



languages

Special Issue Reprint

Investigating Language Variation and Change in Portuguese

Edited by
Patrícia Amaral

mdpi.com/journal/languages



Investigating Language Variation and Change in Portuguese

Investigating Language Variation and Change in Portuguese

Editor

Patrícia Amaral



Basel • Beijing • Wuhan • Barcelona • Belgrade • Novi Sad • Cluj • Manchester

Editor

Patrícia Amaral
Department of Spanish
and Portuguese
Indiana University
Bloomington
United States

Editorial Office

MDPI AG
Grosspeteranlage 5
4052 Basel, Switzerland

This is a reprint of articles from the Special Issue published online in the open access journal *Languages* (ISSN 2226-471X) (available at: https://www.mdpi.com/journal/languages/special_issues/HAXETJT8N2).

For citation purposes, cite each article independently as indicated on the article page online and as indicated below:

Lastname, Firstname, Firstname Lastname, and Firstname Lastname. Article Title. <i>Journal Name</i> Year , Volume Number, Page Range.
--

ISBN 978-3-7258-2062-7 (Hbk)

ISBN 978-3-7258-2061-0 (PDF)

doi.org/10.3390/books978-3-7258-2061-0

© 2024 by the authors. Articles in this book are Open Access and distributed under the Creative Commons Attribution (CC BY) license. The book as a whole is distributed by MDPI under the terms and conditions of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) license.

Contents

About the Editor	vii
Patrícia Amaral Introduction to the Special Issue: Investigating Language Variation and Change in Portuguese Reprinted from: <i>Languages</i> 2024, 9, 288, doi:10.3390/languages9090288	1
Tjerk Hagemeijer, Rita Gonçalves and Nélia Alexandre Spatial Locative Relativization in Three African Varieties of Portuguese: Unity in Diversity and Diversity in Unity Reprinted from: <i>Languages</i> 2024, 9, 83, doi:10.3390/languages9030083	8
Diana Santos On the Variability of Portuguese Duration Adverbials with <i>Por</i> and <i>Durante</i> Reprinted from: <i>Languages</i> 2024, 9, 94, doi:10.3390/languages9030094	32
Cíntia Pacheco, Ana Carvalho and Marta Pereira Scherre When a New Pronoun Crosses the Border: The Spread of <i>A Gente</i> on the Brazilian-Uruguayan Frontier Reprinted from: <i>Languages</i> 2024, 9, 98, doi:10.3390/languages9030098	48
Scott A. Schwenter, Lauren Miranda, Ileana Pérez and Victoria Cataloni Reanalyzing Variable Agreement with <i>tu</i> Using an Online Megacorporus of Brazilian Portuguese Reprinted from: <i>Languages</i> 2024, 9, 197, doi:10.3390/languages9060197	67
Kendra V. Dickinson Regularization and Innovation: A Usage-Based Approach to Past Participle Variation in Brazilian Portuguese Reprinted from: <i>Languages</i> 2024, 9, 52, doi:10.3390/languages9020052	82
Ana Guilherme <i>Fostes tu?</i> : Analogical Change in European Portuguese and the Case of the Second Person Singular in the Simple Past (Indicative) Reprinted from: <i>Languages</i> 2024, 9, 176, doi:10.3390/languages9050176	101
Camila Lívio and Chad Howe Text Mining Approaches to Language Use in Social Media: The Case of Portuguese <i>Buê</i> Reprinted from: <i>Languages</i> 2024, 9, 82, doi:10.3390/languages9030082	123
Manuel Delicado Cantero and Patrícia Amaral Marching towards Contrast: The Case of <i>ao passo que</i> in Portuguese Reprinted from: <i>Languages</i> 2024, 9, 6, doi:10.3390/languages9010006	140
Lukas Müller European Portuguese <i>lá</i> : Use-Conditional Meaning and Pragmaticalization Reprinted from: <i>Languages</i> 2024, 9, 189, doi:10.3390/languages9060189	157
Clara Pinto On the Emergence of Portuguese FCI <i>qualquer</i> : A Diachronic Perspective Reprinted from: <i>Languages</i> 2023, 8, 290, doi:10.3390/languages8040290	178
Ana Regina Calindro A Diachronic Overview of the Prepositional Accusative in Portuguese Reprinted from: <i>Languages</i> 2024, 9, 194, doi:10.3390/languages9060194	207

Aroldo Leal de Andrade and Lara da Silva Cardoso

Word Order in Colonial Brazilian Portuguese: Initial Findings

Reprinted from: *Languages* 2024, 9, 269, doi:10.3390/languages9080269 227

About the Editor

Patrícia Amaral

Patrícia Amaral is a Professor in the Department of Spanish and Portuguese at Indiana University in Bloomington. Her research areas are semantics/pragmatics and historical linguistics. Her work has focused in particular on typologies of meaning, focus particles, and aspectual adverbs. She has also worked extensively on syntactic and semantic change, with a particular emphasis on Portuguese and Spanish.

Editorial

Introduction to the Special Issue: Investigating Language Variation and Change in Portuguese

Patrícia Amaral

Department of Spanish and Portuguese, Indiana University, Bloomington, IN 47405-1105, USA; pamaral@iu.edu

1. This Special Issue

This Issue focuses on morphosyntactic variation and change in Portuguese. It features twelve papers based on data from multiple corpora of Portuguese, representing different regional and historical varieties: (contemporary) Angolan, Brazilian, European, Mozambican, Santomean, and Uruguayan Portuguese, as well as Medieval and Classical Portuguese. Several of these varieties have been underrepresented in the literature. This broad coverage makes it possible to showcase both the range of variation and the shared features of Portuguese. The papers adopt various theoretical approaches and use a range of current methods. However, there are also some common trends and connections among the presented studies, which this collection aims to bring to the fore. In this introduction, I highlight those connections and provide some context on the contribution of the papers to the study of Portuguese and more broadly to linguistic theories.

In the preceding decades, several reference works on Portuguese have been published, providing overviews of the language's subsystems and including descriptions of grammatical structures that display variation (Wetzels et al. 2016; Carrilho and Martins 2016). Some studies focus on specific varieties (e.g., Gonçalves 2010; Bouchard 2017), examine certain phenomena from a variationist perspective (e.g., Malvar and Poplack 2008; Scherre et al. 2018), or highlight the role of pragmatic factors constraining the use of variable structures (e.g., Schwenter and Silva 2002; Posio 2021). However, there are few entire works devoted to variation in Portuguese—a notable exception being the work of (Barbosa et al. 2017). More broadly, there is a need to consider Portuguese data to examine the connection between language variation and change, and to reflect upon the implications of data from Portuguese for theories of syntactic and semantic change. This Special Issue aims to fill this gap.

An additional goal of this volume is to showcase research on topics that have been underexplored in the literature on Portuguese. A case in point is the study of semantic change, which has received significantly less attention than phonetic or morphosyntactic change in the Portuguese context. This volume includes several studies focusing on changes in meaning: these studies examine the change undergone by intensifiers, adversative connectives, free-choice items, and items conveying expressive meaning. All these categories have been understudied in Portuguese, both in terms of synchrony and diachrony. From the perspective of variation, topics that have not received much attention in the literature on Portuguese and are examined in this volume include prepositional accusatives (more broadly known within the phenomenon of differential object marking), variation involving morphological paradigms (such as the case of the second-person singular form of the Preterit—*Pretérito Perfeito Simples*—and the two forms of past participles), variation in the domain of tense and aspect (including temporal adjuncts of duration), and variants of locative relative clauses.

It is also worth mentioning that this volume showcases innovative methods in linguistics. The studies presented examine data from new corpora of private correspondence, corpora of various genres, and data from social media. As for the new methodological

Citation: Amaral, Patrícia. 2024. Introduction to the Special Issue: Investigating Language Variation and Change in Portuguese. *Languages* 9: 288. <https://doi.org/10.3390/languages9090288>

Received: 31 July 2024

Revised: 4 August 2024

Accepted: 21 August 2024

Published: 25 August 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

approaches represented in this volume,¹ there are several corpus-based studies using distributional approaches and a range of statistical analyses.

From a theoretical perspective, there are some common threads across the papers in this volume. Some articles (e.g., Schwenter et al., Dickinson) highlight the importance of considering lexical factors in phenomena of variation and change, a trend found in other research on Romance, and more developed within studies using a constructional or usage-based approach. Other articles show the different ways in which analogy plays a role in variation and change, an active area of research in studies of language change. The research in the following papers reveals how analogical processes are constrained by various types of language-internal factors: Dickinson, Guilherme, and Delicado Cantero and Amaral. Another set of articles explores the mechanisms of and motivations for semantic change: these are authored by Livio and Howe, Müller, and Delicado Cantero and Amaral. In addition, some articles identify factors at the semantics–pragmatics interface that affect morphosyntactic phenomena, e.g., the research of Hagemeyer et al., and Pinto.

Several papers in this Special Issue highlight similarities and differences between varieties of Portuguese that had not been previously noticed, e.g., in the domain of intensifiers, relative constructions, and the variation between the first-person plural pronouns *a gente* and *nós*. Other papers contribute to improving the characterization of previous stages in the history of Portuguese or the identification of new stages. All in all, the studies cover multiple grammatical interfaces, e.g., between syntax and information structure, and between morphosyntax and pragmatics.

In the following, I provide a summary of the papers, starting with those that focus on variation.

2. Summary of the Contributions to This Special Issue

Tjerk Hagemeyer, Rita Gonçalves, and Nélia Alexandre examine variation in relative constructions denoting spatial location, i.e., the use of *onde* ‘where’ vs. *que* ‘that’, in spoken corpora of urban varieties of Angolan, Mozambican, and Santomean Portuguese. Their paper shows that these varieties are rather homogeneous in this domain, mainly displaying *pied-piping* as the main mechanism, and secondarily showing P(reposition)-chopping. As for the examined variants, there are some differences among the three African varieties. Overall, *onde* tends to be used with locatives and areas, but Angolan Portuguese also uses *onde* for goals/sources, and Santomean Portuguese displays more variation between both relative markers in terms of areas and containers. Among the semantic factors analyzed, it was found that antecedents that are [-definite] and display less specificity tend to favor the relativizer *que* in the three varieties, and especially so in Santomean Portuguese. The authors note that, despite the variation described in this study for these African varieties, the main patterns identified are also found in European Portuguese. This study paves the way for more research on the semantic properties of head nouns as well as research on relative constructions using experimental methods.

In a study comparing European and Brazilian Portuguese, Diana Santos examines the variation between temporal adverbials of duration containing *por*, ‘for’, and *durante*, ‘during’. Using corpora data from the Linguatca resource encompassing different genres (newspaper texts, literary fiction, and oral interviews), the author first shows the overall differences in frequency across varieties, with *durante* being more frequent than *por* in European Portuguese, and *por* being more frequently used than *durante* in Brazilian Portuguese. Importantly, this paper is the first to empirically test previous claims—based on introspection—about the distribution of these temporal adverbials and to provide a quantitative analysis of corpus data. This study delves into their distributional properties and the sensitivity of duration adverbials to semantic and pragmatic factors. *Por* and *durante* show distinct preferences for aspectual classes of verbs: *por* may co-occur with achievement predicates, and in such cases measures the duration of the result of the event, while *durante* occurs with states and activities. In addition, *por* and *durante* are sensitive to semantic notions like duration length, with *por* being used more often for short durations.

This is reflected both in the co-occurrence of *por* with nouns denoting short time periods (as determined by their semantic value or by a combination of their encoded meaning and world knowledge) and in its compatibility with the copula verbs *ser* and *estar*. The author also discusses the patterns reported in the literature with respect to a preference of *por* for the expression of planned duration², but finds no evidence for a categorical choice in such contexts. An interesting connection identified by this article is the relation between these adverbials and the reinforcement of negation through what could be called temporal minimizers, like *um momento*, ‘one moment’, or *um instante*, ‘an instant’, under the scope of negation (e.g., *nem por um instante*, ‘not even for an instant’).

Another paper that uses corpus data to analyze variation is **Scott Schwenter, Lauren Miranda, Ileana Pérez, and Victoria Cataloni**’s study of agreement with the pronoun *tu* in Brazilian Portuguese. Unlike previous studies, which used sociolinguistic interviews to study this phenomenon, the authors examine almost 5000 tokens from an online mega-corpus of BP that contains data from multiple varieties in terms of region and register. Their study finds that the most common pattern is the co-occurrence of the pronoun *tu* with a verb form in the third-person singular (e.g., *tu é*), with agreement in person between pronoun and verb form (e.g., *tu és*) being infrequent. The statistical analyses performed show that this type of morphosyntactic variation is mostly constrained by lexical factors, with certain verbs showing a strong tendency to lack second-person singular agreement, while other verbs tend to display agreement. The authors also identify “significant constructional effects”, i.e., patterns of variation that arise in combination with sequences involving tense/aspect and auxiliary verbs. Not only do these results reveal an overwhelming tendency to mark person and number features through personal pronouns rather than verbal inflection in BP, but they also confirm a diachronic tendency in this variety to erode verbal morphology.

The paper by **Cíntia Pacheco, Ana Carvalho, and Marta Pereira Scherre** also focuses on morphosyntactic variation and the pronominal system of Portuguese. This study adopts the approach of comparative sociolinguistics to investigate the variation between the first-person plural pronouns *a gente* and *nós* in Uruguayan Portuguese (a variety of Portuguese spoken in northern Uruguay by Spanish–Portuguese bilinguals) and the neighboring monolingual BP variety. The data come from sociolinguistic interviews conducted in two communities at the southern border of Brazil: Aceguá, Brazil, and Aceguá, Uruguay. This study finds that while both pronominal forms are used by bilingual speakers, the form *a gente* is used less frequently in Uruguayan Portuguese than in Brazilian Portuguese. Nevertheless, the variationist analysis shows that the grammaticalized form *a gente* is favored in both linguistic systems by the same linguistic constraints. While the focus of this Special Issue is not on language contact, this study shows how certain linguistic factors (e.g., verb tense, the phonic salience of a verb form, discourse persistence of a variant) regulate the use of this pronominal variable, both for monolingual and bilingual speakers.

The papers by **Kendra Dickinson** and **Ana Rita Guilherme** look at morphosyntactic variation and specifically discuss analogy and the role of language-internal factors in language variation and change. **Kendra Dickinson** investigates the variation in past participle forms (i.e., the co-existence of irregular participles and regularized ones, or *participios duplos*, e.g., *aberto* and *abrido* for the verb *abrir* ‘open’) in contemporary Brazilian Portuguese. Adopting a usage-based approach, the author examines the role of analogy in the regularization of participial forms, which has had a lasting effect on the history of the language. Using Sketch Engine’s 2011 Portuguese Web Corpus (ptTenTen11), a corpus of 4.6 billion words, the author finds a set of language-internal constraints on the analogical processes affecting participles, whereby certain constructions provide conditioning environments. She finds that Perfect tenses (e.g., the *Pretérito Perfeito Composto*) promote the regularization of Latinate irregular past participles and are a conserving environment for regular past participles that compete with short-term forms. On the other hand, short-form participles (i.e., innovative forms) tend to occur in adjectival contexts with *ser* and *estar*. In addition, some of the variation seems to be lexically constrained, which aligns with other findings

in this Issue (cf. **Schwenter et al.**). More broadly, this paper sheds light on how usage patterns and system-internal factors affect analogy, an important mechanism of language change.

Another paper that highlights the role of analogy in language change is the study by **Ana Rita Guilherme**. The author focuses on the variation in the form of the second-person singular simple past (*Pretérito Perfeito Simples*) indicative, e.g., *estiveste* vs. *estivestes* ‘you were’, which has been attested since the 18th century in the history of European Portuguese. Using two corpora of private correspondence from the 16th to the 20th centuries (*Post Scriptum—PS—and Forgotten Letters—Fly*), the author traces the process of the analogical extension of the *-s* form. **Guilherme** argues that both morphological and pragmatic factors should be taken into account to explain this analogical extension throughout the history of the language. She presents the following factors: (i) regularization based on other second-person singular forms; (ii) ambiguity and variation within verbal paradigms in Medieval and Classical Portuguese as a consequence of the deletion of the suffix *-d* (*-de/-des*) of the second-person plural form in intervocalic contexts; and (iii) changes in the Portuguese system of address forms, involving competition between the pronominal forms *tu* and *vós* and the verbal inflection associated with them. The research by **Guilherme** straddles the study of variation and the study of change. In the following, I summarize the articles that focus more on change.

Lukas Müller investigates non-adverbial uses of *lá* in European Portuguese (e.g., *Acaba lá de comer a sopa!* ‘(Just) eat the soup!’) and their diachrony. Adopting the framework of multidimensional semantics (Gutzmann 2013, 2015), the author proposes that these uses of *lá* do not contribute to the truth conditions of a sentence, and hence should be analyzed as use-conditional. Specifically, in contemporary EP, use-conditional *lá* contributes illocutionary modification. Using data from the *Corpus do Português*, the author provides a diachronic study of semantic change, showing that, over time, the adverb conveying spatial deictic meaning underwent a process of pragmaticalization (Traugott 1989). This process yields a type shift, changing from contributing to truth conditions to contributing to use conditions: $lá_{TC} > lá_{UC}$. The author proposes a three-stage diachronic path whereby an implicature of illocutionary modification becomes conventionalized over time, creating a hybrid, two-dimensional item, which is then split into two homophonous items, yielding the contemporary situation of polysemy found in EP. Importantly, this is one of the few studies that analyzes semantic change in Portuguese adopting a multidimensional semantics approach, and more broadly focusing on the emergence of expressive meaning in Portuguese.

Another study of semantic change in European Portuguese is the paper by **Manuel Delicado Cantero** and **Patrícia Amaral**. This paper traces the syntactic and semantic shifts undergone by the PP *ao passo que*, ‘at the step/pace that’. This originally appeared with a relative clause modifying the noun *passo* and denoted a comparison in pace. It then became a temporal expression denoting a simultaneous gradual change in two dynamic eventualities, and eventually became a contrastive connective. Using data from the *Corpus do Português*, the authors identify contexts in which the expression conveys both temporal–proportional meaning and a contrast between two events. Over time, this led to a broadening of the aspectual classes that were compatible with the expression as it came to encode just contrast (hence losing aspectual restrictions connected to its source meaning). In addition, the authors examine the internal structure of the expression and the meaning of the noun *passo*, ‘step’, and discuss how these data can help adjudicate between competing theories regarding the nature of complex categories. This paper fills a gap in our understanding of the diachrony of the expression *ao passo que*. Overall, there is a lack of studies on the diachrony of contrastive connectives in Portuguese (an exception being the diachronic study of adversatives in the work of Amaral et al. 2023).

Camila Livio and **Chad Howe** use a Digital Humanities approach to study the interplay between variation and change in the system of degree modifiers in Portuguese. Specifically, they mine Twitter (now X) data to investigate the distribution and mean-

ing of the intensifier *bué*, possibly a borrowing from Kimbundu, in Angolan Portuguese. Their analysis shows the widespread use of *bué*, with functions of intensification and quantification. In a broader comparison among varieties of Portuguese with respect to the expression of intensification, the authors argue that the usage patterns of *bué* correspond to the use of *bem* (vs. *muito*) in Brazilian Portuguese. Their study also examines the use of *bué* in European Portuguese, a borrowing resulting from a well-attested motivation for change, which is innovation and expressivity in the domain of intensification. In this variety, *bué* has undergone grammaticalization and nowadays co-occurs with both gradable and non-gradable predicates. This paper adopts a distributional approach to advance our understanding of intensification and, more broadly, degree modification in Portuguese, an area that has only recently begun to be explored (see e.g., Gomes 2011; Pinto 2020; Amaral 2020).

Clara Pinto investigates the diachrony of the free-choice item (FCI) *qualquer*, ‘any’, from *qual*, ‘which’, and a form of the verb, *querer*, ‘to want’, in Portuguese, an understudied topic despite the widespread interest in FCIs in the Romance languages. The author focuses on data from Old Portuguese to examine the syntactic and semantic properties of the original forms, adopting a formal syntax approach. She analyzes the different structures that appear in the medieval data and proposes that there are two different diachronic sources for *qualquer*, underlying the different contemporary interpretations of the word in European Portuguese. Specifically, **Pinto** argues for the existence of two items in the historical data: (i) prenominal *qualquer* displays the properties of a quantifier in the 13th century, while (ii) postnominal *qualquer* should rather be considered an adjectival form resulting from the reanalysis of an appositive relative clause. This paper not only advances our knowledge of FCIs in Portuguese, but it also furthers our understanding of understudied domains of the grammar of Portuguese, like the syntax and semantics of polarity items as well as the properties of determiners that defy classification (e.g., *outro*, ‘other’, and its cross-linguistic counterparts).

Ana Regina Calindro’s study examines the variation and change in *a*-marked arguments in the history of Portuguese. Overall, it is shown that *a*-marked accusatives (“prepositional accusatives”) are much less frequent than bare accusatives, and this is true both for European and Brazilian Portuguese. Using the *Tycho Brahe Corpus of Historical Portuguese*, the author analyzes texts from the 16th to the 19th century and confirms previous claims that *a*-marked accusatives increased in Portuguese in the 17th century, and then decreased in the 18th century. **Calindro** examines specific verb classes and the prepositions associated with their arguments in European and Brazilian Portuguese, noting the differences between these varieties. Finally, the author proposes an analysis of psychological verbs, where contemporary European Portuguese displays a structural marked dative corresponding to the semantic role of Experiencer (*O filme agradou ao João/agradou-lhe*), while Brazilian Portuguese does not display overt marking for this role (*O filme agradou o João/agradou ele*). First, this study provides further empirical evidence that differential object marking is a marginal phenomenon in Portuguese, unlike what happens in other Romance languages. Second, it brings to the fore the behavior of specific verb classes with respect to argument structure, a topic deserving further study using both historical and lexical semantic approaches.

