    Method

The AC/DC project (*Acesso a Corpos/Disponibilização de Corpos* "Access to Corpora/Making Corpora Available", https://www.linguateca.pt/ACDC/) is a collection of 35 corpora of Portuguese, both spoken and written, in various genres, in both European and Brazilian Portuguese, as well as other varieties (such as Mozambican), totalling over 1.5 billion units, annotated using the PALAVRAS parsing system (Bick, 2000). PALAVRAS allows for a variety of tagging methods, including part-of-speech (PoS) tagging and syntactic function tagging. These corpora can be queried using the IMS Open Corpus Workbench (CWB; Evert, 2016).

For the present study, three corpora of a varied nature (see table 1) were searched for instances of the four postverbal NIs *nunca*, *ninguém*, *nenhum(a)* and *nada* using the CWB query syntax (see appendix). The output for each of the three corpora was then searched again for the preverbal negative licensors (hereafter: NLs) *não* and *nunca*. This yielded two sets of data for each corpus: one of NC sentences and one of non-NC sentences. In addition, I added another query for each corpus to include a preposition between the verb and the NI, since the NI may occur in combination with a preposition such as *p(a)ra nada* 'for nothing', *com ninguém* 'with nobody, without anybody' or *em nenhum(a)* 'in any [in negative polarity sentences in English]'. The output from these ad-

ditional queries was once again divided into sets of NC and non-NC sentences. Unintelligible, uninterpretable or doubtful results were manually excluded.

One weakness of this method is that it ignores punctuation and prosody, which means some results that looked like cases of verb + NI at first sight actually consisted of two sentences[1]. These were also manually excluded. Results were tabulated for each corpus and for each postverbal NI (see below) and analyzed for statistical significance.

**Table 1**. Corpora used, with their sizes.

| Corpus name | Type of data | Size (units) |
|---|---|---|
| Corpus Brasileiro (genre subcworpus: en) | texts from film and television scripts | 254,352 |
| Corpus Brasileiro (genre subcorpus: ed) | texts from magazines | 494,263 |
| C-Oral-Brasil | informal, spontaneous conversations (Minas Gerais) | 431,081 |
| **Total** | | **1,179,696** |

Lack of NC being a colloquial phenomenon, the best data to study it is found in spoken language. Nonetheless, I justify the choice of corpora here on the grounds that film and television scripts consist mostly of dialogue and thus, while not themselves spoken language, are a kind of representation of spoken language, often intended to sound colloquial; furthermore, that the language used in magazines, while written, tends to be somewhat more popular and informal than other written language. We can thus use the different corpora as a proxy for the formality of the language, where the spontaneous conversations are the least formal and the texts from magazines are the most formal.

Geographically, Minas Gerais is in the same region as São Paulo (see fig. 1), and we therefore expect the findings for this region to be similar to Agostini & Schwenter's findings for São Paulo. Unfortunately, the data from C-Oral-Brazil does not include information about the speakers, meaning that other demographic variables (race, gender, age, education level) cannot be investigated.

I did not look at the variation between the three different kinds of negation NEG1, NEG2 and NEG3 (as described by Agostini & Schwenter 2015, p. 14), only at NC versus strictly postverbal negation. This means I ignored sentences of the form *não* V *não* and *nunca* V *não*. I made this choice because these constructions represent a different kind of negation[2], whereas the present study is

---

1 Example: Mas o meu amigo me falou nunca falou que tem que desligar o pen drive não (C-Oral-Brasil), which the CWB counts as a case of V+NI, is actually Mas o meu amigo me falou – nunca falou que tem que desligar o pen drive não 'But my friend told me w– he definitely never said you had to disconnect the flash drive'.

