

Negative pronominal forms in Brazilian Portuguese: from lexical classes to corpus annotation

Isaac Souza de Miranda Jr.

Université catholique de Louvain/Universidade Federal de São Carlos

isc_jr@live.com

Negative forms such as *nada*, *ninguém*, and *nenhum* are traditionally grouped together as negative pronouns in descriptions of Brazilian Portuguese. Yet their pronominal status is not a property of lexical form alone: the same item may function as a nominal argument, a determiner, or an adverbial modifier. This raises a basic question for the study of Portuguese pronominal systems and for corpus annotation: what counts as a negative pronoun when lexical identity, syntactic function, and negative meaning do not align transparently? This paper addresses this question through a corpus-based analysis of seven Universal Dependencies (UD) Portuguese treebanks — namely: UD_Portuguese Bosque Rademaker et al. (2017), CINTIL-UDep (Branco et al. 2022), DANTEStocks (Felippo et al. 2024), Portuguese-GSD (McDonald et al. 2013), Portuguese-PUD (Zeman et al. 2017), PetroGold (Souza & Freitas 2022), and Porttinari-base (Duran et al. 2023) — and proposes a function-sensitive enrichment of the Porttinari-base corpus.

The contribution is threefold. First, we show that Portuguese UD treebanks differ in how they encode negative pronominal forms, especially in the use of `PronType=Neg` versus `PronType=Ind`. Second, we argue that the annotation problem is not whether a lexical item belongs to a list of negative pronouns, but whether a token realizes a pronominal function in context. Third, we apply this criterion to Porttinari-base, a gold-standard Brazilian Portuguese treebank with 8,418 sentences and 168,080 tokens, enriching 179 tokens.

UD (Nivre et al. 2020; de Marneffe et al. 2021) provides crosslinguistic guidelines for parts of speech, morphological features, and syntactic relations. It includes `PronType=Neg`, a feature designed to encode negative pronominal values.¹ In principle, this feature distinguishes negative pronominal forms from merely indefinite ones. In practice, however, the Portuguese treebanks investigated here use it unevenly. `PronType=Neg` is found in Bosque for *nada* and *nenhum*, and in GSD only for *nenhum*, as shown in Appendix Appendix 1. Other occurrences are annotated with `PronType=Ind`, which preserves indefiniteness but not negative polarity. This variation shows that current corpus annotation does not consistently distinguish lexical negativity, indefiniteness, and pronominal function.

To define the relevant contexts for enrichment, we combine the UD guidelines with the descriptive taxonomy of negative elements in Brazilian Portuguese proposed by Miranda Jr. (2022). This taxonomy shows that the three forms investigated here do not behave uniformly. *Nada* may function as a pronoun, as in example A2i (*Não disse nada*), or as an adverbial modifier, as in example A2ii (*Não estava nada à espera disto*). *Nenhum* has pronominal uses, especially in partitive constructions with *de*, as in example A2iii (*Nenhum dos homens mostrou clemência*), but may also function as a determiner, as in example A2iv (*Nenhum aluno foi aprovado*). By contrast, *ninguém* is restricted to pronominal use, as in example A2v (*Não veio ninguém*).

On this basis, we propose assigning `PronType=Neg` to tokens of *nada*, *ninguém*, and *nenhum* when they realize a negative pronominal value while retaining `Definite=Ind` to preserve their indefinite interpretation. The analysis, therefore, implies that negative pronouns should not be defined as members of a fixed lexical class but as context-dependent realizations in which pronominal function and negative value coincide. Their classification must conse-

¹Universal Dependencies `PronType`

quently distinguish pronominal tokens from determiner and adverbial uses of the same lexical forms, while theoretical models should represent lexical identity, syntactic function, indefiniteness, and negative value as distinct but interacting dimensions. In practical terms, this function-sensitive description provides a principled basis for the consistent identification and computational encoding of negative pronominal values in natural language processing.

Appendix

A1. Values of PronType for investigated pronouns in the corpora

Corpus	nada	ninguém	nenhum
Bosque	Neg	Ind	Neg
CINTIL	Ind	Ind	Ind
DANTEStocks	Ind	Ind	Ind
GSD	Ind	Ind	Neg
PUD	—	—	—
PetroGold	—	—	Ind
Porttinari-base	Ind	Ind	Ind

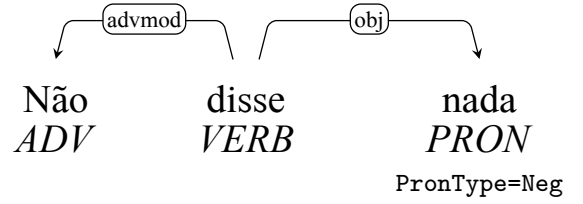
Table 1: ‘Ind’ indicates PronType=Ind, ‘Neg’ indicates PronType=Neg, and ‘—’ indicates that the item has no PronType annotation.