In their research, **Aroldo Leal d’Andrade** and **Lara da Silva Cardoso** explore a recent hypothesis that Brazilian Portuguese may have evolved from a grammatical system, Colonial Brazilian Portuguese, which should be distinguished from Classical Portuguese. The authors focus on word order as they compare two texts from the 17th century, one written by a Brazilian author (Eusébio de Matos), and another written by a Portuguese author (Padre António Vieira). The texts are similar from both a content and a literary perspective, but they display differences that allow **Andrade** and **Cardoso** to propose an emerging distinctiveness in the grammar of their authors. Specifically, they show that in main clauses Colonial Portuguese has more V2 sentences, while Colonial Brazilian Portuguese displays a higher percentage of clauses with V1 word order. The authors also present other differences

in information structure configurations with various types of constituents, arguing that there is less variation in the types of marked constructions in Colonial Brazilian Portuguese than in Colonial Portuguese. Adopting a generative diachronic syntax approach, the authors conclude that the frequency differences observed in these two texts are compatible with the potential microparametric changes that have been proposed for BP and EP.

3. Looking Ahead

First and foremost, the contributions of this Special Issue highlight the need for further research on underrepresented varieties of Portuguese. Beyond those studied in this volume, I would mention, e.g., the importance of learning more about Judeo-Portuguese. Second, the articles in this Issue open the way to research on domains like the syntax–semantics of degree expressions, polarity, indefinites, and tense and aspect, which have been generally understudied in Portuguese (and usually limited to descriptions of European or Brazilian Portuguese). In this respect, the range of methodologies used to investigate these phenomena in this Special Issue are a good indication of the potential and vitality of the field. Third, much work remains to be conducted on the history of Portuguese. From the point of view of data, it would be important to expand diachronic corpus data to include genres and texts that go well beyond canonical and literary sources. Having a broader range of texts would allow us not only to better date changes in the language but also to gain a more fine-grained understanding of factors affecting different grammatical subsystems over time. From the point of view of subjects, I would underscore the critical need for more historical and diachronic studies on the pronominal system(s) of Portuguese, the evolution of temporal and aspectual categories, and the modal system (including both the verbal and the adverbial domain).

Acknowledgments: All the papers in this Issue underwent a double-blind peer-review process. I cannot thank enough the anonymous reviewers for contributing their time and expertise. I also thank Macy Maas for her help with editing. Finally, I want to express my gratitude to the internal editor from MDPI, whose guidance and assistance has been invaluable in bringing this project to light.

Conflicts of Interest: The author declares no conflict of interest.

Notes

- ¹ For studies that adopt NLP tools to investigate language change in Portuguese, see e.g., (Amaral et al. 2023; Bico et al. 2024).
- ² The notion of planned duration, which is identified by Rodrigues (1994); Móia and Alves (2013), provides further evidence for a linguistically relevant category of plan; see Amaral and Del Prete (2016) on the concept of plan at an ontological level and its relevance for the use of modal adverbs in EP and Italian.

References

- Amaral, Patrícia. 2020. Bocado: Scalar semantics and polarity sensitivity. *Zeitschrift für romanische Philologie* 136: 1114–36. [CrossRef]
- Amaral, Patrícia, and Fabio Del Prete. 2016. On truth unpersistence: At the crossroads of epistemic modality and discourse. *Natural Language and Linguistic Theory* 34: 1135–65. [CrossRef]
- Amaral, Patrícia, Zuoyu Tian, Dylan Jarrett, and Juan Escalona Torres. 2023. Tracing semantic change in Portuguese: A distributional approach to adversative connectives. *Journal of Historical Linguistics* 13: 1–34. [CrossRef]
- Barbosa, Pilar, Maria da Conceição de Paiva, and Celeste Rodrigues, eds. 2017. *Studies on Variation in Portuguese*. IHL. Amsterdam: John Benjamins.
- Bico, Maria Inês, Esperança Cardeira, Jorge Baptista, and Fernando Baptista. 2024. Support Verb Constructions in Medieval Portuguese: Evidence from the CTA Corpus. Paper presented at the 16th International Conference on Computational Processing of Portuguese, Santiago de Compostela, Spain, March 14–15; Kerrville: Association for Computational Linguistics, vol. 2, pp. 118–29.
- Bouchard, Marie-Eve. 2017. Linguistic Variation and Change in the Portuguese of São Tomé. Ph.D. dissertation, New York University, New York, NY, USA.
- Carrilho, Ernestina, and Ana Maria Martins, eds. 2016. *Manual de linguística portuguesa*. Boston: De Gruyter Mouton.
- Gomes, Ana Quadros. 2011. Uma proposta de distinção semântica para os intensificadores muito e bem. *Estudos Linguísticos* 41: 379–94.
- Gonçalves, Perpétua. 2010. *A génese do Português de Moçambique*. Lisboa: Imprensa Nacional Casa da Moeda.
- Gutzmann, Daniel. 2013. Expressives and beyond. An introduction to varieties of use-conditional meaning. In *Beyond Expressives. Explorations in Use-Conditional Meaning*. (Current Research in the Semantics Pragmatics Interface—CRiSPI 28). Edited by Daniel Gutzmann and Hans-Martin Gärtner. Leiden: Brill, pp. 1–58.

- Gutzmann, Daniel. 2015. *Use-conditional meaning. Studies in multidimensional semantics*. Oxford: Oxford University Press.
- Malvar, Elisabete, and Shana Poplack. 2008. O presente e o passado do futuro no Português do Brasil. In *Anthony Julius Naro e a Linguística no Brasil: Uma homenagem académica*. Edited by Votre Sebastião and Cláudia Roncarati. Rio de Janeiro: Universidade Federal do Rio de Janeiro Press, pp. 186–206.
- Móia, Telmo, and Ana Teresa Alves. 2013. Tempo adjunto e tempo discursivo. In *Gramática do Português*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura and Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, pp. 557–81.
- Pinto, Clara. 2020. Polarity, expression of degree and negation: The vernacular form caraças. *Estudos De Linguística Galega* 12: 25. [CrossRef]
- Posio, Pekka. 2021. A pessoa and uma pessoa: Grammaticalization and functions of a human impersonal referential device in European Portuguese. *Journal of Portuguese Linguistics* 20: 2. [CrossRef]
- Rodrigues, Rosinda. 1994. Os adverbais durante Q N DET e por Q N DET. Duas formas de quantificar a duração. In *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*. Évora: APL, pp. 497–508.
- Scherre, Maria Marta, Lilian Yacovenco, and Anthony Julius Naro. 2018. Nós e a gente no português brasileiro: Concordâncias e discordâncias. *Estudos de Linguística Galega* Número Extraordinário 0: 13–27. [CrossRef]
- Schwenter, Scott A., and Gláucia Silva. 2002. Overt vs. Null Direct Objects in Spoken Brazilian Portuguese: A Semantic/Pragmatic Account. *Hispania* 85: 577–86. [CrossRef]
- Traugott, Elizabeth C. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65: 31–55. [CrossRef]
- Wetzels, W. Leo, Sergio Menuzzi, and João Costa, eds. 2016. *The Handbook of Portuguese Linguistics*. Chichester: John Wiley & Sons, Inc.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Spatial Locative Relativization in Three African Varieties of Portuguese: Unity in Diversity and Diversity in Unity

Tjerk Hagemeijer *, Rita Gonçalves and Nélia Alexandre *

School of Arts and Humanities, University of Lisbon, 1649-004 Lisbon, Portugal; ritamgg@letras.ulisboa.pt

* Correspondence: t.hagemeijer@letras.ulisboa.pt (T.H.); n_alexandre@letras.ulisboa.pt (N.A.)

Abstract: This paper investigates the formation of spatial locative relativization in three African varieties of Portuguese. While research on relative constructions in Portuguese has deserved considerable attention in the literature, it tends to focus on the European and Brazilian varieties, with locative relativization being only marginally addressed. Using data extracted from spoken corpora of contemporary, urban varieties of Angolan, Mozambican, and Santomean Portuguese, we aim to discuss whether there is a correlation between syntactic and semantic variables and the selection of the two main locatively used relative morphemes, *onde* ‘where’ and *que* ‘that’. Overall, the three varieties at stake behave similarly with respect to the analyzed syntactic variables and follow the tendency found in Portuguese varieties toward the use of pied-piping and P-chopping as the dominant relativization mechanisms, independent of the syntactic relation between the antecedent and the relative clause. Semantically, we identified some fine-grained differences between the three varieties, with Santomean Portuguese generally being the outlier or one of the outliers. Crucially, definiteness of the head noun stands out as the one variable that plays a major role in the selection of the relative morpheme: [–definite] head nouns show a proportionally higher preference for *que* in both AP and STP, which is particularly visible with bare nouns in the latter. This motivates the hypothesis that less specified head nouns show a preference for the un(der)specified relative morpheme *que*. We further show that the role of language contact is at best very limited.

Keywords: locative relativization; African varieties of Portuguese; syntax; semantics; definiteness; language variation; language contact

Citation: Hagemeijer, Tjerk, Rita Gonçalves, and Nélia Alexandre. 2024. Spatial Locative Relativization in Three African Varieties of Portuguese: Unity in Diversity and Diversity in Unity. *Languages* 9: 83. <https://doi.org/10.3390/languages9030083>

Academic Editor: Patrícia Amaral

Received: 1 November 2023

Revised: 1 February 2024

Accepted: 7 February 2024

Published: 29 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Relative clauses have been extensively studied from a wide range of perspectives, including typology and variation. Portuguese is no exception, with research focusing mostly on the European and Brazilian varieties (e.g., Alexandre 2000; Brito 1991; Kato 1993; Kenedy 2007; Peres and Mória 1995; Tarallo 1985; Veloso 2013), whereas relativization in African varieties of Portuguese (AVP) has been the object of a relatively limited number of studies (e.g., Alexandre et al. 2011a, 2011b; Alexandre and Lopes 2022; Brito 2001, 2002; Chimbutane 1996; Gonçalves 1996). The research agenda on relativization in Portuguese features topics such as non-standard relativization mechanisms, in particular, P-chopping and resumption, and the nature of the relativizer (pronoun or complementizer).

In this paper we aim to investigate variation in the formation of the relative clause of spatial location in three urban African varieties of Portuguese (AVP)—Luanda/Angola (AP), Maputo/Mozambique (MOP), and São Tomé/São Tomé and Príncipe (STP). These three varieties are particularly interesting because they are undergoing a process of nativization in a context of language contact and shift toward Portuguese. Based on spoken, contemporary corpus data, we seek to expand the knowledge on these varieties and to contribute to the understudied topic of locative relativization in Portuguese. We are particularly interested in understanding the distribution of the two main relative morphemes heading locative

relative clauses, *onde* ‘where’ and *que* ‘that’, and whether and how the use of these two forms is constrained.

One hypothesis is that language contact plays a role since this factor has been argued to be the driving force behind linguistic features of AVP. The fact that the three AVP are historically in contact with different language typologies, including in the domain of locative relativization, provides an opportunity to assess the role of language contact from a cross-comparative perspective. In addition, we explore a more language-internal hypothesis, aiming to determine whether there are specific syntactic and/or semantic variables that motivate the selection of one or the other relativizer. We therefore selected and analyzed variables which had either been previously discussed for the domain of relativization or showing potential to correlate with the use of the two different relativizers. Addressing these hypotheses will also give insight into the language variation that characterizes the three varieties at stake, internally and among each other.

The paper is structured as follows. Section 2 briefly reviews work on locative relativization in Portuguese; Section 3 provides a brief background on the AVP in question and lays out the methodology used; Section 4 focuses on the properties of locative relativization in the main contact languages and what they predict with respect to the AVP if contact plays a role; Section 5 discusses, on the one hand, syntactic mechanisms involved in locative relativization, including the distinction between head nouns with an argument and an adjunct status, and, on the other hand, semantic variables that were tested in order to determine whether there is a (statistical) correlation between the semantic nature of the head noun and the selection of the main locative relativizers *onde* and *que*, in particular the semantic role of the head noun, the type of location it describes, and its definiteness; Section 6 discusses the results; and Section 7 concludes the paper.

2. Locative Relativization in Portuguese

Locative relative clauses are a type of relative clause that provide essential information about the location or place associated with a noun in a sentence. There is considerable variation in relativization across different languages (see, for example, Comrie 1981; Vries 2002 on relative clauses in general). For example, the structure and use of relative clauses can vary cross-linguistically in terms of word order and the syntactic mechanisms used; the choice of relativizers, i.e., different languages may have different relative pronouns or other lexical material to introduce relative clauses, such as conjunctions or complementizers and classifiers, the type of agreement between the head noun and the relativizer, and the semantic features of the head noun and the relative clause (e.g., Wiechmann 2015).

Portuguese exhibits postnominal relative clauses that are standardly introduced by invariable morpheme *que* or by morphemes showing Case or number features, such as *quem* ‘whom’ and *qual/quais* ‘which’, as illustrated, respectively, in (1a–c). The latter two examples further show that Portuguese exhibits pied-piping (i.e., a relative pronoun preceded by prepositions) in cases of PP-relativization.

- (1) a. *Perdi o livro que me deste.*
I.lost the book REL me you.gave
‘I lost the book you gave me.’
- b. *Não conhecia a pessoa a quem dei o livro.*
NEG I.knew the person to REL I.gave the book
‘I didn’t know the person to whom I gave the book.’
- c. *Vi o filme do qual toda a gente fala.*
I.saw the movie of.the REL every the people talks
‘I watched the movie everybody talks about.’

Languages often display specialized relative markers that encode location, such as the English *where*. In Portuguese the standard locative relativizer is *onde*, ‘where’, but other forms are also used, namely the invariable morpheme *que*, ‘that’, preceded by a pied-piped preposition, which is usually *em*, ‘in’, as in (2), and by strategies involving the variable relativizer *o(a) qual/os(as) quais* ‘which’, which exhibits gender and number features, illustrated in (3), and whose determiner contracts with prepositions used for locative purposes.

- (2) A casa ondelem que eu moro é bonita.
 the house REL/in REL I live is beautiful
 'The house where/in which I live is beautiful.'
- (3) As esplanadas em que/nas quais gosto de ler um livro
 the terraces in REL/in.the which I.like to read a book
 estão viradas a poente.
 are turned the west
 'The terraces where/in which I like to read a book are facing west.'

Locative relativization in Portuguese has generally only deserved brief mention in the extensive literature on other types and properties of relativization in this language. It is commonly stated that the relative morpheme *onde* has a locative or place feature (Alexandre 2000; Brito 1991; Corrêa 2001; Mória 1992; Veloso 2013) and that the head noun in these constructions does not refer exclusively to a dimension of physical space (Peres and Mória 1995, p. 305), as in (4) below. In Caboverdean Portuguese, a variety not analyzed in this paper, the non-standard use of *onde* in Portuguese can also carry additional (underspecified) features related to tense and event/situation (Alexandre and Lopes 2022), as illustrated in (5) and (6).

- (4) *A parte do discurso onde a Ana foi mais convincente foi aquela em que enumerou as promessas não cumpridas do Governo.*
 (Peres and Mória 1995, p. 305)
 'The part of the speech in which Ana was most convincing was the one in which she listed the government's unfulfilled promises.'
- (5) *"Seria bom que a Expo ficasse na História como um momento onde a cultura e a ciência portuguesas se encontrassem com o futuro", disse.* (Alexandre and Lopes 2022, p. 4)
 "'It would be nice if the Expo was remembered as a moment when Portuguese culture and science met the future", he said.'
- (6) *Nessa altura, os agentes da Judiciária de Tomar são chamados a actuar, num caso onde tinham já algum trabalho em marcha.*
 (Alexandre and Lopes 2022, p. 4)
 'At that point, the officers of the Tomar Judicial Police were called into action, in a case on which they already had some work underway.'

Despite the widespread use of *onde*, work on relativization in Portuguese has shown that there is a tendency to use the relativizer *que*, especially in non-standard mechanisms such as P-chopping and resumption, to the detriment of strategies that privilege relativizers with agreement features, such as *quem* or *qual/quais* (e.g., Elisabeth and Rinke 2017; Alexandre 2000; Alexandre et al. 2011a; Peres and Mória 1995; Tarallo 1985). It has been argued that, due to its lack of phi-features, *que* is being reanalyzed as a complementizer including in EP and AVP (e.g., Alexandre 2000; Alexandre et al. 2011a).¹ While it has been established that *onde* is increasingly used outside of locative relative structures, a fact that has been diachronically reported for other languages (e.g., Ballarè and Inglese 2022)², the role of *que* in locative relativization has not been adequately assessed, i.e., is it the case that the general tendency to favor an invariable relativizer reflects or brings about changes in the domain of locative relativization?

The puzzle about Portuguese locative relative clauses in general is therefore that *onde*, 'where', extends its functions to semantic roles different from [+locative], while *que*, 'that', enters the domain of locative relative clauses, in particular in non-standard relativization strategies. With this study we attempt to shed new light on the distribution of these two relative markers in contemporary, spoken AVP, focusing on the role of putative syntactic and semantic variables and language contact.

3. Background and Methodology

The more widespread use of Portuguese in the former Portuguese colonies in Africa is mostly a 20th century phenomenon, related to the effective occupation, exploitation, and administration of these spaces (e.g., Gonçalves 2010, 2013; Hagemeyer 2016). During the colonial period, it was mainly the language of a (privileged) minority and most typically an L2 for those without roots in the metropole. After the independences in the 1970s, when Portuguese became the exclusive official language of the new countries supported by the democratization of education in this language, increased social mobility, as well as other

aspects of modernization such as exposure to audiovisual means, Portuguese became not only more widespread, but also increasingly nativized in Angola, Mozambique, and São Tomé and Príncipe. The boom of Portuguese is underscored by data from the national censuses that were held after the independences (Hagemeijer 2016, p. 46). According to latest national censuses, Portuguese is spoken by 98.4% of the population of São Tomé and Príncipe (INE 2013); 71.15% of the Angolan population speaks Portuguese at home, which reaches 85% in urban areas (INE 2016); and Portuguese is spoken by 47.4% of the Mozambican population, which includes 16.6% of L1 speakers, again, with strong prevalence in urban areas (77%) (INE 2019). Although the censuses in Angola and São Tomé and Príncipe do not distinguish between L1 and L2 speakers of Portuguese and the other languages, the combined percentages of the number of speakers of the other languages show the role of nativization. In the latest census in São Tomé and Príncipe (INE 2013), for example, only slightly over 50% of the speakers were indicated to be speakers of one of the creole languages (Hagemeijer 2018, p. 178), which also confirms that there is a growing body of Portuguese monolinguals.

Since Portuguese is (historically) an L2, language contact has been argued to play a prominent role in the shaping of AVP, in the sense that features (and lexicon) from the L1 languages have been transferred to L2 varieties of Portuguese (e.g., Gonçalves 2010; Inverno 2011; Mingas 2000). While the data do show (some) evidence in support of language contact, quantitative-based research has shown that AVP exhibit substantial intra and interspeaker variation (e.g., Gonçalves and Chimbutane 2004; Gonçalves et al. 2022), which also includes substantial convergence with EP, the target language, as well as innovative patterns with respect to both EP and the relevant contact languages (e.g., discussion in Gonçalves et al. 2022; Hagemeijer et al. 2022a). The growing number of L2 and especially L1 speakers and the increased role of schooling may actually be taken as forces that increasingly counter the putative role of language contact.

With respect to the data, this case study is based on spoken, urban corpora of AP, MOP, and STP that were prepared within the project *Possession and Location: microvariation in African varieties of Portuguese* (PALMA). The semi-structured interviews that form the corpora were collected in the capitals Maputo, Luanda, and São Tomé between 2008 and 2020, and were as much as possible balanced according to level of education, age, and gender (cf. Hagemeijer et al. 2022b). Portuguese is the L1 or primary language of most of the informants, especially in the case of urban AP and STP, confirming the ongoing tendency toward nativization of Portuguese and its role as a lingua franca. Table 1 summarizes the basic information of the corpora.

Table 1. Profile of the PALMA corpora.

	Interviews	Hours	Tokens	Year of Recording
Angola	58	34	393,745	2012–2013, 2019
Mozambique	70	42	380,958	2010, 2020
São Tomé and Príncipe	77	32	322,999	2008, 2011–2012
Total	205	108	1,097,702	

For the purpose of this paper, we proceeded to extract all the contexts with the relative markers *onde*, ‘where’, *que*, ‘which/that’, and *qual/quais*, ‘which’, from the searchable CQPweb platform (Hardie 2012) hosting the three corpora, and then manually excluded all the contexts which do not concern relative clauses of spatial location, as well as unclear contexts.³ In doing so, we identified a total of 645 relevant contexts: 217 for AP; 271 for MOP; and 157 for STP, as shown in Figure 1 below, which includes the numbers and percentages per variety.

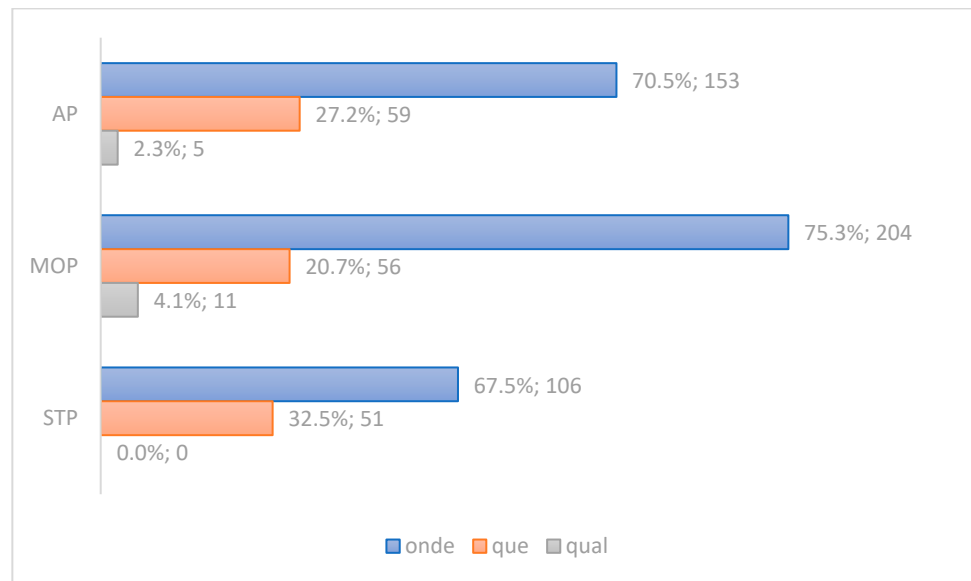


Figure 1. Distribution of relativizers *onde*, *que*, and *qual/quais* in the corpora.