2 The second, postverbal não in sentences like Eu não gosto disso não 'I don't like this (no/at all)' is not an NC item, but rather a separate negator added to strengthen the negation. This can be seen from the fact that it can only come at the end of the sentence, whereas postverbal NIs in an NC construction can only occur as part of the VP (*Eu não gosto não disso is ungrammatical, cf. NC Eu não gostei nunca disso which is grammatical); as well as from the fact that there is an equivalent affirmative construction with sim, litt. 'yes', as in Eu gosto disso sim 'I do like this (definitely)'.

focused only on NC. However, I did occasionally find cases where post-verbal *não* is preceded by an NI (V *nada não* etc.), which, although they are cases of NEG2 or NEG3, I did include in the analysis because they still represent variation between NC and lack of NC. In other words, both NEG2 and NEG3 can occur either with or without a postverbal NI; only the cases where they occur with a postverbal NI are relevant to the present study. I have counted these separately.

I also chose not to look at other NC items such as *nem* 'not even, nor, neither' and *sem* 'without', because their semantics preclude them from being used as preverbal NLs in non-NC constructions, thus they only occur in this way in NC constructions. They can also occur postverbally, but this was not found in the current study. A small number of sentences with preverbal *nunca mais* 'never again, never … anymore' were found, as well as one sentence with highly formal *jamais* 'never'; I decided to include them because *nunca mais* and *jamais* behave the same way as *nunca*.

## 3.    Results

Results are tabulated below. The numbers for the queries with and without prepositions were added together. Any manually excluded cases have already been subtracted; they are not shown in the tables.

**Table 2.** Corpus Brasileiro (text from various genres), subcorpus en (film and television scripts): NC vs. lack of NC.

| | |
|---|---|
| Negative concord | 125 (97.66%) |
| Lack of negative concord | 3 (2.34%) |
| **Total** | **128** |

**Table 3.** Corpus Brasileiro (text from various genres), subcorpus ed (magazines): NC vs. lack of NC.

| | |
|---|---|
| Negative concord | 141 (97.92%) |
| Lack of negative concord | 3 (2.08%) |
| **Total** | **144** |

**Table 4.** C-Oral-Brasil (informal, spontaneous conversations, Minas Gerais):NC vs. lack of NC.

| | |
|---|---|
| Negative concord | 155 (79.49%) |
| Lack of negative concord | 40 (20.51%) |
| **Total** | **195** |

**Table 5.** Corpus Brasileiro (text from various genres), subcorpus en (film and television scripts): NC vs. lack of NC for each NI

| postverbal NI only | with NC | without NC | Total |
|---|---|---|---|
| nada | 103 | 3 | 106 |
| ninguém | 0 | 0 | 0 |
| nenhum(a) | 16 | 0 | 16 |
| nunca | 5 | 0 | 5 |
| **postverbal NI only: total** | **124** | **3** | **127** |
| **postverbal NI + additional não** | | | |
| nada não | 1 | 0 | 1 |
| ninguém não | 0 | 0 | 0 |
| nenhum(a) NP não | 0 | 0 | 0 |
| nunca não | 0 | 0 | 0 |
| postverbal NI + additional não: total | 1 | 0 | 1 |
| **Total** | **125** | **3** | **128** |

**Table 6.** Corpus Brasileiro (text from various genres), subcorpus ed (magazines): NC vs. lack of NC for each NI.

| postverbal NI only | with NC | without NC | Total |
|---|---|---|---|
| nada | 86 | 2 | 88 |
| ninguém | 0 | 0 | 0 |
| nenhum(a) | 50 | 1 | 51 |
| nunca | 5 | 0 | 5 |
| **postverbal NI only: total** | **141** | **3** | **144** |
| **postverbal NI + additional não** | | | |
| nada não | 0 | 0 | 0 |
| ninguém não | 0 | 0 | 0 |
| nenhum(a) NP não | 0 | 0 | 0 |
| nunca não | 0 | 0 | 0 |
| **postverbal NI + additional não: total** | **0** | **0** | **0** |
| **Total** | **141** | **3** | **144** |