A2. Examples

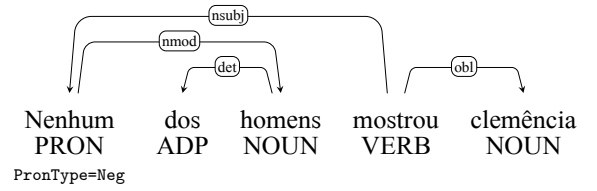
- (i) Não disse nada.
not say.PST.3SG nothing
‘S/he didn’t say anything.’
- (ii) Não estava nada à espera disto.
not be.IMPF.3SG nothing to.the expectation of.this
‘S/he was not expecting this at all.’
- (iii) Nenhum dos homens mostrou clemência.
none of.the men show.PST.3SG mercy
‘None of the men showed mercy.’
- (iv) Nenhum aluno foi aprovado.
no student be.PST.3SG approve.PTCP
‘No student passed.’
- (v) Não veio ninguém.
not come.PST.3SG nobody
‘Nobody came.’

A3. Trees

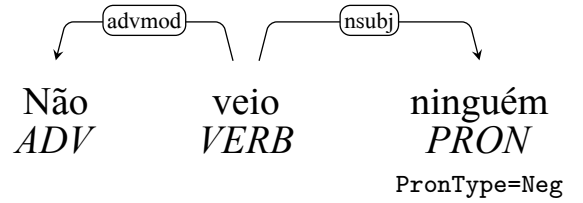
- (i) Não disse nada



- (ii) Nenhum dos homens mostrou clemência



- (iii) Não veio ninguém



Bibliographical References

- Branco, António, João Ricardo Silva, Luís Gomes & João António Rodrigues. 2022. [Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support](#). In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022)*, 5617–5626. Marseille, France: European Language Resources Association.
- Duran, Magali, Lucelene Lopes, Maria das Graças Nunes & Thiago Pardo. 2023. [The dawn of the porttinari multigenre treebank: Introducing its journalistic portion](#). In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 115–124. Belo Horizonte, MG, Brazil: SBC.
- Felippo, Ariani Di, Maria das Graças Volpe Nunes & Bryan Khelven Barbosa. 2024. [A dependency treebank of tweets in Brazilian Portuguese: Syntactic annotation issues and approach](#). In Daniela Barreiro Claro & Adriana Pagano (eds.), *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, 86–94. Belém do Pará, Brazil: Association for Computational Linguistics.
- de Marneffe, Marie-Catherine, Christopher D. Manning, Joakim Nivre & Daniel Zeman. 2021. [Universal dependencies](#). *Computational Linguistics* 47(2). 255–308.
- McDonald, Ryan, Joakim Nivre, Yvonne Quirnbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Beldini, Núria Bertomeu Castelló & Jungmee Lee. 2013. [Universal dependency annotation for multilingual parsing](#). In Hinrich Schuetze, Pascale Fung & Massimo Poesio (eds.), *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 92–97. Sofia, Bulgaria: Association for Computational Linguistics.
- Miranda Jr., Isaac Souza de. 2022. [Não é nada não: uma análise das lexias negativas do português para anotação nas Universal Dependencies](#): Universidade Federal de São Carlos dissertation.
- Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. [Universal dependencies v2: An evergrowing multilingual treebank collection](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference (LREC 2020)*, 4034–4043.
- Rademaker, Alexandre, Fabricio Chalub, Livy Real, Claudia Freitas, Eckhard Bick & Valeria Paiva. 2017. [Universal dependencies for Portuguese](#). In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, 197–206.
- Souza, Elvis & Claudia Freitas. 2022. [Polishing the gold – How much revision do we need in treebanks?](#) In Thiago Alexandre Salgueiro Pardo, Ariani Di-Felippo & Norton Trevisan Roman (eds.), *Proceedings of the Universal Dependencies Brazilian Festival*, 1–11. Fortaleza, Brazil: Association for Computational Linguistics.
- Zeman, Daniel, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg

Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj & Josie Li. 2017. [CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies](#). In Jan Hajič & Dan Zeman (eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 1–19. Vancouver, Canada: Association for Computational Linguistics.