Locative relative clauses formed with *qual/quais* are practically absent (AP and MOP) or inexistent (STP) in our data, which is in line with tendencies observed in other work on relativization in European Portuguese (e.g., Rinke and Aßmann 2017; Alexandre et al. 2011a; Arim et al. 2004; Selas 2014). We therefore excluded a total of 16 occurrences of *qual/quais* from further analysis, focusing exclusively on data involving relativizer *onde*, which is dominant in the three AVP (68–75%), and *que*, which is also common (21–33%). This yields a final total of 629 contexts: 212 for AP; 260 for MOP; and 157 for STP.

We then proceeded to manually annotate the remaining extracted contexts with respect to the follow syntactic and semantic variables⁴, which we briefly address further below:

- i. The syntactic relativization mechanism;
- ii. The argument vs. adjunct status of the antecedent head noun;
- iii. The semantic role of the antecedent head noun (Locative vs. Goal/Source);
- iv. The type of location expressed by the antecedent head noun: physical (container or area) vs. non-physical;
- v. Definiteness of the antecedent: [+definite] or [−definite].

A large amount of crosslinguistic work has been produced on the syntactic mechanisms of relativization (e.g., Comrie 1981). Hence, we aim to discuss to what extent the relativization mechanisms identified in previous work on Portuguese in general and AVP in particular occur in locative relative constructions, namely (i) pied-piping, (ii) P-chopping, (iii) resumption, and (iv) defective copying (Alexandre 2012), how they correlate with relative markers *que* and *onde*, and whether there is evidence of contact-induced effects. In the domain of syntax, we will also assess whether the adjunct or argument status of the head noun leads to any prediction with respect to the use of *onde* and *que*.

We will further analyze the data considering semantic properties of the antecedent of the locative relatives to test whether semantic variables play a role in the selection of relativizers *onde/que*. The semantic features that will be assessed are (i) the semantic role of the head noun, distinguishing between Locatives, on the one hand, and Goal/Source on the other; (ii) the type of location expressed by the antecedent, i.e., physical location or non-physical location and, within the former type, well-delimited head nominals (‘containers’) and those lacking sharp boundaries (‘areas’), a distinction adapted from work by Nikitina (2008) on spatial Goals; and, finally, we investigate whether the definiteness status of the head nominal is a predictor for the use of *onde/que*.

Assuming that *onde* has a basic locative feature (Peres and Mória 1995; Veloso 2013), whereas *que* is unspecified, sharing properties with complementizers (e.g., Alexandre 2000;

Alexandre et al. 2011a; Tarallo 1985), we will discuss whether this difference is reflected in the form of the relativizer that accompanies the head noun. In particular, we will assess whether *onde* is more prone to occur with Locatives, physical locations, and definite locations, and *que* in other contexts. Despite crosslinguistic research on the semantic features of the antecedent and/or the properties of the relative clause (e.g., Wiechmann 2015), these semantic variables have not been explored for locative relativization in Portuguese.

The validity of the syntactic and semantic variables will be subjected to statistical analysis through the online statistical software Jamovi (The Jamovi Project 2023).

We do not distinguish between restrictive (the bulk of our data set) and appositive relative clauses, because we do not consider this distinction critical to our discussion.

Finally, standard European Portuguese (EP), as the target grammar for these varieties, is used for comparative purposes.

4. Language Contact

Although the domain of relativization has not been thoroughly investigated, research focusing on its syntactic properties, in particular Brito (2001, 2002) on genitive relatives in MOP and Alexandre et al. (2011a) on PP relativization in Caboverdean and Santomean Portuguese, has argued against a major role for language contact. For these two types of relativization, this assessment follows from the fact that the corresponding syntactic properties of the main contact languages, which differ typologically from Portuguese, occurred only marginally in these AVP.

Considering the importance that is traditionally assigned to language contact in the literature, this section aims to briefly address the main properties of locative relativization in the contact languages that are typically linked to the urban varieties under discussion, namely the Bantu languages Mbundu⁵ for AP and Changana/Ronga for MOP, and the creole language Santome for STP.

4.1. Mbundu

Many Bantu languages still exhibit the three specific locative markers within the noun class system that have been reconstructed for proto-Bantu (class 16 **pà-*, 17 **kù-*, and 18 **mù-*), each with specific semantic functions (e.g., Zeller, forthcoming). This is the case of Mbundu, where locative relatives are expressed by an initial prefix in the verbal complex that is sensitive to the semantic role of the head noun: with Goal/Source antecedents, prefix *ku-* occurs, whereas Locatives occur with stative *mu-*, as shown in (7) and (8), respectively⁶.

- | | | | | | |
|-----|---|-----------------------|----------------|-----------|-----------------|
| (7) | <i>Inzo</i> | <i>ku-nga-ya</i> | <i>ya-kala</i> | <i>mu</i> | <i>Luwanda.</i> |
| | 5.house | LOC17-1SG.TAM-go | 5.TAM-be | LOC18 | Luanda |
| | 'The house where I go to is in Luanda.' (elicited with Afonso Miguel) | | | | |
| (8) | <i>Inzo</i> | <i>mu-nga-tungila</i> | <i>ya-kala</i> | <i>mu</i> | <i>Luwanda.</i> |
| | 5.house | LOC18-1SG.TAM-live | 5.TAM-be | LOC18 | Luanda |
| | 'The house where I live is in Luanda.' (elicited with Afonso Miguel) | | | | |

Differently from Portuguese, Mbundu thus uses a mechanism that is primarily morphological to yield a locative relative interpretation. Therefore, if language contact is a relevant factor for locative relativization in AP, the prediction is that the semantics of this morphologically encoded Goal/Source vs. Locative contrast could carry over to AP.

4.2. Changana/Ronga

Differently from Mbundu, Bantu languages of the Tsonga cluster, and Changana and Ronga in particular, only exhibit traces of the proto-Bantu tripartite locative noun class system (the morpheme *ka-* in the examples below). The locative relativization strategy in this cluster consists of the use of class-agreeing relativizers identical to demonstratives⁷, the use of verbal prefixes or suffixes that show agreement with tense-marking (*-taka-* and *-nga-* in the examples below), and the use of resumption, which is expressed by the locative + pronoun complex *ka-xone* and *ka-drone* in (9) and (10) (e.g., Macaba 1996; Vondrasek 1999; Chimbutane 2002; Zeller, forthcoming).

- (9) *Axi-tramu* *lexi* *u-taka-trama* *ka-xone*.
 7-chair REL7 2SG-REL.FUT-sit LOC-7PRON
 ‘The chair on which you will sit.’ (lit. ‘The chair that you will sit on it.’)
 (Ronga, adapted from Vondrasek 1999, p. 134)
- (10) *Adoropa* *ledri* *ni-nga-kulela* *ka-drone*.
 5.vila REL5 1SG-REL.PST-crescer LOC.5PRON
 ‘The town in which I grew up.’ (Lit. ‘The town that I grew up in it.’)
 (Ronga, adapted from Vondrasek 1999, p. 134)

Differently from Mbundu, the examples show that Changana/Tsonga exhibits syntactic relativizers, which manifest themselves in different forms according to the noun class the head noun belongs to⁸. (Mozambican) Portuguese of course lacks a noun class system and does not present special relative tenses. The use of resumptive pronouns, on the other hand, has been attested in all varieties of Portuguese and has been discussed for MOP in Chimbutane (1996), who analyzes instances of Direct Object and (mostly non-locative) Oblique relatives in this variety from the perspective of Universal Grammar, without discussing the hypothesis of language contact. If this factor plays a role, we expect it to manifest itself in the form of resumption, which is arguably the only feature of locative relativization in Changana/Ronga available for transfer to MOP.

4.3. Santome

In Santome (Forro, Sãotomense), the main creole language spoken on the island of São Tomé, relative clauses in general are headed by the relativizer *ku*. In locative relatives with locative adjuncts and stative verbs of locative use, such as *vivê* ‘to live’, *ta* ‘to live, to be at’, and *sa* ‘to be (at)’, as shown in (11) and (12), this form can be accompanied by the defective copy *nê*⁹. Defective copying, i.e., an invariable third-person singular form *ê*, which lacks number agreement with the head noun (which is particularly visible with plural head nouns), corresponds to the canonical PP-relativization mechanism in Santome (e.g., Alexandre and Hagemeyer 2002, 2013; Alexandre et al. 2011a). The following examples were extracted from a Santome corpus (Hagemeyer et al. 2014).

- (11) *Kabla sêbê xitu ku kabla ka kume nê*. (Santome)
 goat know place REL goat HAB eat in-3SG
 ‘Goats know where goats eat.’
 (Lit. ‘Goats know the place that goats eat in it’)
- (12) *ke ku êlê tan ku anzu se saka vivê nê*. (Santome)
 house REL 3SG only with child DEM PROG live in-3SG
 ‘the house where only he and his child are living in.’
 (Lit. ‘The house that only he and his child are living in it.’)

Verbs of directed movement, on the other hand, introduce Goal arguments directly, as shown in (13a) by the sequence *ba ke*, ‘go home’. Therefore, relativization of Goal arguments of these verbs shows the same properties as regular Direct Objects (leaving a gap) and does not trigger a defective copy, as shown in (13b). In the specific case of ‘to go’, there are two allomorphs, *ba* and *be*, whose distribution is determined by syntactic properties (Hagemeyer 2004).

- (13) a. *Mina ba ke ka sola potopoto*. (Santome).
 child go home TAM cry IDEOPHONIC
 ‘The child went home crying cats and dogs.’
- b. *kwa ku n mêsê sa sêbê xitu ku bô be* (Santome)
 thing REL 1SG want be know place REL 2SG go
 ‘what I want to know is where you went’

If Santome plays a role in the patterns observed in STP, the prediction is that this variety will exhibit defective copying as a relevant strategy and possibly also a tendency toward the use of *que*, since the Santome counterpart of Portuguese *onde* (*andji* or its short form *an*) is only used in locative interrogatives. Moreover, since directed motion verbs in Santome are transitive, under a contact-induced hypothesis we might also expect cases of (apparent) P-chopping when Goal arguments are transitivized.

4.4. Summary

This brief incursion in the mechanisms of locative relativization in the languages that have been historically in contact with AP, MOP, and STP shows substantial differences among them, not only between the Bantu languages and Santome, whose typologies are considerably different, but also among a western Bantu language (Mbundu) and southeastern Changana/Ronga. Some of the features of locative relatives in these contact languages, such as special tenses or specific locative morphology, are of course not available in Portuguese and therefore not good candidates for transfer; but others, especially those in the syntactic domain, such as resumption and defective copying or the use of an exclusive relative marker, are within reach of the typology of Portuguese and therefore potential candidates for transfer. Since the contact languages display different mechanisms and features in locative relativization, the role of language contact, if relevant, would tendentially lead to different outcomes in each of the AVP in this study.

5. Syntactic and Semantic Variables in AVP Data

In the subsections below, we discuss the syntactic mechanisms of relativization in the AVP, the syntactic relation between the head noun and the relative clause (argument or adjunct), as well as several semantic features, namely the semantic role of the head noun (Locative, Goal, Source) with respect to the predication, the type of location expressed by the head noun (physical, i.e., area or container, or non-physical), and definiteness of the head noun. The distribution of these variables is crossed with the use of *onde/que* and tested for statistical significance.

5.1. Syntactic Mechanisms of Locative Relativization

Pied-piping in locative relatives in AVP involves the relativizers *onde* and *que* (14a,b), which can be preceded by a preposition that is pied-piped from inside the relative clause. Most commonly, this preposition is *em*, ‘in’ (14b), in which case it is accompanied by relative *que*, but other combinations also occur, for example *a*, ‘to(wards)’, and *de*, ‘from’, for Goals and Sources or *por*, ‘by, through’ (e.g., *por onde*). We treat locative relative clauses introduced by *onde* as instances of intrinsic pied-piping.

- (14) a. *fui para uma oficina de marcenaria onde estive durante sete anos* (STP)
 I.went to a woodwork studio REL I.was for seven years
 ‘I went to a woodwork studio where I stayed for seven years’
- b. *é difícil arranjar emprego no país em que nós nos encontramos*
 it.is hard to.find job in.the country in REL we REFL are (AP)
 ‘it is hard to find a job in the country where we are’

In the cases of locative relatives, P-chopping, i.e., the deletion of argumental and non-argumental prepositions, primarily involves the deletion of the locative preposition *em*, ‘in’, with relativizer *que*, as illustrated in (15). However, the absence of other prepositions is also attested in the data, as shown in (16), in particular with *que* (16a,b), but also with *onde* (16c,d)¹⁰.

- (15) a. *e é uma província que fala-se muito quimbundo* (AP)
 and it.is a province REL speak.IMP a lot of Mbundu
 ‘and it is a province where they speak a lot of Mbundu’
- b. *é como São Tomé é um país que chove muito* (STP)
 and since São Tomé is a country REL rains a lot
 ‘and since São Tomé is a country where it rains a lot’
- c. *há sítios que não havia escolas* (MOP)
 there.are places REL NEG were schools
 ‘there are places that didn’t have schools’

- (16) a. *há sítios que carro mesmo não vais* (AP)
 there.are places REL car even NEG you.go
 ‘there are places where you don’t even go by car’
- b. *uma região ali do norte também que eu nunca fui.* (STP)
 a region there of.the north also REL I never went
 ‘a region in the north where I never went to.’
- c. *Para os ovimbundos onde eu sou originário* (AP)
 to the Ovimbundu people Ø REL I am originating
 ‘To the Ovimbundu people where I belong to.’
- d. *um serviço alargado de agenda pública* Ø *onde os ouvintes trazem*
 a service extended of agenda public REL the listeners bring
os seus avisos ou anúncios (STP)
 the their notifications or announcements
 ‘an extended service of the public agenda to which the listeners bring their notifications or announcements’

Locative resumption entails the semantic recovery of the head noun by a deictic, in particular *lá*, ‘there’, as in (17a–c) but less commonly also a (repeated) full-fledged NP, as in (17d). Resumption is observed with both the relative marker *que* and *onde*, which also shows that different mechanisms, such as pied-piping and resumption (17c) or chopping and resumption (17d), may cooccur¹¹.

- (17) a. *volto para minha província que já não vivo lá há quinze anos* (AP)
 I.return to my province REL already NEG I.live there for fifteen years
 ‘I return to my province where I haven’t lived for fifteen years’
- b. *assemelha-se a uma indústria que eu estava lá* (MOP)
 it.resembles-REFL to a industry REL I was there
 ‘it resembles an industry where I used to work’
- c. *uma zona famosa onde dizem que tem lá algumas pessoas* (STP)
 a zone famous REL they.say that has there some people
 ‘a famous region where they say some people live’
- d. *é um lugar que hoje ou amanhã pode sair daquele lugar* (AP)
 it.is a place REL today or tomorrow can leave from.that place
 ‘it’s a place which he may leave today or tomorrow’

Finally, locative defective copying consists of a stranded preposition that is accompanied by a third-person singular pronoun which lacks gender and number agreement features with the head noun, as illustrated in (18).

- (18) *própria escola que estudei nele* (STP)
 self school REL I.studied in-3SG
 ‘the very school I studied in’ (lit. the very school I studied in it)

Figure 2 below sums up the results based on the extracted corpus data with the relativization mechanism and the relative markers *onde* and *que*.

The first conclusion that can be drawn from the figure is that, in all three AVP, pied-piping is preferred (AP = 65.1%, MOP = 71.5%, STP = 63.3%), being followed by chopping, a mechanism that mainly affects *que* locative relatives (AP = 13.2%, MOP = 14.2%, STP = 24.2%). Percentage wise, the ratio of use of these mechanisms in the AVP shows similarities, although P-chopping is more prominent in STP. Chopping with *onde*, on the other hand, is almost inexistent in STP and more prominent in AP. Figure 2 further shows that locative resumption and especially locative defective copying, which is exclusive to STP, are residual mechanisms in our data. Nevertheless, it should be noted that locative resumption is overall more common with *onde*. Moreover, the fact that these two residual mechanisms are the canonical processes in the main contact languages Changana/Ronga and Santome, respectively (cf. Sections 4.2 and 4.3 above), constitutes a strong argument against the role of language contact.

In order to assess whether there is a correlation between relativizer *onde/que* and the relativization mechanisms, as suggested by the data and the results of previous research, which show a correlation between non-standard mechanisms and the use of *que*, and given the characteristics of our data, in particular the limited occurrence of resumption and

defective copying, we applied a Fisher’s exact test. Taking relativization mechanisms as independent variables and relative markers *onde* and *que* as dependent variables, the results indicate that the correlation is indeed highly significant (p -value ≤ 0.001).¹² We further found that the correlation between relativizers *onde/que* and the relativization mechanism is also highly significant in each variety (p -value ≤ 0.001).

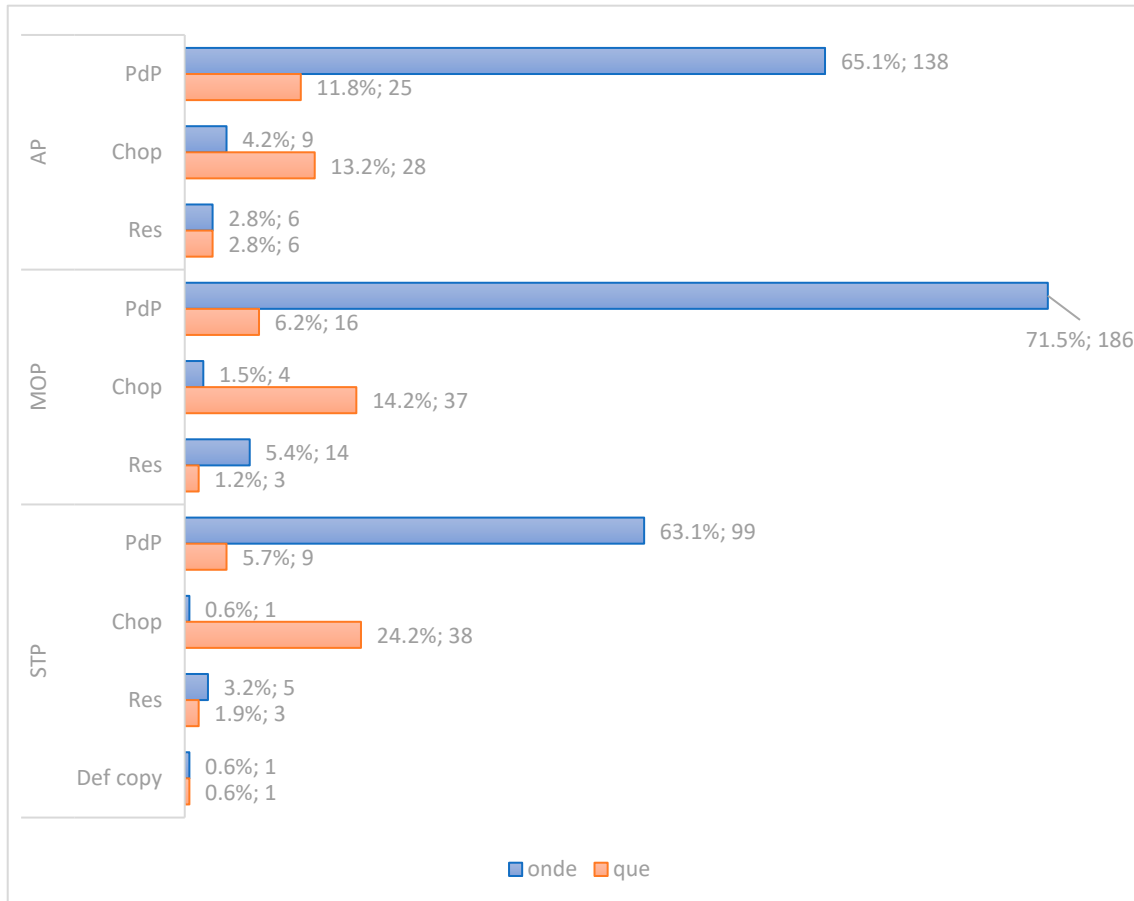


Figure 2. Distribution of relativizers *onde/que* by relativization mechanisms and AVP.

5.2. Syntactic Relation between the Head Noun and the Relative Clause

Figure 3 shows the distribution of relative markers *onde/que* and the syntactic relation between the antecedent and the relative clause, i.e., whether the antecedent is an argument or an adjunct within the relative clause. We base the distinction between argument and adjunct on standard assumptions about argument structure. Locative objects of motion verbs such as *ir*, ‘to go (to)’, *vir*, ‘to come (to, from)’, *chegar*, ‘to arrive (at)’ or *entrar*, ‘to enter’, as well as the object of verbs such as *viver* and *morar*, ‘to live (in)’, were treated as arguments.

The figure shows that the use of the relativizer *onde* in the three AVP is largely preferred in locative relatives, independently of the syntactic relation between the head noun and the relative clause, i.e., irrespective of whether the antecedent is an argument, as in (14a), or an adjunct within the relative clause, as in (17c). With respect to the argument/adjunct distinction and the selection of relativizer *que*, AP and MOP show a proportionately higher number of contexts of locative head nouns that correspond to an adjunct in the relative clause (AP = 9.0% arguments vs. 18.9% adjuncts; MOP = 8.8% arguments vs. 12.7% adjuncts); in STP this proportion is more balanced and slightly favors arguments (17.8% arguments vs. 14.6% adjuncts).