**Table 7.** C-Oral-Brasil (informal, spontaneous conversations, Minas Gerais): NC vs. lack of NC for each NI.

| postverbal NI only | with NC | without NC | Total |
|---|---|---|---|
| nada | 102 | 24 | 126 |
| ninguém | 17 | 2 | 19 |
| nenhum(a) | 13 | 2 | 15 |
| nunca | 1 | 0 | 1 |
| **postverbal NI only: total** | **133** | **28** | **161** |
| | | | |
| **postverbal NI + additional não** | | | |
| nada não | 18 | 10 | 28 |
| ninguém não | 3 | 2 | 5 |
| nenhum(a) NP não | 1 | 0 | 1 |
| nunca não | 0 | 0 | 0 |
| **postverbal NI + additional não: total** | **22** | **12** | **34** |
| **Total** | **155** | **40** | **195** |

Chi-square tests of independence were performed on these results. A significant association was found between corpus and proportion of NC vs. lack of NC, $\chi^2(2) = 42.87$, $p < .001$. To make sure the association was only between C-Oral-Brasil on the one hand and the other two corpora on the other hand, another chi-square test of independence was performed only on the two Corpus Brasileiro subcorpora. No significant association was found here, $\chi^2(1) = .02$, $p = .884$. This confirms that the two Corpus Brasileiro subcorpora do not differ significantly in their distribution of NC vs. lack of NC compared to each other, but they do compared to C-Oral-Brasil.

Moving on to the individual NIs, a significant association was found between corpus and the frequency of the individual NIs, $\chi^2(12) = 119.86$, $p < .001$. Because this can be interpreted several ways, the test was performed again on the two Corpus Brasileiro subcorpora only, then on the single postverbal NIs only, and then on the postverbal NIs + não only. A significant association was found in the first case, $\chi^2(3) = 20.08$, $p < .001$, as well as the second, $\chi^2(6) = 72.07$, $p < .001$, but not in the third, $\chi^2(2) = .21$, $p = .899$. This indicates that the corpora differ significantly in their proportions of the single postverbal NIs, but not in their proportions of the NIs + não, although this latter finding is hard to interpret considering NI + não does not occur at all in the magazines subcorpus, and only once in the film and TV scripts subcorpus.

In addition, a significant association was found between the individual NI and the proportion of NC vs. lack of NC, $\chi^2(6) = 30.02$, $p < .001$; but when controlled for corpus, this association did not differ significantly between corpora.

Finally, a binomial test was performed to see if the postverbal frequen-

cies of ninguém and nenhum(a) differed significantly. This indicated that the proportion of nenhum(a) .81 was significantly higher than the expected .50 on the whole, $p < .001$; however, this difference disappeared when looking only at C-Oral Brasil, where the proportion was only .56, $p = .608$.

# 4.    Discussion and conclusion

## 4.1 Discussion

The most immediately striking finding is that nenhum(a) as a postverbal NI appears so much more often than ninguém: 82 times (17.56% of the total amount), whereas ninguém only occurs 19 times (4.07%). At first sight, this seems to be a much greater difference than that reported by Agostini & Schwenter; however, the difference disappears in the spoken language corpus, meaning their findings are in fact replicated. Curiously, postverbal ninguém does not occur at all in the two Corpus Brasileiro subcorpora. This seems to indicate that ninguém as a postverbal NI is markedly informal. Moreover, the magazines subcorpus has a relatively higher incidence of nenhum(a) and a lower incidence of não postverbally than the film and TV scripts subcorpus. Why these frequencies differ remains an open question that cannot be adequately answered without a qualitative analysis.

Another remark regarding frequency is that the combination of NI + não is much more frequent in spoken language; it barely occurs at all in the written language corpora. This should be seen in light of NEG3 (see above) being characteristic of informal language. The language in film and TV scripts does not differ from that in magazines here, which is perhaps surprising if we think of scripts as imitations of spoken language. Not all the lines from the film and TV script subcorpus contained dialogue, though; some consisted of stage directions.