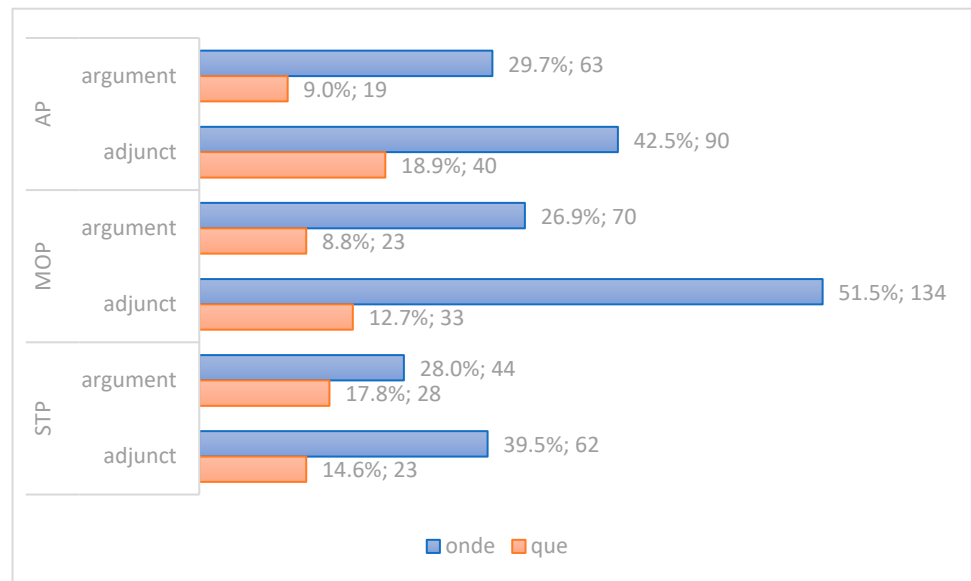


Figure 3. Distribution of relativizers *onde/que* by arguments/adjuncts and AVP.

In order to confirm the correlation between relativizers and the syntactic relation between the antecedent and the relative clause, we applied a Chi-square (χ^2) test, taking argument and adjunct as independent variables and relativizers *onde* and *que* as dependent variables. Results indicate that differences observed in the production of those mechanisms are not significant (p -value = 0.373) and the correlation is, in fact, not significant in each AVP independently (AP, p -value = 0.229; MOP, p -value = 0.350; STP, p -value = 0.115).

In sum, the two sections on syntactic properties of locative relatives show that the three AVP do not differ substantially from one another in what concerns the relativization mechanisms and the distinction between argument and adjunct. The data further confirm the claim that P-chopping correlates with *que* in varieties of Portuguese. Locative resumption, however, is overall more common with *onde* in our data, but this finding requires further research, since it is based on a rather small number of occurrences.

5.3. Semantic Role of the Antecedent

Regarding the semantic role of the antecedent, a distinction was made between Locative in (19), which corresponds to static locations, and Goal/Source, which corresponds to locations introduced by verbs of movement, as shown, respectively, in (20) and (21). By considering this distinction, we aim to assess whether the selection of the relativizers *onde/que* correlates with the semantic role of the antecedent.

- (19) a. *frequentei uma escola do meu bairro onde eu cresci* (MOP)
 I.attended a school of.the my neighborhood REL I grew.up
 'I attended a school in the neighborhood I grew up in'
- b. *é uma zona que há sempre conflitos* (STP)
 it.is a region REL there.are always conflicts
 'It's a region where there are always conflicts'
- (20) a. *nos musseques onde eu vivo* (AP)
 in.the slums REL I live
 'in the slums where I live'
- b. *diria que a cidade que eu provavelmente me mudaria...* (MOP)
 I.would.say that the city REL I probably REFL I.would.move.to
 'I'd say that the city I'd probably move to...'

- (21) a. *vivi na província donde veio um padrasto meu que me criou* (AP)
 I.lived in.the province from.REL came a stepfather my REL me raised
 'I lived in a province from which came a stepfather of mine who raised me'
- b. *há palmeira que pode sair cinco filhos* (STP)
 there.is palm tree REL can go.out five offshoots
 'There are palm trees that may develop five offshoots.'

Figure 4 below sums up the results, showing that Locative antecedents by far outnumber Goal/Source antecedents and that the use of these semantic roles with *que* and *onde* is fairly proportional across the three AVP. Moreover, the expression of Locative is highly preferred with *onde* across varieties. The main difference between the varieties concerns the use of *que* vs. *onde* with Goal/Source antecedents. Here, AP and STP are at opposite ends: AP shows a strong preference for *onde* and STP for *que*, with MOP in between (AP = 8.0% vs. 1.9%; MOP = 3.1% vs. 2.3%; STP = 1.9% vs. 7.0, for *onde* and *que*, respectively). As in the case of locative resumption in the previous section, this finding requires confirmation based on a larger data set.

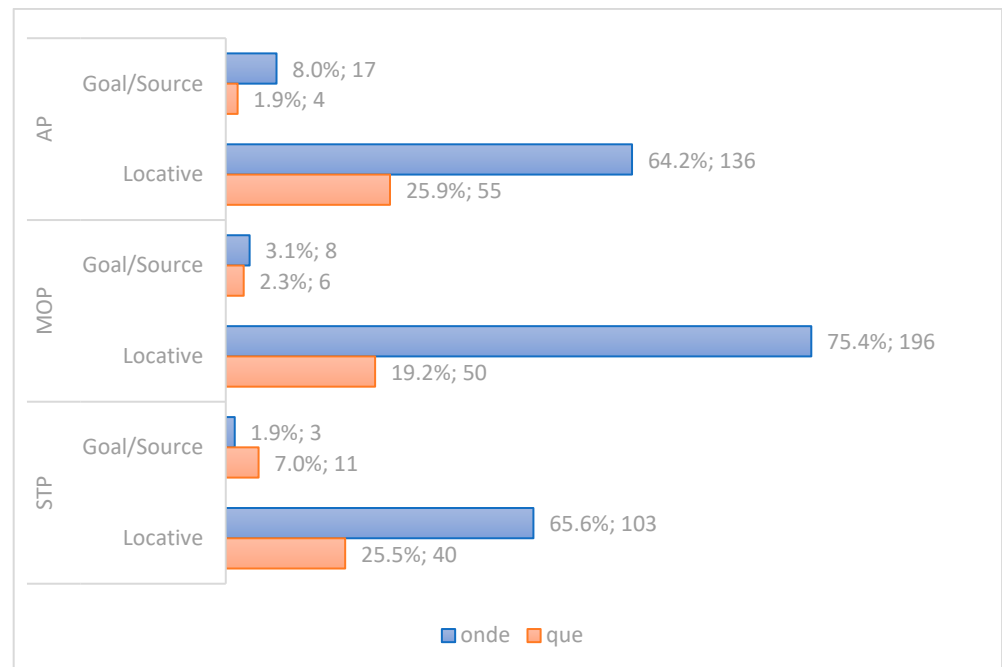


Figure 4. Distribution of relativizers *onde/que* by semantic role and AVP.

Results of a Chi-square test, taking Goal/Source and Locative as independent variables and relativizers *onde* and *que* as dependent variables, indicate that the general correlation between these variables is significant (p -value = 0.006). However, the correlation between variables differs among the AVP. The correlation between the relative markers *onde/que* and the semantic role of the antecedent is significant in STP (p -value \leq 0.001, result of a Chi-square test) but not in AP (p -value = 0.344) and MOP (p -value = 0.086) (results of a Fisher’s exact test).

Additionally, the Fisher’s exact test established a correlation between the relativization mechanism and the semantic role of the antecedent, since pied-piping is mostly registered with Locatives (<0.001). However, this correlation is significant in STP (p -value \leq 0.001) and MOP (p -value \leq 0.001), but not in (AP, p -value = 0.140).

At this point we should recall that it was shown that Goal/Sources and Locatives in Mbundu occur with different noun class prefixes, that is, *ku-* for the former and *mu-* for the later (cf. Section 4.1). If language contact were playing a role in locative relativization, we might expect AP to exhibit a tendency toward the use of different relativizers for these different semantic roles. However, the more widespread use of *onde* in AP with both

Goals/Sources and Locatives than in MOP and STP shows that this prediction was not borne out. On the other hand, the more extensive overall use of *que* in STP than in AP and MOP, especially with Goals, is potentially related to the use of the exclusive etymologically related relativizer *ku* and the direct transitivity of Goal-selecting verbs in Santome (cf. Section 4.3), thus arguably showing a mild contact-induced effect.

5.4. Type of Location Expressed by the Antecedent

The second semantic variable that was tested for the *onde/que* distinction consists of the type of location expressed by the antecedent. Here, we adapted the typology proposed by Nikitina (2008) in her work on spatial Goals with motion event (‘in’ vs. ‘into’) to spatial locative relatives. This work was previously used in research on Goal arguments of verbs of directed motion in AVP (Hagemeyer et al. 2022a), with interesting results with respect to the use of the type of location selected by the (non-standard) preposition *em* with two verbs of directed motion.¹³ With respect to our data, a primary distinction was made between head nouns corresponding to physical and non-physical locations, with the former type being further divided into two categories: containers and areas¹⁴. Following Nikitina’s work, containers correspond to locations with well-defined borders lacking a transitional zone (e.g., *house, store, lunch box, sea*)¹⁵, whereas areas are locations without well-defined borders and therefore typically exhibit a transitional zone (e.g., *neighborhood, field, countries, market*). The following examples illustrate the classification concerning the nature of the antecedent: (22) showcases non-physical locations, (23) containers, and (24) areas.

- (22) a. *há um outro fenómeno aí que as pessoas trocam de*
 there.is a other phenomenon there REL the people change of
salário durante o mês (AP)
 salary during the month
 ‘There is another phenomenon there in which people change salaries during the month’
- b. *num ambiente característico de dificuldades onde havia falta*
 in.a environment characteristic of difficulties REL there.was lack
de quase tudo (MOP)
 of almost everything
 ‘In a typical environment of difficulties where almost everything was lacking’
- c. *Há muitas histórias tradicionais onde o tartaruga aparece (STP)*
 there.are many stories traditional REL the turtle shows.up
 ‘there are traditional stories in which the turtle shows.up’
- (23) a. *há sempre uma capoeira onde se cria galinhas (AP)*
 there.is always a hen house REL IMP breed chicken
 ‘there is always a hen house where chicken are bred’
- b. *estou numa loja onde vendo telefones (MOP)*
 I.am at.a store where I.sell phones
 ‘I’m at a store where I sell phones.’
- c. *dirigindo um centro onde trabalhávamos cerca de trinta mulheres (STP)*
 managing a center REL we.worked around of thirty women
 ‘managing a center where around thirty women used to work’
- (24) a. *estamos num sítio em que entrasse um mais velho ... (AP)*
 we.are in.a place in REL enter.SUBJ a more old
 ‘we are at a place where if an elderly person would enter. ...’
- b. *o tal sítio onde o meu filho escreveu (MOP)*
 the DEM place REL the my son wrote
 ‘that place where my son wrote’
- c. *eu estou mais habituado ao bairro onde eu nasci (STP)*
 I am more used in.the neighborhood REL I was.born
 ‘I’m more used to the neighborhood in which I was born’

Figure 5 sums up the results for this semantic variable.

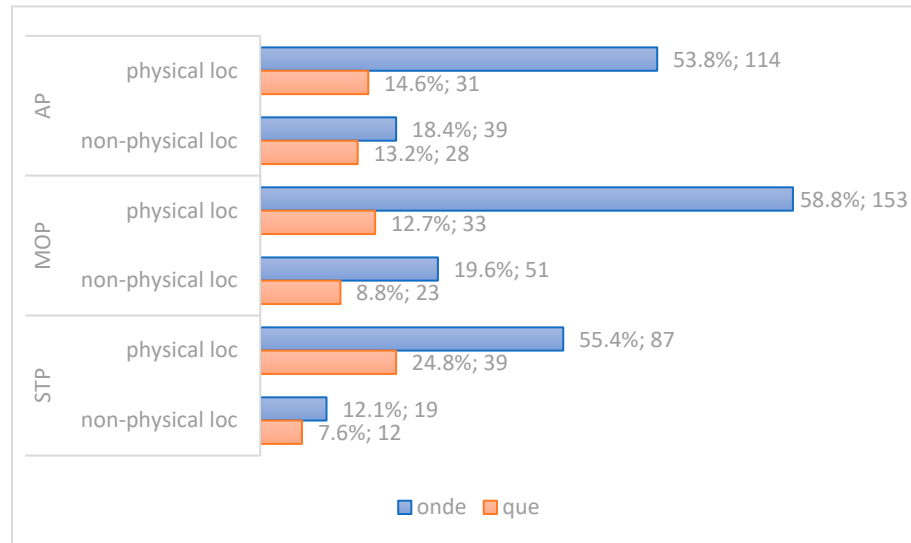


Figure 5. Distribution of relativizers *onde/que* by physical and non-physical location and AVP.

The figure shows that *onde* outnumbers *que* with both physical and non-physical locations in the three AVP and that the contexts with *que* in all three varieties are proportionally greater with non-physical locations than with physical locations, which is especially visible in AP (13.2% vs. 18.4%). It also follows that *onde* is commonly used with non-physical locations, similarly to what has been observed for EP (e.g., Mória 1992; Peres and Mória 1995), a finding we do not explore in this paper.

Zooming in on the distribution of relativizers according to the categories containers, areas, and non-physical, we observe several fine-grained differences between the varieties, which are represented in Figure 6 below.

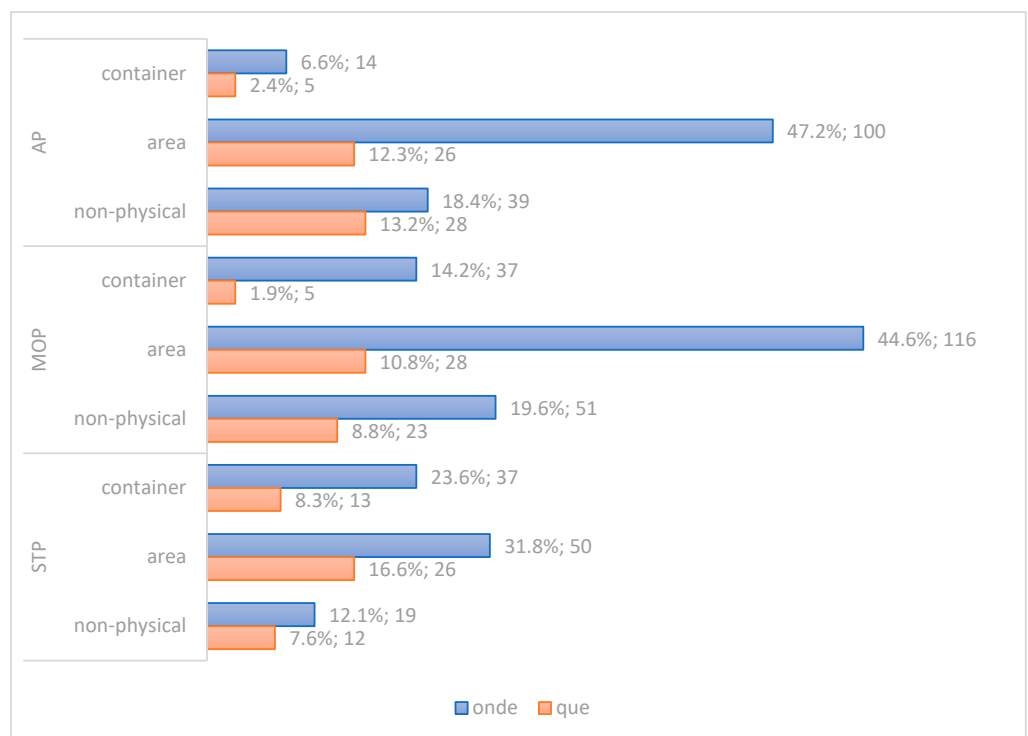


Figure 6. Distribution of relativizers *onde/que* by containers, areas, and non-physical locations and AVP.

STP shows a proportionally greater number of cases in which antecedents that are areas cooccur with *que* instead of *onde* (16.6% vs. 31.8%), whereas this proportion is quite a bit lower in AP (12.3% vs. 47.2%) and MOP (10.8% vs. 44.6%). With containers, on the other hand, MOP stands out for the proportionally lesser use of *que* than *onde* (1.9% vs. 14.2%) than in AP (2.4% vs. 6.6%) and STP (8.3% vs. 23.6%).

Results of a Chi-square test, taking physical locations and non-physical locations as independent variables and relativizers *onde* and *que* as dependent variables, indicate that there is a significant correlation between these variables (p -value ≤ 0.001). Note, however, that the significance is not observed in each AVP. The correlation is significant in MOP (p -value = 0.018) and AP (p -value = 0.002), but not in STP (p -value = 0.409). An additional Chi-square test using areas, containers, and non-physical locations as independent variables confirms the correlation between relative marker and type of location in MOP (p -value = 0.036) and AP (p -value = 0.008), but not in STP (p -value = 0.447). The former varieties clearly prefer *onde* with areas, a tendency that is not observed in STP.

5.5. Definiteness of the Antecedent

Finally, we investigated whether the definiteness status of the head nominal is a predictor for the use of *onde* and *que*. Following Wiechmann (2015, p. 88), “[a] head was treated as definite if the entity referred to was specific and identifiable in a given context of utterance.” In most cases, the locative heads acquire this feature through definite articles and/or the presence of demonstrative and possessive pronouns. We did not analyze the variable ‘specificity’, which accompanies definiteness (e.g., Lyons 1999), because a relative clause by itself is a specification of the head noun. Some examples of [+definite] and [–definite] antecedents are illustrated in (25) and (26), respectively.

- (25) a. *lá no escritório onde faço o part-time* (MOP)
 there in.the office REL I.do the part-time
 ‘there in the office where I have a part-time job’
- b. *acho que deve ser essa escola em que estamos agora* (STP)
 I.think that it.should be that school in REL we.are now
 ‘I think it should be this school in which we are right now’
- (26) a. *o namoro é uma fase em que duas pessoas do*
 the dating is a stage in REL two person of.the
sexo opostos vão conhecendo-se (AP)
 sex opposite go.3PL knowing-REFL
 ‘dating is a stage in which two people of the opposite sex are getting to know each other’
- b. *é um país livre onde as pessoas conseguem fazer tudo que*
 it.is a country free REL the people can do all that
 ‘It’s a free country where people can do whatever they need to.’
eles precisa (MOP)
 they need

Figure 7 below shows the overall results for the three AVP.

The figure shows that [+definite] contexts largely outnumber the [–definite] contexts in all three AVP. Moreover, the relativizer *onde* is strongly preferred in [+definite] contexts. On the other hand, *que* correlates more strongly with [–definite] than with [+definite] in AP and STP (AP = 17.9% vs. 9.9%; STP = 20.4% vs. 12.1%), with MOP showing the weakest correlation (10.4% vs. 11.2%).

Results of a Chi-square test, taking [+definite] and [–definite] as independent variables and relative markers *onde* and *que* as dependent variables indicate that there is a significant correlation between these variables, which is observed in each AVP (AP p -value ≤ 0.001 ; MOP p -value = 0.008; STP p -value ≤ 0.001).

In order to further test whether less specified head nouns correlate with *que*, we additionally analyzed the occurrence of singular and plural bare noun locative antecedents with *que* and *onde*. The following examples illustrate bare nouns with *que*, which occur commonly in existential constructions with *haver*, ‘to be’.

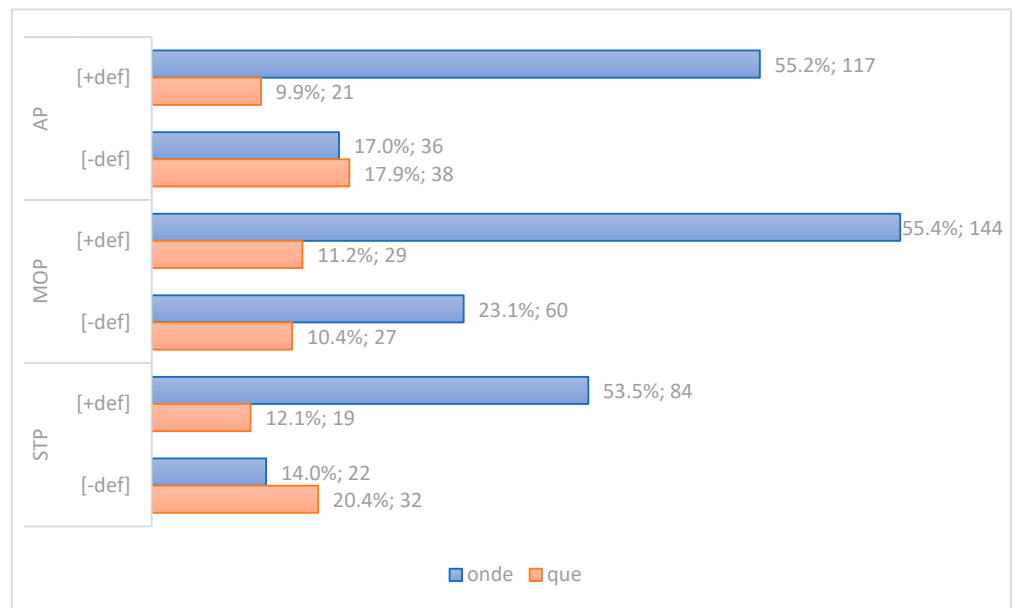


Figure 7. Distribution of relativizers *onde/que* by definiteness of the antecedent and AVP.

- (27) *porque há áreas que pedem um preço elevado* (AP)
 because there.are zone REL they.ask a price high
 ‘because there are zones where they ask a high price’
- (28) *há sítios que nem dá para falar changana* (MOP)
 there.are place REL not.even is.possible to speak Changana
 ‘there are places where it is not even possible to speak Changana’
- (29) *mas há casa ainda que entra água* (STP)
 but there.are house still REL enters water
 ‘but there are also house where the water enters’

If our hypothesis is on the right track, the prediction is that bare nouns—the least specified head nouns—should cooccur more commonly with *que* than *onde*. Figure 8 shows the overall results for the three AVP.

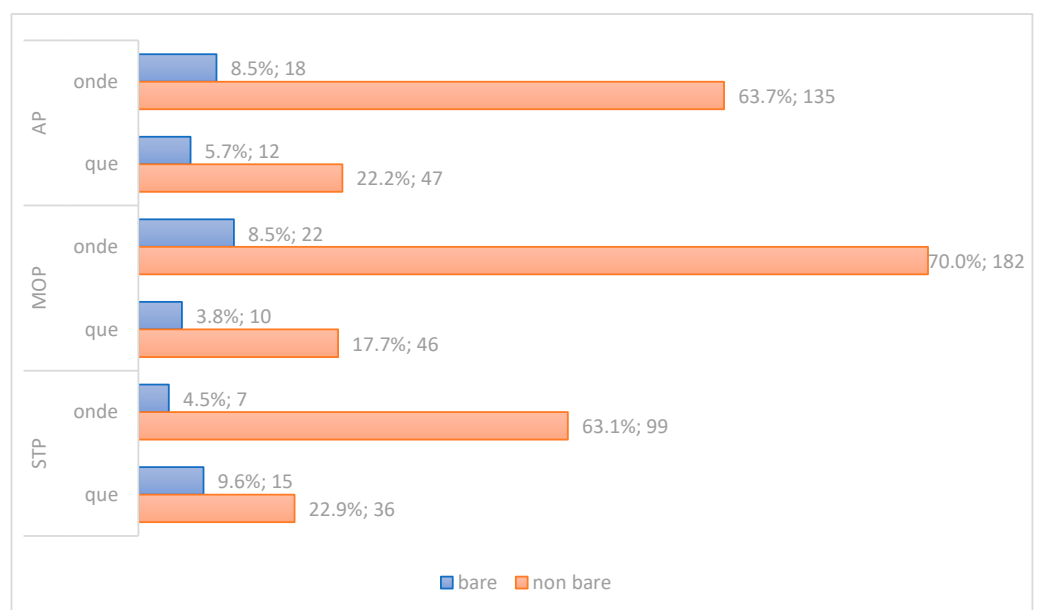


Figure 8. Distribution of relativizers *onde/que* with bare/non-bare NP antecedents and by AVP.

While locative bare nouns do occur with both *que* and *onde*, they are proportionally more common with the former relativizer. The main difference is observed in STP, which exhibits a higher number of bare noun locative antecedents with *que* than with *onde*, whereas the reverse situation applies to AP and MOP (STP: 9.6% vs. 4.5%, AP: 5.7% vs. 8.5%; MOP: 3.8% vs. 8.5%). These two findings underscore the hypothesis that less specified locative antecedents increase the likelihood that *que* is selected, and that this tendency is particularly relevant in STP. In fact, results of a Chi-square test, taking [bare] and [non-bare] as independent variables and relative markers *que* and *onde* as dependent variables indicate that there is a significant correlation between these variables (p -value ≤ 0.001). However, the correlation is only significant in STP (p -value ≤ 0.001). No correlation was found in MOP (p -value = 0.154) and AP (p -value = 0.108).

5.6. Summary

The previous sections have shown that, generically speaking, the three AVP at stake do not differ substantially from each other with respect to the syntactic and semantic variables under discussion, although some tendencies are worth emphasizing.

For the syntactic variables, the data convincingly show that the preferred relativization mechanism is (standard) pied-piping, which is mainly observed with *onde*, followed by P-chopping in contexts with the relative *que*, particularly in STP. With respect to syntactic relation of the antecedent, no correlation was found between the distinction argument/adjunct and the relativizers *que/onde*.

Regarding the semantic variables, the data lead to three main observations. In the first place, locative relatives, whose antecedent head noun has the semantic role of Locative, mainly occur with *onde* across varieties; when Goals/Sources are involved, AP prefers *onde* whereas STP prefers *que*. The second observation concerns the type of location. Considering the distinction between physical and non-physical locations, AP and MOP behave alike in their preference for areas with *onde*; STP, on the other hand, exhibits a higher variation between the two relativizers with head nouns that are areas. Third, the data clearly demonstrate a correlation between the definiteness of the antecedent and the relativizer in the three AVP: *onde* is strongly preferred with [+definite] antecedents, whereas *que* correlates with [−definite], a correlation that is particularly strong in STP, where it is also observed with bare head nouns.

In the next section, we discuss the results in more detail, and consider the extent to which our working hypotheses have been confirmed.

6. Discussion

The analysis of the syntactic and semantic variables in the previous sections has shown that the three AVP show considerable uniformity, differing among each other with respect to a few smaller details. Given the differentiated typological relativization properties of their main (historical) contact languages, the role of language contact, often considered a major factor with respect to the linguistic patterns found in these varieties, is not supported. First, the primary syntactic mechanisms in Changana/Ronga (resumption) and Santome (defective copying) are the exception rather than the rule in MOP and STP. In these varieties, as well as in AP, pied-piping, followed by P-chopping, are by and large the most common mechanisms. Second, the (morphological) distinction between Goal/Source and Locative relativization characterizing Mbundu lacks a counterpart in AP, which rather shows a tendency to generalize the relativizer *onde*, also for Goal/Source, differently from MOP and STP. Third, the [\pm definite] status of the antecedent head noun does not lead to different strategies in the contact languages, where head nouns require noun class morphology (Mbundu and Changana/Ronga) or exhibit an invariable relativizer (Santome). On the other hand, the more extensive use of relativizer *que* in STP, as compared to AP and MOP, appears to reveal a mild effect of language contact, possibly due to the etymological overlap with the generalized relativizer *ku* in Santome. In addition, the greater number of cases of P-chopping of the preposition *a*, ‘to’, with head nouns bearing the semantic role of Goal

in STP, may also show the effect of language contact, since verbs of directed motion in Santome are transitive (V + Goal), as shown in Section 4.3. Note, however, that these cases are restricted to relative clauses (and more generally to cases of fronted locatives), since Goal arguments of directed motion verbs in STP, in particularly the most common verb *ir*, ‘to go’, are typically introduced by a preposition (*para*, *a*, *em*) when they occur in the canonical object position (Hagemeyer et al. 2022a). In other words, it cannot be argued that P-chopping in STP relatives is a consequence of a general change to the subcategorization properties of directed motion verbs.

With respect to the syntactic mechanisms, we have shown that the three AVP follow standard European Portuguese in generally using pied-piping (mainly through *onde*). While this finding also applies to STP, in this variety *que* shows a proportionally higher use than in AP and MOP. Regarding the frequency of P-chopping (and resumption), the findings from the AVP are in line with what has been observed in the literature for EP and BP (e.g., Alexandre 2000; Espírito Santo et al. 2023; Selas 2014; Veloso 2013 for EP; and Corrêa 2001; Kato 2010; Ribeiro 2009; Tarallo 1985 for BP), including the loss of agreeing or Case-specified relative markers and the rise of *que*, associated to non-standard strategies, in particular P-chopping.¹⁶ Locative defective copying is the exclusive territory of STP. As expected, the correlation between relativizers *onde/que* and relativization mechanisms are statistically significant, since *onde* is mainly found with pied-piping and *que* with P-chopping. A couple of fine-grained differences arise with respect to the use of *que* and the distribution of contexts of P-chopping with respect to arguments and adjuncts, but they were not statistically significant. Some of the more specific results are, on the one hand, the greater number of relatives with *que* and chopping of the preposition *a*, ‘to’, and *em*, ‘in’, introducing arguments in STP, which we addressed above, and, on the other hand, the higher frequency of P-chopping with *em* over pied-piping of this preposition in MOP and STP, with AP showing a more balanced distribution of these two mechanisms.

We further found that the distinction Goal/Source vs. Locative is proportionally balanced across the three AVP with respect to the occurrence of *onde* and *que*. In light of the [+locative] feature of this relative marker and the absence of this feature from *que*, this result was expected. The correlation between relativizer and semantic role of the antecedent was found to be statistically significant in MOP and STP, but not in AP, where *onde* is more extensively used with Goal/Source as well.

In what concerns the type of location, the data point to a significant correlation between the choice of relativizer and physical locations, specifically areas, which is observed exclusively in AP and MOP. The absence of this correlation in STP can be accounted for by the more balanced use of *que* with areas and containers in this variety than in AP and MOP. Moreover, and similarly to other varieties of Portuguese, *onde* is also commonly used with non-physical locations.

Finally, the most relevant and novel finding is that of a significant correlation between *onde* and [+definite] head nouns and between *que* and [−definite] head nouns. To account for this finding, we propose that (overtly) less specified antecedents show a greater preference for relativizer *que*. In the literature on relativization in Portuguese, this form has been argued to lack feature specification as compared to other relativizers, such as *quem*, *cujo/a(s)*, *qual/quais*, and *onde*, which has led several scholars to treat it as a complementizer (e.g., Alexandre 2000; Tarallo 1985)¹⁷. Although these studies do not focus on locative relativization, the idea of relativizers with a different featural makeup can be extended to the contrast between *onde*, which arguably bears Oblique Case (Alexandre 2000) and a locative feature (Peres and Mória 1995; Veloso 2013), and *que*, which would of course lack such feature specification.

The hypothesis that that less functionally specified head nouns select the least specified relativizer requires further investigation, in order to determine whether this is yet another property that ultimately leads back to restructuring in the functional domain. It is well established that some of the most functional features of EP have undergone significant restructuring in AVP, and also in BP, which encompasses, for instance, the following

tendencies: the loss of accusative clitics (e.g., Gonçalves et al. 2023); the partial replacement of locative preposition *a*, ‘to’ (Hagemeijer et al. 2022a); and a reduction in overt number agreement (Brandão 2011; Jon-And 2011). In this sense, the definiteness effect observed in locative relativization of AVP can be seen as a case of language-internal restructuring based on an agreement relation which involves the degree of feature specification and the available relativizers in Portuguese: each of the two relative markers specializes, to some extent, for a different semantic environment. Of course, as in many other domains of the grammar of AVP, these are tendencies within the large spectrum of variation which are unlikely to fully crystallize.

7. Final Remarks

The empirical data discussed in this paper resulted in a survey of locative relativization, a topic that has hardly been explored for Portuguese. We provided a first, corpus-based description and analysis of syntactic and semantic properties of spatial locative relativization in the nativizing, urban varieties of Portuguese spoken in Angola, Mozambique, and São Tomé and Príncipe. In addition, we developed a statistically supported discussion of the syntactic and semantic features of the head noun and the relative clause in order to determine what, if anything, drives the selection of the relativizer *que* or *onde*.

Overall, our analysis shows that the three AVP are considerably homogeneous in the domain of locative relativization: (i) the three varieties mainly exhibit *pied-piping* as a primary mechanism and P-chopping as a secondary mechanism, independently of the distinction between argument and adjunct; (ii) *onde* correlates with Locatives and areas, although AP also extends this relative marker to Goals/Sources, and STP exhibits a higher degree of variation between *onde* and *que* with areas and containers (physical locations); and (iii) [–definite] and generally less specified antecedents show a greater preference for relativizer *que* in all three AVP, a preference that is especially pronounced in STP. Interestingly, while *que* is gaining space in the domain of locative relativization, we also notice that *onde* is extending its functions to other semantic roles (cf. Ballarè and Inglese 2022) beyond physical locations, showing the dynamics of relativization and locative relativization in particular.

The results of our research also led us to argue that the role of language contact, which is often considered a driving factor in studies on AVP, should be largely dismissed, since the mechanisms of relativization that characterize the main contact languages are not only distinct from the AVP but also from each other. It also shows that language contact may be more or less prominent, or more or less direct, according to the (sub)domain of grammar under analysis. Using the same corpora, Hagemeijer et al. (2022a) also downplay the role of language contact in the domain of PP selection of directed motion verbs, although in other studies based on the same data, Gonçalves et al. (2022, 2023) show that language contact plays a role in the expression of dative objects and anaphoric direct objects (clitics and pronouns).

The larger spectrum of variation that typically characterizes AVP as compared to EP, and arguably—but perhaps less clearly—also BP, can be assigned to the fact that Portuguese in Africa was until recently, or still is, an L2 in complex multilingual settings. Nevertheless, several quantitative studies in the domain of syntax and morphosyntax show that, despite robust variation, the dominant patterns within the variation of certain features in the AVP are often also the ones that are also dominant in EP, which ultimately implies that variation does not necessarily lead to a new outcome, i.e., the consummation of change.

Finally, with the working hypotheses on the table, the findings of this paper can be further investigated for these and other varieties of Portuguese, especially EP and BP. Moreover, some of the findings require a larger data set or, for example, the application of elicitation tasks, in order to further test and confirm the observed tendencies.

Author Contributions: Conceptualization: T.H., R.G. and N.A.; Data curation: T.H., R.G. and N.A.; Formal analysis: T.H., R.G. and N.A.; Investigation: T.H., R.G. and N.A.; Methodology: T.H., R.G. and N.A.; Visualization: T.H., R.G. and N.A.; Writing—original draft: T.H.; Writing—review & editing: T.H., R.G. and N.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Portuguese Foundation for Science and Technology (FCT), grant number PTDC/LLT-LIN/29552/2017 (Project Possession and location: microvariation in African varieties of Portuguese).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: The data used in this paper are available under “Palma” at gamma.clul.ul.pt/CQPweb/ (accessed on 30 October 2023). Access to the corpora requires previously requested login credentials.

Acknowledgments: We are grateful to the anonymous reviewers for their enlightening feedback on the first version of this paper and to Afonso Miguel for his feedback on the Mbundu data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

DEM = demonstrative, FUT = future tense; HAB = habitual aspect; IMP = impersonal; LOC = locative; N = noun; NEG = negation; PRES = present tense; PROG = progressive aspect; PRON = pronoun; PST = past tense; REFL = reflexive; REL = relative marker; TAM = tense-aspect-mood.

Notes

¹ The same can be said for other languages, such as German, which prefer to select a complementizer rather than a Case-marked relative pronoun in P-chopping relative clauses (see Tarallo and Myhil 1983 for the acquisition of four typologically distinct languages by native English speakers). Similarly, in the L1 acquisition of relative clauses by French and Italian children, relatives that required a preposition and a relative pronoun (i.e., pied-piping), such as Indirect Object, genitive or locative relative clauses, were generally introduced only by the complementizer *che/que* ‘that’ and followed by a resumptive pronoun, rather than being formed by pied-piping as in standard French and Italian (Guasti and Cardinaletti 2003).

² Ballarè and Inglese (2022, p. 11) emphasize that in several typologically unrelated languages, the semantically locative relativizers in its origin expand their function to all the roles in Keenan and Comrie’s (1977) Accessibility Hierarchy (AH), “so that it comes to be the standard marker for the relativization of all roles on the AH in a given language.” This is illustrated by Trinidad English Creole in (i):

(i) *Hi doz be seein a gyal weh live down S. Fernando.*
 he HAB PROG seeing the girl where live Prep S. Fernando
 ‘He dates/goes out with a girl who lives in San Fernando.’ (adapted from Mühleisen 2013, p. 67)

³ Since our search was restricted to *onde*, *que* and *qual/quais*, it is virtually possible that we overlooked instances of non-standard spatial locative relatives. However, if they occur, we expect these cases to be residual, since we are not aware of non-standard relativization mechanisms without these forms.

⁴ We did not include the extralinguistic variable ‘schooling’ since a preliminary analysis of this factor failed to produce meaningful results.

⁵ Mbundu is also known in the literature as Kimbundu, i.e., with overt realization of class 7 prefix *-ki*, which refers to the language spoken by the Ambundu people.

⁶ Mbundu exhibits considerable variation (cf. Chatelain 1888/1889, pp. xii–xiv), for example between its urban and rural use. Locative relativization in Mbundu has not been addressed in the literature, but it came to our attention that the locative morphology may also, additionally, occur on the head noun according to the Goal/Locative distinction (*ku-inzo* vs. *mu-inzo*), triggering the corresponding class 17 and 18 locative agreement on the verb-*kala* ‘to be, to stay, to remain’.

⁷ Chimbutane (2002, p. 151) suggests that these relative markers are indeed canonical post-nominal demonstratives, which follows from the fact that these two identical forms cannot cooccur (i.e., *N DEM REL). The use of demonstrative as relativizers is also attested in other Bantu languages (e.g., Van de Velde 2021).

⁸ Changana’s noun classification system is discussed in work by Langa (2012) and Ngunga and Simbine (2012), among others. Roughly 15 different classes have been proposed.

⁹ *Nê* is the contraction of the locative preposition *ni* and 3SG pronoun *ê* (*ni + ê > nê*).

¹⁰ Note that most cases of locative P-chopping with *que*, such as the examples in (15) and (16a,b), could also be analyzed as cases of replacement of *onde* by *que*. In some contexts, such as the following appositive locative relative of a toponym, this would arguably be the only possibility according to standard Portuguese:

- (i) *mesmo* *na* *Água Bobô,* *que* *vive* (STP)
 even in.the Água Bobô REL he.lives
 ‘even in Água Bobô where he lives’

We also came across a few examples of the opposite tendency, i.e. cases where *onde* is used instead of *que*, such as (ii):

- (ii) *Quais* *são* *as* *província do* *país* *onde* *você* *conhece?* (AP)
 which are the province of.the country where you know
 ‘Which are the provinces of the country you know?’

- 11 Cases where resumption co-occurred with pied-piping, as in (17c), were classified as resumption only. Similarly, cases of P-chopping with *onde*, as in (16c,d) were treated as chopping only.
- 12 It is usually assumed that $p < 0.05$ is statistically significant and $p < 0.001$ statistically highly significant (a probability of less than one in a thousand of being wrong) (Field et al. 2012).
- 13 It was shown that objects with well-defined boundaries, i.e., containers, favor directional readings of preposition *em* in the same three AVP. However, the type of predication and physical movement were shown to be statistically even more significant semantic properties.
- 14 Another distinction that has been used is between concrete and abstract, where concrete is defined as being perceivable by the human senses (Wiechmann 2015, p. 91). Concrete would then apply to containers and possibly also to areas, and abstract to non-physical locations.
- 15 *School* as a physical building. *School* as an abstract concept (as in ‘All children *school* should go to school’ = be entitled to education) is a non-physical location.
- 16 The question here concerns the categorial status of *onde* and *que* in locative relative clauses. If we follow Rinke and Aßmann (2017), in prepositional (locative) relative clauses, *que* is an intransitive demonstrative similar to an e-type pronoun, and therefore it can combine with a preposition that sometimes ends up being chopped. We will leave the discussion of the status of *onde* for further work, since we need to take into account its adverbial nature and the fact that it is able to combine with Goal/Source prepositions (e.g., *aonde*, *donde*).
- 17 However, Rinke and Aßmann (2017) argue that the criteria typically used to distinguish relative pronouns from complementizers are not always valid. Since our goal is not to further discuss the nature of the relative morpheme, this issue is left for future research.