Looking at the syntactic behaviour of the individual postverbal NIs, we see that they differ with respect to their occurrence in NC versus non-NC constructions. In particular, nada is the only postverbal NI that occurs somewhat commonly in non-NC constructions. This is in accordance with Agostini & Schwenter's findings (cf. their discussion of this phenomenon on pp. 18–19 and p. 20).

We can now answer the two research questions as follows. As for the occurrence of postverbal NIs in non-NC constructions, my findings replicate those of Agostini & Schwenter in that nada is preferred for this construction and nunca is dispreferred, while ninguém and nenhum(a) are in the middle. This observation holds across all corpora, with the curious side note that ninguém does not occur postverbally at all in the written language corpora. Thus, hypothesis 1 has been confirmed.

As for non-NC constructions themselves, they occur considerably more in spoken language compared to written language, in line with expectations; however, their occurrence in film and TV scripts do not differ significantly from that of magazines. Thus, hypothesis 2 has been partly confirmed and partly rejected, but this says more about the relatively formal language used in Brazilian films and TV series compared to spontaneous conversations.

wFinally, I hypothesized that the data from Minas Gerais would match Agostini & Schwenter's data for São Paulo. Unfortunately, their data consists of acceptability judgments, whereas I only have access to frequencies. This may be

considered a methodological weakness of the present study. Even if we allow acceptability to function as something of a proxy for frequency, the data are difficult to compare because the other two corpora do not contain geographical information. That said, we can state that lack of NC is relatively frequent in Minas Gerais. However, this is weak and no definitive conclusion can be drawn. Thus, hypothesis 3 is tentatively supported, but more research is needed.

## *4.2 Conclusion*

In summary, we can conclude that the factors determining the presence or absence of NC are a combination of language-external (formality, genre, perhaps geographical) and language-internal ones (specific postverbal NIs occur more often in non-NC constructions than others, which may have semantic as well as syntactic reasons). These factors show complex interactions regarding individual NIs, cf. the variable occurrence of nenhum(a) in magazines compared to film and TV scripts and that of ninguém in spoken compared to written language. This matter calls for thorough, especially qualitative, research that looks at the pragmatics, semantics and syntax of the variation in the context of the situation, communicative genre, and the properties of the individual speakers.

# **References**

Acesso a corpos de português: Projeto AC/DC [Access to Portuguese corpora: Project AC/DC]. 2019, January 13. Linguateca. Web page. Retrieved from: https://www.linguateca.pt/ACDC/.

Agostini, T. & S. Schwenter. 2015. Variable Negative Concord in Brazilian Portuguese: Acceptability and Frequency, presented at the Hispanic Linguistic Symposium, Urbana, September 24–27, 2015. Retrieved from: https://www.researchgate.net/profile/Scott_Schwenter/publication/292615304_Variable_Negative_Concord_in_Brazilian_Portuguese_Acceptability_and_Frequency/links/56b0430e08ae-9c1968b496d6/Variable-Negative-Concord-in-Brazilian-Portuguese-Acceptability-and-Frequency.pdf.

Bick, E. 2000. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Thesis. Aarhus: Aarhus University Press.

Evert, S. & The CWB Development Team. 2016. The IMS Open Corpus Workbench (CWB): CQP Query Language Tutorial, CWB Version 3.4. Retrieved from: http://cwb.sourceforge.net/files/CQP_Tutorial.pdf.

Polon, L. (n.d.). Quantos estados tem o Brasil? Veja a resposta! [How many states does Brazil have? See the answer!] Estudo Kids. Web page. Retrieved from: https://www.estudokids.com.br/quantos-estados-tem-o-brasil-veja-a-resposta/.

SoldadoTrifadon. 2018, November 17. "12 jeitos de dividir o Brasil [12 ways of dividing Brazil]". Reddit. Online forum. Retrieved from: https://www.reddit.com/r/brasil/comments/9xvs76/12_jeitos_de_dividir_o_brasil/