References

- Alexandre, Nélia. 2000. A Estratégia Resumptiva em Relativas Restritivas do Português Europeu [The Resumptive Strategy in European Portuguese Restrictive Relative Clauses]. Master’s dissertation, University of Lisbon, Lisbon, Portugal.
- Alexandre, Nélia. 2012. *The Defective Copy Theory of Movement: Evidence from Wh-Constructions in Cape Verdean Creole*. Amsterdam: John Benjamins Publishing Company.
- Alexandre, Nélia, and Francisco Lopes. 2022. Onde[REL] em PCV: Contacto linguístico ou mudança interna? [Onde[REL] in Caboverdean Portuguese: Language contact or internal change?]. In *Novas dinâmicas do português no Atlântico: África e Brasil* [New Dynamics of Portuguese in the Atlantic: Africa and Brazil]. Edited by Carlos Figueiredo, Rita Gonçalves, Tjerk Hagemeijer and Márcia Oliveira. Lisbon: Lisbon International Press, pp. 135–160.
- Alexandre, Nélia, and Tjerk Hagemeijer. 2002. Pronomes resumptivos e abandono de preposição nos crioulos atlânticos de base lexical portuguesa [Resumptive pronouns and preposition stranding in Atlantic Portuguese lexified creoles]. In *Actas do XVII Encontro Nacional da APL* [Proceedings of the XVII National Meeting of APL]. Lisboa: Colibri, pp. 17–29.
- Alexandre, Nélia, and Tjerk Hagemeijer. 2013. Estratégias de relativização de PPs no mundo luso-atlântico: Crioulos de base lexical portuguesa e variedades do português [PP-relativization strategies in the Luso-Atlantic world: Atlantic Portuguese lexified creoles and varieties of Portuguese]. In *Para a História do Português Brasileiro—Volume III: Sintaxe Comparativa entre o Português Brasileiro e Língua Crioulas de Base Lexical Portuguesa* [For the History of Brazilian Portuguese—Vol. III: Comparative Syntax between Brazilian Portuguese and Portuguese Lexified Creoles]. Edited by Denilda Moura and Marcelo Sibaldo. Maceió: Edufal, pp. 49–71.
- Alexandre, Nélia, Rita Gonçalves, and Tjerk Hagemeijer. 2011a. A formação de frases relativas em português oral de Cabo Verde e de São Tomé [The formation of relative clauses in spoken Portuguese in Cape Verde and São Tomé]. In *Textos Seleccionados do XXVI ENAPL 2010* [Selected Texts from XXVI ENAPL 2010]. Edited by Armanda Costa, Isabel Falé and Pilar Barbosa. Lisboa: APL, pp. 17–34.
- Alexandre, Nélia, Rita Gonçalves, and Tjerk Hagemeijer. 2011b. PP-relativization in African varieties of Portuguese: Between contact and convergent change. Paper presented at Rethinking Contact-Induced Change, Universiteit Leiden, June 9–11.
- Arim, Eva, Maria Celeste Ramilo, and Tiago Freitas. 2004. Estratégias de relativização nos meios de comunicação social portugueses [Relativization strategies in the Portuguese social media]. In *Actas do XIX Encontro da Associação Portuguesa de Linguística* [Proceedings of the XIX National Meeting of APL]. Edited by Tiago Freitas and Amália Mendes. Lisboa: APL, pp. 279–88.
- Ballarè, Silvia, and Guglielmo Inglese. 2022. The development of locative relative markers: From typology to sociolinguistics (and back). *Studies in Language* 46: 220–57. [CrossRef]

- Brandão, Sílvia Figueiredo. 2011. Concordância nominal em duas variedades do português: Convergências e divergências [Nominal agreement in two varieties of Portuguese: Convergences and divergences]. *Veredas* 15: 164–178.
- Brito, Ana Maria. 1991. A Sintaxe das Orações Relativas em Português: Estrutura, Mecanismos Interpretativos e Condições Sobre a Distribuição dos Morfemas Relativos [The Syntax of Relative Clauses in Portuguese: Structure, Interpretative Mechanisms, and Conditions on the Distribution of Relative Morphemes]. Doctoral dissertation, INIC, Porto, Portugal.
- Brito, Ana Maria. 2001. Relativas de genitivo no português europeu e no português de Moçambique [Genitive relatives in European and Mozambican Portuguese]. In *Actas do XVI Encontro Nacional da APL [Proceedings of the XVI National Meeting of APL]*. Edited by Clara N. Correia and Anabela Gonçalves. Lisboa: Colibri, pp. 115–29.
- Brito, Ana Maria. 2002. Relativas de genitivo “estranhas” no Português de Moçambique: Erros ou sinais de mudança? [“Strange” genitive relatives in Mozambican Portuguese: Errors or signs of change?]. In *As Ciências Sociais nos Espaços de Língua Portuguesa: Balanços e Desafios: Actas, 2 [The Social Sciences in the Spaces of the Portuguese Language: Balances and Challenges: Pro-ceedings, 2]*. Edited by António C. Gonçalves and Rui Centeno. Porto: FLUP, pp. 329–36.
- Chatelain, Heli. 1888/1889. *Grammatica elementar do kimbundu ou língua de Angola [Elementary grammar of Kimbundu or language of Angola]*. Genebra: Typ. De Charles Schuchardt.
- Chimbutane, Feliciano. 1996. A estratégia de pronome resumptivo na formação de orações relativas de OD e de OBL do português de Moçambique [The resumptive pronoun strategy in the formation of DO and OBL relative clauses in Mozambican Portuguese]. In *Actas do XI Encontro da APL [Proceedings of the XI National Meeting of APL]*. Lisboa: Colibri, vol. III, pp. 225–48.
- Chimbutane, Feliciano. 2002. Grammatical Functions in Changana: Types, Properties and Function Alternations. Master’s dissertation, The Australian National University, Canberra, Australia.
- Comrie, Bernard. 1981. *Language Universals and Linguistic Typology: Syntax and Morphology*, 2nd ed. Cambridge: Blackwell.
- Corrêa, Vilma. 2001. Variação sintática em Portugal e no Brasil: Orações relativas [Syntactic variation in Portugal and Brazil]. In *Actas do XVI Encontro Nacional da APL [Proceedings of the XVI National Meeting of APL]*. Lisboa: Colibri, pp. 615–26.
- Elisabeth, Aßmann, and Esther Rinke. 2017. Relative clauses in a spoken corpus of European Portuguese: Identifying the factors determining their variation. *Revista de Estudos Linguísticos da Universidade do Porto* 12: 9–39.
- Espírito Santo, Ana, Nélia Alexandre, and Sílvia Perpiñán. 2023. The role of resumption in the acquisition of European Portuguese prepositional relative clauses by Chinese learners. *Second Language Research*. [CrossRef]
- Field, Andy, Jeremy Miles, and Zoë Field. 2012. *Discovering Statistics Using R*. London: SAGE Publication.
- Gonçalves, Perpétua. 1996. *Português de Moçambique: Uma Variedade em Formação [Mozambican Portuguese: A Variety in Formation]*. Maputo: Universidade Eduardo Mondlane.
- Gonçalves, Perpétua. 2010. *A Génese do Português de Moçambique [The Genesis of Mozambican Portuguese]*. Lisbon: INCM.
- Gonçalves, Perpétua. 2013. O português em África [Portuguese in Africa]. In *Gramática do Português [Grammar of Portuguese]*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura and Amália Mendes. Lisbon: Fundação Calouste Gulbenkian, vol. 1, pp. 157–78.
- Gonçalves, Perpétua, and Feliciano Chimbutane. 2004. O papel das línguas bantu na génese do português de Moçambique: O comportamento sintático de constituintes locativos e direcionais [The role of the Bantu languages in the genesis of Mozambican Portuguese: The syntactic behavior of locative and directional complements]. *Papia* 14: 7–30.
- Gonçalves, Rita, Inês Duarte, and Tjerk Hagemeyer. 2022. Dative microvariation in African varieties of Portuguese. *Journal of Portuguese Linguistics* 21: 1–39. [CrossRef]
- Gonçalves, Rita, Inês Duarte, and Tjerk Hagemeyer. 2023. Objetos diretos em variedades africanas do português: Um estudo de caso de microvariação [Direct objects in African varieties of Portuguese: A case-study of microvariation]. In *Para o Estudo Comparativo de Variedades do Português: Questões Teórico-Metodológicas e Análises de Dados [For the Comparative Study of Portuguese Varieties: Theoretical and Methodological Issues and Data Analyses]*. Edited by Sílvia Figueiredo Brandão and Sílvia Rodrigues Vieira. Berlin: De Gruyter, pp. 53–83.
- Guasti, Maria Teresa, and Anna Cardinaletti. 2003. Relative clause formation in Romance child’s production. *Probus* 15: 47–89. [CrossRef]
- Hagemeyer, Tjerk. 2004. Going in the clause: *ba* and *be* in Santome. *Journal of Portuguese Linguistics* 3: 71–95. [CrossRef]
- Hagemeyer, Tjerk. 2016. O português em contacto em África [Portuguese in contact in Africa]. In *Manual de Linguística Portuguesa [Handbook of Portuguese Linguistics]*. Edited by Ana Maria Martins and Ernestina Carrilho. Berlin: De Gruyter, pp. 43–67.
- Hagemeyer, Tjerk. 2018. From creoles to Portuguese: Language shift in São Tomé and Príncipe. In *The Portuguese Language Continuum in Africa and Brazil*. Edited by Laura Álvarez López, Perpétua Gonçalves and Juanito Avelar. Amsterdam and Philadelphia: John Benjamins, pp. 169–84.
- Hagemeyer, Tjerk, Amália Mendes, Rita Gonçalves, Catarina Cornejo, Raquel Madureira, and Michel Génereux. 2022b. The PALMA corpora of African varieties of Portuguese. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*. Edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani and et al. Paris: European Language Resources Association (ELRA), pp. 5047–53.
- Hagemeyer, Tjerk, António Leal, Raquel Madureira, and João Cordeiro. 2022a. Goal arguments of *ir* ‘to go’ and *chegar* ‘to arrive’ in three African varieties of Portuguese. *Journal of Portuguese Linguistics*. 21: 1–40. [CrossRef]

- Hagemeyer, Tjerk, Haldane Amaro, Michel Génereux, Iris Hendrickx, Amália Mendes, and Abigail Tiny. 2014. *Santome Corpus*. Lisbon: Centro de Linguística da Universidade de Lisboa.
- Hardie, Andrew. 2012. CQPweb-combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17: 380–409. [CrossRef]
- INE (Instituto Nacional de Estatística de Angola). 2016. *Resultados Definitivos do Recenseamento Geral da População e da Habitação de Angola 2014 [General Population and Housing Census of Angola 2012]*. Luanda: Instituto Nacional de Estatística.
- INE (Instituto Nacional de Estatística de Moçambique). 2019. *IV Recenseamento Geral da População e Habitação 2017 [IV General Population and Housing Census 2017]*. Maputo: Instituto Nacional de Estatística.
- INE (Instituto Nacional de Estatística de São Tomé e Príncipe). 2013. *IV Recenseamento Geral da População e da Habitação 2012. Características Educacionais da População [IV General Population and Housing Census 2012. Educational Characteristics of the Population]*. São Tomé: Instituto Nacional de Estatística.
- Inverno, Liliana. 2011. Contact-Induced Restructuring of Portuguese Morphosyntax in Interior Angola. Doctoral dissertation, Universidade de Coimbra, Coimbra, Portugal.
- Jon-And, Anna. 2011. Variação, Contato e Mudança Linguística em Moçambique e Cabo Verde: A Concordância Variável de Número em Sintagmas Nominais do Português [Variation, Contact, and Linguistic Change in Mozambique and Cape Verde: Variable Number Agreement in Noun Phrases in Portuguese]. Doctoral dissertation, Stockholm University, Stockholm, Sweden.
- Kato, Mary. 1993. Recontando a História das Relativas em uma Perspectiva Paramétrica [Retelling the history of relative clauses in a parametric perspective]. In *Português Brasileiro: Uma Viagem Diacrônica—Homenagem a Fernando Tarallo [Brazilian Portuguese: A Diachronic Journey—Tribute to Fernando Tarallo]*. Edited by Ian Roberts and Mary Kato. Campinas: Ed. Unicamp, pp. 223–61.
- Kato, Mary. 2010. Optional prepositions in Brazilian Portuguese. In *Romance Linguistics 2008—Interactions in Romance*. Edited by Karlos Arregi, Zsuzsanna Fagyal, Silvina Montrul and Annie Tremblay. Amsterdam: John Benjamins Publishing Company, pp. 171–84.
- Keenan, Edward, and Bernard Comrie. 1977. Noun phrase accessibility and Universal Grammar. *Linguistic Inquiry* 8: 63–99.
- Kenedy, Eduardo. 2007. A Antinaturalidade de Pied-Piping em Orações Relativas [The Antinaturalness of Pied-Piping in Relative Clauses]. Doctoral dissertation, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.
- Langa, David. 2012. Morfofonologia do Verbo em Changana [Morphophonology of the Verb in Changana]. Doctoral dissertation, Universidade Eduardo Mondlane, Maputo, Mozambique.
- Lyons, Christopher. 1999. *Definiteness*. Cambridge: Cambridge University Press.
- Macaba, Fátima João. 1996. Para uma Sistematização das Orações Relativas no Changana [Toward a Systematization of Relative Clauses in Changana]. Bachelor thesis, Universidade Eduardo Mondlane, Maputo, Mozambique.
- Mingas, Amélia. 2000. *Interferência do Kimbundu no Português Falado em Lwanda [Interference of Kimbundu on Portuguese Spoken in Luanda]*. Porto: Campo das Letras.
- Móia, Telmo. 1992. A Sintaxe das Orações Relativas sem Antecedente Expresso do Português [The Syntax of Relative Clauses without Overt Antecedent in Portuguese]. Master's dissertation, University of Lisbon, Lisbon, Portugal.
- Mühleisen, Susanne. 2013. Trinidad English Creole. In *The Survey of Pidgin and Creole Languages*. Edited by Susanne Michaelis, Philippe Maurer, Martin Haspelmath and Magnus Huber. Oxford: Oxford University Press, vol. 1, pp. 61–69.
- Ngunga, Armindo, and Madalena Simbine. 2012. *Gramática Descritiva da Língua Changana [Descriptive Grammar of Changana]*. Maputo: Center of African Studies, Universidade Eduardo Mondlane.
- Nikitina, Tatiana. 2008. Pragmatic factors and variation in the expression of spatial goals. The case of into vs. in. In *Syntax and Semantics of Spatial P*. Edited by Anna Asbury, Jakub Dotlačil, Berit Gehrke and Rick Nouwen. Amsterdam: John Benjamins Publishing Company, pp. 175–95.
- Peres, João, and Telmo Móia. 1995. *Áreas Críticas da Língua Portuguesa [Critical Areas of the Portuguese Language]*. Lisboa: Caminho.
- Ribeiro, Ilza. 2009. As sentenças relativas [Relative clauses]. In *Português Afro-Brasileiro [Afro-Brazilian Portuguese]*. Edited by Dante Lucchesi, Alan Baxter and Ilza Ribeiro. Salvador: EDUFBA, pp. 185–208.
- Rinke, Esther, and Elisabeth Aßmann. 2017. The syntax of relative clauses in European Portuguese. Extending the determiner hypothesis of relativizers to relative *que*. *Journal of Portuguese Linguistics* 16: 1–26. [CrossRef]
- Selas, Catarina. 2014. Relativas Cortadoras no Português Europeu Falado: Interação com as Variáveis Sociais [Chopping in Spoken European Portuguese: Interaction with Social Variables]. Master's dissertation, University of Minho, Braga, Portugal.
- Tarallo, Fernando. 1985. The filling of the gap: Pro-drop rules in Brazilian Portuguese. In *Selected Papers from the XIIIth Linguistic Symposium on Romance Languages*. Edited by Larry D. King and Catherine A. Maley. Amsterdam: John Benjamins Publishing Company, pp. 355–75.
- Tarallo, Fernando, and John Myhill. 1983. Interference and natural language processing in second language acquisition. *Language Learning* 33: 55–76. [CrossRef]
- The Jamovi Project. 2023. Jamovi (Version 2.3) [Computer Software]. Available online: <https://www.jamovi.org> (accessed on 30 October 2023).
- Van de Velde, Mark. 2021. The Bantu relative agreement cycle. *Linguistics* 59: 981–1015. [CrossRef]
- Veloso, Rita. 2013. Subordinação relativa [Relative subordination]. In *Gramática do Português [Grammar of Portuguese]*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura and Amália Mendes. Lisbon: Fundação Calouste Gulbenkian, vol. 2, pp. 2061–133.

- Vondrasek, Rose. 1999. Ronga relative clauses and Noun Phrase accessibility. In *Ronga Linguistics. University of Washington Working Papers in Linguistics*. Edited by Sharon Hargus and Manuel da Conceição. Seattle: University of Washington, vol. 16, pp. 115–40.
- Vries, Mark de. 2002. *The Syntax of Relativization*. Utrecht: LOT.
- Wiechmann, Daniel. 2015. *Understanding Relative Clauses*. Berlin: De Gruyter.
- Zeller, Jochen. Forthcoming. Locatives. In *The Oxford Guide to the Bantu Languages*. Edited by Lutz Marten, Nancy Kula, Ellen Hurst and Jochen Zeller. Oxford: Oxford University Press. Available online: www.jzeller.de/pdf/Loc.pdf (accessed on 30 October 2023).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

On the Variability of Portuguese Duration Adverbials with *Por* and *Durante*

Diana Santos

Linguatca & Department of Literature, Area Studies and European Languages, University of Oslo,
0315 Oslo, Norway; d.s.m.santos@ilos.uio.no

Abstract: The fact that there are (at least) two different translations of the English preposition *for* used to describe duration in Portuguese, namely *durante* and *por*, shows that duration is an interesting area in the tense and aspect panorama of this language, which deserves close study. In this paper, I present an empirical study based on four different corpora to investigate the matter, looking at the Portuguese varieties from Brazil and Portugal and different text genres (general and specialised newspapers, oral interviews, and fiction). The material spans circa two centuries. This paper identifies some specific uses and preferences of the two adverbials and confirms and challenges other claims in the linguistic literature. It shows that narrative fiction is definitely different from newspaper text due to its need to describe short duration and that planned duration can be expressed using both kinds of adverbials.

Keywords: tense and aspect; corpus linguistics; Portuguese; duration adverbials

1. Introduction

Duration adverbials are part and parcel of the tense and aspect system of any language and are usually key ingredients for uncovering an aspectual system. Vendler (1967)'s seminal work on defining the English aspectual system and coining the quadripartition states–activities–accomplishments–achievements did exactly that. Santos (1996b, 1996a, 2004) replicated Vendler's methodology for Portuguese.

However, Santos (1996b, 1996a, 2004) was concerned with the big picture of the overall systems and did not have access to large corpora. In this paper, I will concentrate specifically on particular adverbials, in this case, describing duration, to have a closer look at their behaviour, looking at a large number of occurrences.

Sentences with *for* in English have two very different translations into Portuguese depending on whether the duration includes the present moment:

- (1) I have lived in Oslo for two years.
Vivi em Oslo durante dois anos. [period totally in the past, using Perfeito¹]
live-Perfeito;1S in Oslo for two year-PL
Vivo em Oslo há dois anos. [period including now, using Presente²]
live-Presente;1S in Oslo since two year-PL

This is one of the first instances of contrastive data that any foreign learner of Portuguese has to deal with. In fact, this is a general contrast between Germanic and Romance languages, but for ease of exposition, I will only deal with Portuguese and English in the present paper.

The second contrastive data, and one which we will be especially concerned with here, is that a *for* adverbial can, in fact, be translated into Portuguese in two additional ways using the verb in Perfeito ('vivi'), namely without a preposition or with the preposition *por*. Cf. the following examples, also translating *I have lived in Oslo for two years*:

- (2) Vivi em Oslo por dois anos. [period totally in the past, using Perfeito]
live-Perfeito;1S in Oslo for two year-PL

Citation: Santos, Diana. 2024. On the Variability of Portuguese Duration Adverbials with *Por* and *Durante*. *Languages* 9: 94. <https://doi.org/10.3390/languages9030094>

Academic Editor: Patrícia Amaral

Received: 8 October 2023

Revised: 29 February 2024

Accepted: 4 March 2024

Published: 12 March 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Vivi em Oslo dois anos. [period totally in the past, using Perfeito]
live-Perfeito;1S in Oslo two year-PL³

It is this variability—or rather, the alternation between *por* and *durante* followed by temporal duration—that I want to study closer in this paper, using empirical data from distinct corpora.

This is an example of what Talmy (1983, p. 277f) has beautifully pointed out in his paper about how languages structure space:

Rather than a contiguous array of specific references, languages instead exhibit a smaller number of such references in a scattered distribution over a semantic domain. [...] Their locations must nevertheless be to a great extent arbitrary, constrained primarily by the requirement of being “representative” of the lay of the semantic landscape, as evidenced by the enormous extent of non-correspondence between specific morphemes of different languages, even when these are spoken by the peoples of similar cultures.

So, although Portuguese and English are obviously related, they mark in their grammars different details of temporal specification.

This paper is structured as follows. In Section 2, I discuss grammars and research papers on the subject of these adverbials. In Section 3, I discuss the objectives of the present study and describe the corpora used in Section 4. Then, Sections 5–7, respectively, discuss duration length, tense usage, and more specific questions, such as planned vs. non-planned periods of time. This paper ends with a study of convergence and divergence between European Portuguese (EP) and Brazilian Portuguese (BP) in Section 8, before concluding in Section 9.

2. What Grammars and Previous Research Say about Duration Adverbials

2.1. Reference Grammars

It is a well-known fact that one can use \emptyset /*durante*/*por* to specify a period of time in Portuguese. However, there are not many works that discuss, let alone explain, why there are three competing forms. In general grammars of Portuguese, we see this presented without further comment.

In subsection §16.3.1 *Duração de estados e processos* (‘duration of states and processes’), Mória and Alves (2013, p. 575ff) stated the following:

A duração de estados e processos é tipicamente marcada—em português europeu contemporâneo—por sintagmas com a preposição *durante*, que indicam o tempo durante o qual uma situação se mantém. [...] Um facto linguístico interessante é que a preposição *durante* pode muitas vezes ser omitida sem alteração substancial de significado. (The duration of states and processes is typically marked—in contemporary European Portuguese—by phrases with the *durante* preposition, which indicate the time during which the situation is maintained. (...) An interesting linguistic fact is that the preposition *durante* can often be omitted without significantly changing the meaning. (my translation))

As for *por*, Mória and Alves (2013, p. 577) added the following:

Quanto às expressões com *por*, apesar de não serem frequentes no português europeu atual como marcadores de duração equivalentes a *durante*, ao contrário do que acontece no português brasileiro, são possíveis em alguns contextos especiais; *A emissão esteve no ar apenas por alguns segundos.*

A preposição *por* tem, entretanto, um uso corrente como marcador de uma forma particular de duração, que designamos duração planeada.

A Ana saiu por meia hora.

A Ana foi para Paris por duas semanas. (As for the expressions with *por*, although they are not frequent in present-day European Portuguese as duration

markers equivalent to *durante*, in contrast to what happens in Brazilian Portuguese, they are possible in some special contexts: *The emission was on air just for some seconds*. The preposition *por* has, meanwhile, current use as a marker of a particular kind of duration, which we call planned duration. (my translation)

As for Brazilian grammars, I was only able to find a short mention of *por* in the context of its relevant meaning by Bechara ([1971] 1999, p. 318), where the grammarian cited 14 meanings or uses of *por*, of which the tenth referred to “time, duration”. As for *durante*, Bechara ([1971] 1999, p. 299) only stated that it is considered a derived preposition from the verb *durar* (‘to last’) but did not discuss its meaning or use.

This distinction was discussed in passing by (Santos 1993, pp. 401–2), who claimed that the difference between *durante* and *por* is related to the original aspectual class before it becomes an accomplishment (in both cases). In this analysis, *por* naturally applies to temporary states and *durante* applies to activities (using a mix of new and Vendler’s categories).⁴

2.2. Research Works

The first research I am aware of about *durante* and *por* and their differences was published by Rosinda Rodrigues in 1994 (Rodrigues (1994)). She looked at the felicity of these duration adverbials with Vendler’s four aspectual classes, relying on her own judgements in basic sentences. She made the following claims:

- Because of the subinterval property, states and activities can occur with duration adverbials with *durante* (p. 500).
- She discussed event-states (events that include a resulting state, in terms of the aspectual classification by Borillo (1984)). The examples given were *parar*, *cessar*, *partir*, *vir*, *deixar*, *fechar*, *abandonar* (‘stop’, ‘quit’, ‘leave’, ‘come’, ‘leave’, ‘close’, ‘abandon’) on page 501, and she stated that they are felicitous with *durante*.
- Accomplishments in the progressive can occur with *durante* adverbials (p. 502).
- Pluralised or iterative events occur with *durante* adverbials (p. 503).
- Temporary situations can be used with *por* but not with *durante*.
- When used with achievements, *por* implies saturation. In other words, instead of indicating iterativity, it describes a total amount. This analysis was inspired by Berthonneau (1991).⁵ The example given was:

(3) A Inês comeu gelados por um mês. (I. ate ice cream for a month (before getting fed up)).

- *Por* can denote a planned interval before it is over.

Later on, Mória (2001) contended that *por* is used by BP speakers in contexts where EP speakers would use *durante*. As for European Portuguese, he claimed the following:

1. *Por* is not felicitous with states or activities but acceptable with achievements (he also used Vendler’s classification), in which case, it measures the duration of the result.
2. *Por* is more felicitous when the situation refers to a prediction for the future.
3. *Por* is more felicitous when the situation can be controlled.
4. *Por* is more felicitous when the period is vague.

Most of these claims were based on acceptability judgements, with three possibilities: “OK”, “?”, or “?”. However, how these judgements were elicited was not explained so they were probably mainly those of the author. Some claims were illustrated with corpus examples but no corpus study was undertaken. Mória (2011) revisited these matters, separating duration adverbials that work as arguments from those that are adjuncts. He suggested a binary partition of adjunct adverbials: those that are anchored and those that are not (referred to, respectively, as temporal location adverbs and strict duration adverbs by Mória (2005)). Anchored adverbials are those that, in addition to specifying the duration, also convey time (“normally coincident with the temporal perspective point implied by the tense” (Mória 2011)).⁶ In the tables presented in the aforementioned paper, *durante* can only

be used to specify the non-anchored duration of atelic situations. However, Mória did not discuss the difference between *durante* and *por*, except in a footnote by Mória (2005, p. 62), which is repeated here:

Modern European Portuguese does not normally use *por*-phrases to express simple atelic duration (unless in some restricted cases, e.g., those expressing very short duration like *só a vi por uns segundos* [‘I only saw her for a second’]). However, there are many instances of this use in classical Portuguese writers.

It should be clarified that in what follows, I deal solely with what Mória calls “strict duration adverbs”, not anchored ones, when annotating and revising the corpus data.

As for Brazilian Portuguese, Basso and Bergamini-Perez (2016, p. 353ff) discussed several duration adverbials and claimed that the main difference between *durante* and *por* is the kind of duration measures accepted: while *por* only accepts “primary” duration measures involving explicit temporal nouns, *durante* also accepts “secondary” measures describing an event, like *jogo* (‘game’), *filme* (‘film’), and *peça* (‘play’). Additionally, they explicitly mentioned the high similarity of the two adverbials, thus supporting Mória’s contention about their equivalence in Brazilian Portuguese.

3. Corpus Analysis

The research I describe here tries to validate (or challenge) these claims, as well as identify some other reasons and/or uncover new linguistic generalisations on these matters. Specifically, I investigate the following factors:

- Length of the duration;
- Tense of the clause;
- Negative semantic prosody⁷;
- The issue of temporariness (something that is extremely important in the aspectual fabrics of Portuguese, as reflected by the famous *ser/estar* distinction).

In some cases, it is possible to automatically identify the features discussed, whereas in other cases, I use a random subset of cases and judge them one by one.

In any case, an important aspect of the present work is that it does not rely on the author’s idiolect to obtain examples, nor does it select examples according to a particular purpose: it is corpus-based. That is, it uses corpora of authentic examples. When, due to the sheer volume of the examples, it is impossible to analyse every example, the choice is again not directed by a particular theory or aim; it is simply random.

However, a corpus-based study does not mean that the linguist’s intuition is not called for. On the contrary, it has to be duly exercised when interpreting a plethora of examples, even though they may not be part of her idiolect at all. This is especially true when the material encompasses different varieties and ages, let alone different literary authors’ styles. Because I am aware that this interpretation may, at times, be challenged, and in order to allow for further work on the subject, all examples used, analysed, and referenced in the present paper are available for inspection at <https://www.linguateca.pt/documentacao/artigoPorDurante.html> (accessed on 1 March 2024).

4. Corpora Used

I used four different corpora, each containing roughly the same amount of Brazilian and Portuguese material, covering four different genres: newspaper texts from 1994–1995 from the CHAVE collection (Santos and Rocha 2005); specialised newspaper texts from the 1950s, 1970s, and 2000s about football, fashion, and health from the ConDiv corpus (Soares da Silva 2008); literary fiction from the year 1500 onwards from the Literateca corpus (Santos 2019); and transcribed interviews from the Museu da Pessoa corpus (Almeida et al. 2001). The number of occurrences of the duration adverbials appear in Table 1.

All these corpora are publicly available for querying on the Web through the AC/DC project (Santos 2014). Some of them are also available for download.

Table 1. Number of occurrences of the duration adverbials *por* and *durante* in the PT/BR corpora.

Corpus	Durante	Por
CHAVE	1709/5003	3420/2220
ConDiv	671/241	152/248
Museu da Pessoa	67/144	3/51
Literateca	506/739	897/1469

The sizes of the different corpora, together with the (unrevised) counts of *durante* and *por* duration phrases in each variety, are presented in Table 2. In Appendix A, I provide the actual search expressions to allow for reproducibility, as well as the concrete versions of the corpora, which can change over time. As the corpora have not been fully revised, it is important to be aware that, notwithstanding the good performance of PALAVRAS (Bick 2000), some cases may be missing, and others spurious. Some of these numbers are revised in the following sections.

Table 2. Overview of the four corpora. The corpus size is given in tokens (i.e., it includes both words and punctuation marks).

Corpus	PT Size	Durante	Por	BR Size	Durante	Por
CHAVE	64,096,543	5003	2220	41,901,785	1709	3420
ConDiv	3,322,364	671	152	2,625,012	241	248
Literateca	25,005,408	989	1415	12,462,820	857	906
Museu da Pessoa	346,458	67	3	1,182,943	144	51

Another way to observe the differences is in Figure 1.

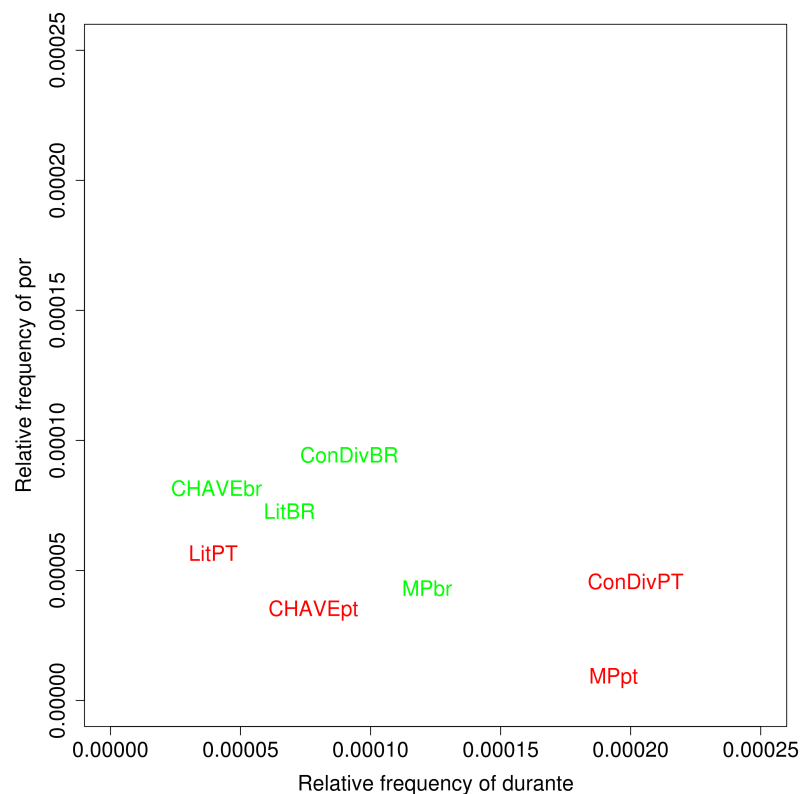


Figure 1. The relative frequency of *durante* and *por* duration adverbials per variety and per corpus, without human revision. Red indicates EP, green BP.

One can immediately see that there is a wide range of differences among the corpora as far as these temporal adverbials are concerned. While in CHAVE, each variety is a mirror image of the other (with *durante* more frequent than *por* in Portugal, and *por* more frequent than *durante* in Brazil), in the other corpora, the two varieties behave more similarly. However, while *durante* predominates in interviews and specialised journalese, *por* is far more frequent in literary texts.

The rest of this paper is an attempt to explain these differences and identify or confirm the reasons for the use of each of these prepositions.

5. Long and Short Durations

I started by inquiring whether *por* emphasises short duration and *durante* emphasises long duration, assuming that no preposition represents the neutral case.

First, I counted the three instances of each duration noun in Literateca. It should be noted that “others” includes more than duration,⁸ and “Total” includes a few other cases. I ordered the nouns by duration length, as shown in Tables 3–6. I also listed the vague cases, notably with the meta-noun *tempo* (“time”), to give a better quantitative characterisation of these durational phrases.⁹

The data confirm that *por* is more frequent for short durations across all corpora and more frequent for vague duration specifications. Most interestingly, durations seem to sharply differ by genre. In the literature, as shown in Table 3, there is a multitude of short durations unparalleled by any other kind of text.

Table 3. Frequency of “por/durante measure duration noun” per noun in Literateca. The numbers with “+” represent cases with no quantifiers, like *durante anos* (‘for years’) or *por horas* (‘for hours’).

	Durante	Por	Others
instante	48 + 6	389 + 278	4616
momento	105 + 11	472 + 357	6978
segundo	42 + 11	58 + 13	3670
minuto	115 + 16	117 + 35	2617
hora	147 + 59	235 + 32	14,050
dia	590 + 36	491 + 51	31,235
semana	97 + 35	31 + 2	1827
mês	198 + 44	155 + 16	5578
ano	528 + 83	415 + 31	21,849
decênio	1	0	7
século	106 + 36	50 + 42	2127
Total	1977	2413	94,554
tempo	567	2809	35,179

One can also observe that the specification of temporal duration (using the adverbials we are concerned with here) is much more relevant in news compared to interviews (as described in Table 4), even if those interviews are supposed to mirror the interviewee’s life and might, therefore, prompt several temporal descriptions.

In Table 6, we can see that daily newspapers mainly report situations that last weeks, months, or years, whereas the specialised newspapers shown in Table 5 are more focused on days, hours, minutes, and seconds. This may occur because of sports reporting—football in this case—but texts on health also seem to address this kind of temporal period).

However, it is not only absolute temporal duration that counts. An examination of the examples shows that what is long and what is short is dependent on the kind of event or situation described. Long and short are relative to the noun they modify (contrast *long way* and *long hair*), and, I would claim, are also often subjective.

Table 4. Frequency of “por/durante + measure duration noun” per noun in Museu da Pessoa.

	Durante	Por	Others
instante	0	0	3
momento	0	3	289
segundo	0	0	12
minuto	1	0	71
hora	0	2	607
dia	59	3	1186
semana	18	0	162
mês	18	8	533
ano	98	29	3614
década	0	1	32
decênio	0	0	1
século	2	0	15
Total	195	46	2911
tempo	51	30	1616

Table 5. Frequency of “por/durante + measure duration noun” per noun in ConDiv.

	Durante	Por	Others
instante	4	13	80
momento	6	19	654
segundo	13	2	153
minuto	222	46	5152
hora	86	25	1466
dia	194	49	1595
semana	110	15	567
mês	63	26	909
ano	132	70	3764
década	0	0	45
decênio	1	0	
século	3	1	47
Total	834	225	14,432

Table 6. Frequency of “por/durante + measure duration noun” per noun in CHAVE.

	Durante	Por	Others
instante	6	100	74
momento	22	262	9860
segundo	52	139	3126
minuto	390	422	12,004
hora	1032	903	23,103
dia	2710	1938	51,926
semana	1517	244	14,962
mês	2048	854	37,100
ano	4888	2517	151,655
década	359	75	3447
decênio	0	2	26
século	131	27	2981
Total	13,155	7483	310,264

The following examples show this clearly, and I believe that the durations are regarded as short given the entire period that the main clause event covers, namely a medieval war, the writing of a novel, and a period in a depressive mood:

- (4) (a) As memórias desses tempos não nos dizem quem quebrou as pazes juradas: só sabemos que a luta interrompida **por dois anos** começou de novo. (“the fight

interrupted for two years started again’).

(b) Em Junho, de novo interrompi A Selva, desta vez não **por alguns dias, mas por dois meses** e sem desgosto algum, com um prazer todo febril e exultante. (‘I interrupted the book, nor for some days, but for 2 months’).

(c) Começou a viver solitário, e desse programa só o carnaval o arrancou **por três dias**. (‘out of that state of mind only Carnival was able to grab him for three days’).

Similarly, the following examples show that different absolute durations can be regarded as long given the context. Three days, seven weeks, and two months can be considered long for plundering, the preparation of a trip, and a convalescence, respectively.

(5) (a) O bairro levantado ficou **durante três dias** entregue ao saco e, expulsos os seus habitantes, foi arrasado. (was for three days ransacked).

(b) Jacinto não conhecia Torges, e foi com desusado tédio que ele se preparou, **durante sete semanas**, para essa jornada agreste. (with a rare tedium he prepared himself for seven weeks’).

(c) Silveira assistiu ao enfermo **durante dois meses** de morosa convalescença. (for two months of slow convalescence).

So, in order to really ascertain whether a particular duration is regarded as short or long definitely requires close reading and manual annotation of each case, as reported in Table 10 in Section 7.

Another important issue is whether *por* marks temporariness, as in the Portuguese expression *por enquanto* (in English, “for the moment” or “for the time being”). Something temporary is obviously shorter than what is considered permanent.

In addition, one tends to wish that bad things take less time than good things, which means that one would expect a preference for *por* when reporting bad things and *durante* when reporting good things. It is, therefore, important to note that if *por* carries with it a negative opinion, it might be rhetorically minimised, and then it would be natural to occur more often with shorter durations anyway. This means that these three features—short period, negative evaluation, and temporariness—may not be independent factors, but all somehow—and possibly even diachronically—related. I revisit this after the next section.

6. Tense with Duration Adverbials

I then tried to ascertain whether tense and aspect had anything to say about duration. Note that morphosyntactic tense in Portuguese is a very rich system, so I am not talking about past, present, and future here, but about distinguishing morphosyntactic tenses (which also encode aspect).

I started by counting the tenses in the smaller amount of material, the interviews. Table 7 shows the number of times the (most frequent) tenses were present in the corpus, and the number of times they occurred with *por* and *durante*.

Table 7. Frequency of *por*/*durante* per tense in Museu da Pessoa. Mnemonics of AC/DC tenses are mentioned in the second column so that users can conduct their own searches for these tenses.

Tense		All	Durante	Por
Presente do indicativo	PRIND	66,942	6	4
Imperfeito do indicativo	IMPF	43,975	11	5
Perfeito do indicativo	PS	42,107	105	23
Infinitivo	INF	35,907	2	3
Gerúndio	GER	6260	1	0
Futuro perifrástico	FUTPER	5528	2	2
Particípio passado	PCP	4547	1	0

The picture was clear: almost only clauses in Perfeito had duration expressed (with *por* or *durante*). Many of the cases of *por* seemed to imply a short period (Example (6) (a)) and/or seemed to be negatively conceived by the speaker (Example (6) (b)):

- (6) (a) Depois porque me faltavam quatro anos para a jubilação, e o que é que eu ia lá fazer **por quatro anos?** ('what would I do there for four years?').
- (b) E nessas missões o senhor ficava longe de casa **por dois anos seguidos?** ('were you far from home for two years in a row?').

But in the Brazilian interviews, some cases with *por* seemed merely a neutral way of stating duration:

- (7) trabalhei em uma loja em Porto Alegre, e depois na Praia dos Ingleses, **por cinco meses.** ('I worked in a shop... for five months').

I then turned to the literature to see whether there were significant differences in the way the duration adverbials were employed, and I ended up closely reading all the examples; therefore, I corrected the initial numbers. Several cases of *por* that were not temporal adverbials were discarded, examples of which I present here:

- (8) (a) Apesar de **endurecido por quarenta anos de caça e carnificinas**, eu próprio sentia um nó na garganta, e creio que me fiz pálido. (hardened by forty years of hunt and carnificine).
- (b) Tinha eu chegado do continente, **prostrado por duas horas de canal da Mancha...** ('tired by two and a half hours of the Channel').
- (c) E a população muçulmana, **enfurecida por nove horas de bombardeamento**, sem polícia para a conter, ('infuriated by nine hours of bombing').
- (d) da Comenda que **mereci, por dezesseis anos de serviço** na guerra. ('...I deserved for 16 years of service in war').
- (e) Era o conselheiro Andrade, **conhecido por quarenta anos de ceias consecutivas**, desde o remoto Rocher de Cancale até os desvairamentos dos atuais. ('known for 40 years of consecutive dinners').

So, Table 8 shows the actual distributions of these temporal adverbials in the literature by tense:

Table 8. Frequency of *por*/*durante* per tense in Literateca, all varieties.

Tense		All	Durante	Por
Presente do indicativo	PRIND	1,185,431	42	62
Perfeito do indicativo	PS	797,760	218	231
Imperfeito do indicativo	IMPF	721,507	29	56
Infinitivo	INF	712,880	50	75
Particípio passado	PCP	384,494	28	32
Gerúndio	GER	236,056	19	38
Futuro perifrástico	FUTPER	33492	1	0

As for the CHAVE collection, the amount of data required a selection, so I just annotated 200 cases (100 per variety) randomly selected with Perfeito. The results are presented in Table 9.

Table 9. Frequency of *por*/*durante* with *Perfeito* in *CHAVE*. Cases that turned out not to be *Perfeito* were removed from the counts.

Variety	Durante	Por
BR	33	60
PT	78	19
all	111	79

These data seem to agree with *Móia*'s claim that *durante* is clearly preferred in European Portuguese, whereas Brazilians may have a preference for *por*.

7. More Specific Questions about Duration

For each case, I annotated whether it was a clearly negative or clearly positive action or situation and whether I regarded it as long or short. In cases where I did not feel it conveyed any such connotation (long or short), I did not add any annotations.

I also annotated the (in this case, quite clear) cases of planned duration. Finally, I also identified the cases where a negative sentence with *por um momento/instante* was an emphatic way to convey the negation through a minimiser, as in the following examples:¹⁰

- (9) Mas o Tomé, servo cumpridor das ordens que lhe davam, nem **por um momento** hesitou em dirigir para ali os seus passos. ('not for a moment he hesitated').
 Devo dizer também que, vendo-a, ouvindo-a, eu não supus nem **por um momento** que no homicídio de que ela se acusava pudesse haver o que se chama verdadeiramente um crime, isto é, uma intenção infame ou perversa. ('I did not suppose for a moment that...').
 José, que tudo ouviu, não se intimidou **por um momento**. ('did not cower for a moment').

First, I present the numbers obtained in Table 10, which correspond to the most frequent tenses in *Literateca*, amounting to 937 sentences.

Table 10. Frequency of *por*/*durante* per annotation in *Literateca*. LONG stands for long time, TEMP stands for short time or temporary, NEGAT stands for semantically negative situation, PLAN stands for planned action, and NAOENF stands for emphatic not.

Annotation	Durante	Por
LONG	273	53
NEGAT LONG	44	26
TEMP	22	275
NEGAT TEMP	1	15
PLAN	37	161
NAOENF	-	7

From these numbers, I concluded that a negative attitude about the event does not play any role, or at least, that the data cannot support this hypothesis. In fact, the data additionally showed that to make such decisions based only on the particular sentence is quite hard, as the three next examples try to illustrate:

- (10) Depois, fuja, abandonem o lugar, a capela, tudo, porque a seca vai continuar ainda **por dois anos** ('run away, leave the place, the chapel, everything, because the drought is going to go on for two years still').
 Ia ficar sozinha **por um mês**, o amigo era chamado a S. Paulo para um negócio urgente. ('She would be alone for one month').
 As memórias desses tempos não nos dizem quem quebrou as pazes juradas: só sabemos que a luta interrompida **por dois anos** começou de novo. ('the fight interrupted for two years').

While the first is probably consensual—it is not good for a drought to continue, and the sentence even exhorts people to flee—in order to ascertain whether it would be good (or negative or neutral) for the feminine character to be alone, one would have to know more about the plot. And although with modern eyes, to fight again would be considered negative, my impression is that in the text in question, the fight is considered good and, therefore, the interruption bad. So, regardless of my own interpretation of badness, what I should annotate is what the author meant. But anyway, I think we can safely conclude that this category plays no role in the choice between *durante* and *por*.

On the other hand, I believe it is fair to claim that *durante* is clearly preferred when the speaker is conveying long duration (317 vs. 79 cases), and *por* is clearly preferred when the speaker is conveying short or temporary periods (90 vs. 23 cases). I am aware that I was not able to distinguish between short and temporary in my subjective annotation.¹¹

But, after considering all these short/temporary cases, a large proportion of which use *um momento*, it was clear that this is a key ingredient of fictional narratives in Portuguese. It is often associated with changes in the disposition or thoughts of the character in question,¹² which are obviously out of place in factual journalese. And this may explain why *por* is much more common in fiction, independent of the epoch.

Given that both Rodrigues (1994) and Mória and Alves (2013) cited planned duration as a clear case of *por* adverbials in Portuguese, and Rodrigues (1994, p. 505) even explicitly said that *durante* is not allowed in such contexts,¹³ I expected that *por* would be categorical with planned duration. However, there were enough instances in the literature of planned duration with *durante* to show that it is—or was—just a preference or that “planned” is not the whole story.

Some examples of planned duration with *durante* are as follows:

- (11) (a) para aí edificarem o teatro do Bairro Alto, pagando anualmente 240000, **durante catorze anos**, renováveis, salvo se o proprietário quisesse continuar a reedificação do palácio. (‘in order to build the theatre, paying annually X for 14 years’).
 (b) Àqueles que seguissem Sancho nas incursões contra os sarracenos ou formassem parte do seu exército concedia ele, papa, **durante quatro anos**, as mesmas indulgências que os concílios haviam decretado para os que se votavam às longínquas expedições de ultramar (‘he, the Pope, would issue for four years the same indulgences...’).
 (c) Liberato e Frederico deviam demorar-se apenas quatro meses com seus pais, seguindo depois para a América do Norte, onde **durante dois** anos estudariam com observação solícita os sistemas, processos, [...] (L. and F. should stay only 4 months with their parents, and then go to North America, where for two years they would study...’).
 (d) Eu também vou em breve atirar fora a minha pena e as minhas declamações, para me fazer **durante três dias** espontâneo e lógico (‘in order to become spontaneous and logical for three days’).

As for the differences between having *durante* or no preposition, I tried to rephrase a considerable amount of *durante* adverbials¹⁴ to ascertain whether this was possible at all, and if yes, what the difference would be, if any.

Interestingly, I was able to find 42 cases (out of 135) where *durante* was not removable, as shown in the following examples:

- (12) (a) Gillooly, que foi preso na semana passada e é considerado o «cérebro» da agressão a Kerrigan, **foi interrogado durante seis horas** pelo FBI. (‘was questioned for the duration of six hours’).
 (b) Passada a fronteira, **não pararam durante duas semanas** e mantiveram-se sempre à frente dos indonésios. (‘did not stop for two weeks’).
 (c) Nós conseguimos ter um escritório da ONG Vitae Civilis somente em 93, na época recebendo um apoio de U\$ 2.500, nem sei quanto valeria hoje, mas com

esse dinheiro a gente alugou uma sala durante um ano e ainda arrumou uma funcionária ('we rent a room for one year').

(d) No silêncio e na solidão dos claustros escapou durante seis séculos o ténue pergaminho que nos conserva a memória de Afonso Mendes Sarracines ('escaped from destruction for six centuries the tenuous parchment').

In all cases, *durante* specifies that one is talking about a consecutive period, not a simple (possibly discontinuous) duration. In addition, in (c), the alternative expression, *um ano*, would mean something like once during a given year.

For the other cases, the large majority of *durante* adverbials convey a long duration compared to those with the bare adverbial.

8. Convergence and Divergence of Portuguese

Until now, we have been looking at different genres produced at different times, given the available corpora, namely contemporary Portuguese for CHAVE and Museu da Pessoa, and mainly the nineteenth century for the literature. Although I have argued that the differences are mainly due to genre, it is definitely an advantage to have another resource that may help us check whether time is also an important feature here.

The use of ConDiv, which was precisely designed to address the study of the convergence and divergence of the varieties from Brazil and Portugal, may help us here. In Table 11, we show the numbers for the three different decades, per variety, as well as the ratio between *durante* and *por*.

Table 11. Frequency of *por*/*durante* per decade and variety in ConDiv.

Variety	Decade	Durante	Por	Ratio
BR	1950	111	62	1.79
BR	1970	65	84	0.774
BR	2000	46	73	0.630
PT	1950	204	38	5.37
PT	1970	255	28	9.1
PT	2000	119	31	3.8

Although these numbers are probably not enough to come to a definite conclusion, it is interesting to observe that the use of *por* consistently increases in Brazilian Portuguese, whereas the predominance of *durante* in European Portuguese seems to diminish.

In order to check whether this is an artefact of different distributions per theme, I repeated the queries separately per domain, as presented in Table 12, respectively, for football, fashion and health.

Table 12. Frequency of *durante* and *por* per decade and variety in ConDiv. For the football and health domains, the *durante*–*por* ratio is also presented.

Variety	Decade	Durante	Por	Ratio	Durante	Por	Durante	Por	Ratio
		Football			Fashion		Health		
br	1950	36	18	2.0	39	26	36	18	2.0
br	1970	35	53	0.66	5	2	25	29	0.86
br	2000	32	43	0.74	1	8	13	22	0.59
pt	1950	92	10	8.5	2	1	110	27	4.07
pt	1970	26	4	6.5	3	0	226	24	9.4
pt	2000	51	22	2.3	1	1	67	8	8.4

We can see no significant differences between the themes, except that in the football domain, the Portuguese practice in the 2000s seems to converge toward the Brazilian style,

with *por* increasingly more frequent. However, this cannot be seen in the health domain (see also Figure 2).

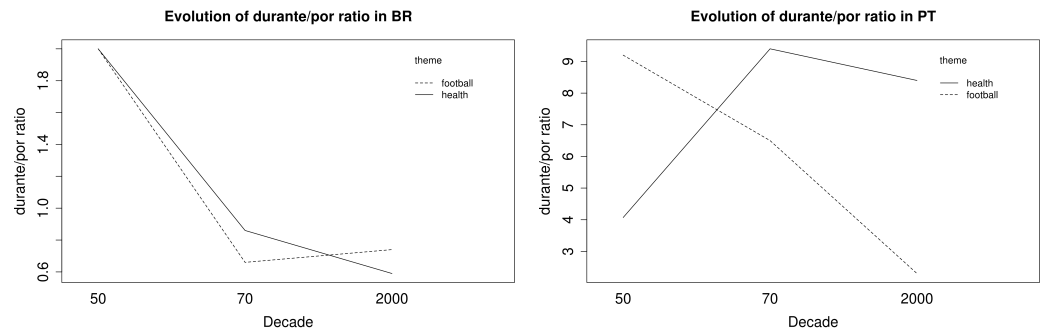


Figure 2. The evolution of the durante–por ratio by variety and theme.

9. Conclusions

This paper reported on a corpus study for the purpose of understanding the use of *por* and *durante* adverbials in Portuguese, both intervarietally and intergenre, as well as over three different decades spanning the 20th and 21st centuries.

This study was more than a simple comparison of counts since a considerable number of examples were closely read and annotated for categories such as “planned activity”, “contextually long duration”, “temporary/short duration”, “negatively seen”, and “emphatic negation”.

The main conclusions were that genre matters, and the marked preference for *por* in literary texts is related to the frequency of occurrence of events with short durations in fictional narratives. Planned activities favour the use of *por* in both varieties, but it is also possible to use *durante*. The use of *durante* compared to the use of no preposition, especially in fiction, seems to convey a long duration. It displays the attitude of the writer towards the period, in addition to expressing its length. *Por*, in fiction, seems to convey temporariness and short duration when not describing planned activities. In informative texts, where the attitude of the writer is less frequently expressed, there seems to be a marked preference for *durante* in texts from Portugal, and a more liberal use of *por* in texts from Brazil, which increased around the 1960s, if we take football- and health-related newspaper texts as good indicators of the language as a whole. In negative contexts, *por um instante* or *por um momento* simply emphasises negation. Duration adverbials with either preposition are mainly expressed using verbs in Perfeito, that is, indicating events that are completely in the past.

It is left for further research to investigate the possible dependence of the duration adverbials on (a) aspectual class, and (b) the existence of iterated readings. This would imply annotation of these two pieces of information, which has not yet been done.

Also, a more fine-grained study of literary texts might uncover (a) different stylistic preferences of different authors, and (b) a chronological map of the *durante–por* variation on the two sides of the Atlantic. The literary corpus itself includes more than just fiction, as discussed in Freitas and Santos (2023), so it is possible to perform a more fine-grained analysis with such material.

All the data compiled for this paper, along with the annotations, are available for inspection¹⁵ so that other researchers can validate and/or improve on them.

Funding: This research received no external funding.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://www.linguateca.pt/documentacao/artigoPorDurante.html> (accessed on 1 March 2024).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. AC/DC Commands Used to Obtain the Examples

Search queries used to compare duration adverbials with *por*, *durante*, and neither for Tables 3–6:

```
[lema="por"] [pos="NUM_card|DET.*"] @[sema=".*dur.*"]
[lema="durante"] [pos="NUM_card|DET.*"] @[sema=".*dur.*"]
[lema!="por|durante"] [pos="NUM_card|DET.*"] @[sema=".*dur.*"]
```

The reason for this simplified query is to obtain a meaningful comparison for the “others” category.

Search queries used to find duration adverbials with *por* in general:

```
[lema="por"] [pos!="N|V.*|PROP.*" & lema!="lá|aí|de\+aqui"]+ [sema=".*dur.*"] within s
```

Search queries used to find duration adverbials with *durante* in general:

```
[lema="durante"] [pos!="N|V.*|PROP.*" & lema!="lá|aí|de\+aqui" & word!="o|a"]+
[sema=".*dur.*"] within s
```

Search queries used to find duration adverbials with *tempo*:

```
[lema="por"] [pos!="N|V.*" & lema!="lá|aí|de\+aqui"] [lema="tempo" ] within s
[lema="durante"] [pos!="N|V.*" & lema!="lá|aí|de\+aqui"] [lema="tempo" ] within s
[lema!="por|durante"] [pos!="N|V.*" & lema!="lá|aí|de\+aqui"] [lema="tempo" ] within s
```

Search queries for finding (preceding) verbs associated with duration adverbials, specifically with the Perfeito indicative tense.

```
[pos="V.*" & temcagr=".*PS.*"] [pos!="V.*"]* "por" [pos!="N|V.*|PROP.*" &
lema!="lá|aí|de\+aqui"]+ [sema=".*dur.*"] within s
```

Note that for simplicity, we assumed that the verb precedes the duration adverbial. But, during the revision and annotation, we assigned the “correct” verb, sometimes following the adverbial.

The corpus versions used are shown in the table below.

Table A1. Overview of the AC/DC corpora used.

Corpus	Version	Date of Creation
Museu da Pessoa	9.2	8 October 2023
CHAVE	10.0	20 July 2023
Literateca	11.4	4 September 2023
ConDiv	12.1	16 July 2023

Notes

- One of the many past tenses in Portuguese, corresponding roughly to both past simple and present perfect in English; named *pretérito perfeito do indicativo* in Portuguese grammar.
- Present tense in Portuguese, corresponding roughly to the English simple present; named *presente do indicativo* in Portuguese grammar.
- Owing to the number and size of the examples, and especially the limited role of morphological features in the overall discussion, I will not provide detailed glosses in what follows.
- This kind of analysis was inspired by the aspectual network of Moens (1987), where grammatical categories like *for* adverbials were modelled as operators that changed aspectual type in grammar.
- It is interesting to note that Portuguese is also different from French, the language Berthoneau analysed, in that *pendant* is similar to *durante* but *pour* is more general than *por*, also translated by *para* in a temporal context.
- This is analogous to the distinction made by Hitzeman (1993, p. 44ff) between instantiated and non-instantiated interval templates for *for* adverbials in English.
- This term, namely “semantic prosody”, was coined by Sinclair and is not relative to prosody proper (see Stubbs (2009)). In a nutshell, it refers to the communicative purpose of an extended lexical unit.

- 8 Notably, temporal anchoring, like in *há dois anos* ('two years ago'), *três dias depois* ('three days later'), and *às sete horas* ('at seven o'clock').
- 9 Interestingly, the noun *época* ('epoch') does not, with *por* or *durante*, describe a duration but simply an imprecise location in time, translatable as "at that time" in English.
- 10 I had initially called this a "never" meaning, but I am grateful to an anonymous reviewer for noting that only the second of the three examples could, in fact, be paraphrased by *nunca* ('never'). Patrícia Amaral has also called my attention to the relevance of Horn (2001).
- 11 This is probably something that can only be decided by eliciting answers with minimal pairs especially construed by only being short, or only being temporary, and asking native speakers to produce the preposition. Such an experiment is left for future work. But, in authentic language examples, it appears that most utterances convey both, and this is why it is hard to decide whether only one is expressed.
- 12 A significant number of cases with *por um momento* concerned perception, bodily changes, and hesitation.
- 13 Cf. "a propriedade de programar a duração do intervalo I que "por" pressupõe mas "durante" recusa" (the property to program the duration of the interval that *por* presupposes but *durante* refuses).
- 14 Specifically, the first 15 cases from Literateca in Perfeito, the first 15 cases of Brazilian interviews from Museu da Pessoa in Perfeito, the first 15 cases of Portuguese interviews in the same corpus and same tense, the first 15 cases from CHAVE in Perfeito for each variety, and the first 15 cases of each of the six subcorpora of ConDiv, divided by decade and variety; altogether 135 cases.
- 15 See the following link: <https://www.linguateca.pt/documentacao/artigoPorDurante.html> (accessed on 1 March 2024).

References

- Almeida, José João, Jorge Gustavo Rocha, Pedro Rangel Henriques, Sónia Moreira, and Alberto Manuel Simões. 2001. Museu da Pessoa—Arquitetura. In *Atas do Congresso Nacional de Bibliotecários, Arquivistas e Documentalistas Arquivistas*. (Porto, 23, 24 e 25 de Maio de 2001). Porto: BAD.
- Basso, Renato, and João Francisco Bergamini-Perez. 2016. Adjuntos temporais e measure phrases: Uma proposta semântica. *Cadernos de ESTUDOS LINGUÍSTICOS* 58: 345–67.
- Bechara, Evanildo. 1999. *Moderna gramática Portuguesa. Edição Revista e Ampliada*. Rio de Janeiro: Editora Lucerna. [First published 1971]
- Berthonneau, Anne-Marie. 1991. Pendant et pour, variations sur la durée et donation de la référence. *Langue Française* 91: 102–24. [CrossRef]
- Bick, Eckhard. 2000. The Parsing System "Palavras": Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Ph.D. thesis, Aarhus University, Aarhus, Denmark.
- Borillo, Andrée. 1984. Pendant et la specification temporelle de duree. *Cahiers de Grammaire* 8: 55–75.
- Freitas, Cláudia, and Diana Santos. 2023. Gender Depiction in Portuguese. Distant reading Brazilian and Portuguese literature. *Journal of Computational Literary Studies* 2: 1–29. [CrossRef]
- Hitzeman, Janet. 1993. Temporal Adverbials and the Syntax-Semantics Interface. Ph.D. thesis, University of Rochester, Rochester, NY, USA. Technical Report 931.
- Horn, Laurence R. 2001. *A Natural History of Negation*. Stanford: CSLI Publications.
- Moens, Marc. 1987. Tense, Aspect and Temporal Reference. Ph.D. thesis, University of Edinburgh, Edinburgh, UK.
- Móia, Telmo. 2001. Sobre a expressão da duração em Português Europeu e Português Brasileiro: O uso de sintagmas com a preposição por. In *Boletim da Associação Brasileira de Linguística, Número Especial, II Congresso Internacional da ABRALIN, Fortaleza, Março de 2001, Anais—Vol I*. Edited by M. E. Soares. Fortaleza: ABRALIN, pp. 415–19.
- Móia, Telmo. 2005. Portuguese Expressions of Duration and its English Counterparts. *Journal of Portuguese Linguistics* 5: 37–73. [CrossRef]
- Móia, Telmo. 2011. Sobre a expressão lexical da duração e da localização temporal em português. In *Linguística do Português: Rumos e Pontes*. Edited by Mathias Arden, Christina Märzhäuser and Benjamin Meisnitzer. München: Martin Meidenbauer, pp. 251–269.
- Móia, Telmo, and Ana Teresa Alves. 2013. Tempo adjunto e tempo discursivo. In *Gramática do Português*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura and Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, pp. 557–81.
- Rodrigues, Rosinda. 1994. Os adverbiais durante Q N DE T e POR Q N DE T. Duas formas de quantificar a duração. In *Actas do X Encontro Nacional da Associação Portuguesa de Linguística*. Évora: APL, pp. 497–508.
- Santos, Diana. 1993. Integrating tense, aspect and genericity. In *Actas do IX Encontro da Associação Portuguesa de Linguística*. Coimbra: APL, pp. 391–405.
- Santos, Diana. 1996a. Tense and aspect in English and Portuguese: A Contrastive Semantical Study. Ph.D. thesis, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa, Portugal.
- Santos, Diana. 1996b. Uma classificação aspectual portuguesa do português. In *Actas do XII Encontro da Associação Portuguesa de Linguística*. Braga: APL, pp. 299–315.
- Santos, Diana. 2004. *Translation-Based Corpus Studies: Contrasting English and Portuguese Tense and Aspect Systems*. Amsterdam: Rodopi.

- Santos, Diana. 2014. Corpora at Linguateca: Vision and Roads Taken. In *Working with Portuguese Corpora*. Edited by Tony Berber Sardinha and Telma de Lurdes São Bento Ferreira. London: Bloomsbury, pp. 219–36.
- Santos, Diana. 2019. Literature studies in Literateca: Between digital humanities and corpus linguistics. In *HUmanists and the Digital Toolbox: In honour of Christian-Emil Smith Ore, Oslo*. Edited by Martin Doerr, Øyvind Eide, Oddrun Grønvik and Bjørghild Kjelsvik. Sofiemyr: Novus Forlag, pp. 89–109.
- Santos, Diana, and Paulo Rocha. 2005. The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE. In *Multilingual Information Access for Text, Speech and Images, 5th Workshop of the Cross-Language Evaluation Forum, CLEF 2004, Bath, UK, September 15-17, 2004, Revised Selected Papers*. Edited by Carol Peters, Paul Clough, Julio Gonzalo, Gareth J. F. Jones, Michael Kluck and Bernardo Magnini. Lecture Notes in Computer Science. Berlin/Heidelberg: Springer, Volume 3491, pp. 821–32.
- Soares da Silva, Augusto. 2008. O corpus CONDIV e o estudo da convergência e divergência entre variedades do português. In *Perspectivas Sobre a Linguateca/Actas do Encontro Linguateca: 10 Anos*. Edited by Luís Costa, Diana Santos and Nuno Cardoso. pp. 25–28.
- Stubbs, Michael. 2009. The search for units of meaning: Sinclair on empirical semantics. Memorial article: John Sinclair (1933–2007). *Applied Linguistics* 30: 115–37. [CrossRef]
- Talmy, Leonard. 1983. How language structures space. In *Spatial Orientation: Theory, Research, and Application*. Edited by Herbert L. Pick and Linda P. Acredolo. New York: Plenum Press, pp. 225–82.
- Vendler, Zeno. 1967. *Linguistics in Philosophy*. Ithaca: Cornell University Press.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

When a New Pronoun Crosses the Border: The Spread of *A Gente* on the Brazilian-Uruguayan Frontier

Cíntia Pacheco¹, Ana Carvalho^{2,*} and Marta Pereira Scherre¹¹ Department of Linguistics, University of Brasília, Brasília 70910-900, Brazil; cintia.pacheco@unb.br (C.P.)² Department of Spanish and Portuguese, University of Arizona, Tucson, AZ 85721, USA

* Correspondence: anac@email.arizona.edu

Abstract: This study shows that the incorporation of the first-person plural pronoun *a gente* has not only reached the southernmost tip of the Brazilian territory, but has crossed the border and entered Uruguayan Portuguese, or varieties of Portuguese spoken in northern Uruguay by Portuguese–Spanish bilinguals. This finding is based on the quantification of the *a gente/nós* variable in sociolinguistic interviews carried out in two border communities: Aceguá, Brazil, and Aceguá, Uruguay. The analysis of interviews recorded on each side of the border yielded a total of 1000 tokens that were submitted to a multivariate analysis. Following the premises of comparative sociolinguistics, we compared the distribution of the variable on both sides of the border and found that although Uruguayans used *a gente* less often than Brazilians, this innovation, preferred by young speakers, is incorporated in both dialects, following similar linguistic paths. These results show that Uruguayan Portuguese has incorporated the pronominal *a gente* in its grammar in a clear sign of convergence towards Brazilian Portuguese and divergence from Spanish, despite the coexistence with Spanish that categorically uses *nosotros* as the first-person plural pronoun and reserves the cognate *la gente* for its purely lexical meaning ‘the people’.

Keywords: pronominal *a gente*; language contact; Uruguayan Portuguese; Border Portuguese; morphosyntactic variation; language change

Citation: Pacheco, Cíntia, Ana Carvalho, and Marta Pereira Scherre. 2024. When a New Pronoun Crosses the Border: The Spread of *A Gente* on the Brazilian-Uruguayan Frontier. *Languages* 9: 98. <https://doi.org/10.3390/languages9030098>

Academic Editor: Patrícia Amaral

Received: 28 December 2023

Revised: 1 March 2024

Accepted: 6 March 2024

Published: 13 March 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In Brazilian Portuguese, the use of the first-person plural pronoun *a gente*, which resulted from the grammaticalization of the noun phrase ‘the people’, is well attested and is increasingly displacing the pronoun *nós* (‘we’) (Omena 1996a, 1996b; Lopes 2003; Zilles 2005; Vianna and Lopes 2015; among others). Example (1) illustrates a case where the speaker clearly uses *a gente* rather than *nós* to refer to himself and his wife:

- (1) O final de semana mesmo que a gente pode se juntar geralmente. Aproveito com a minha família mesmo com meus filho com minha mulher, a gente leva eles para passear.
‘It is on the weekends that we can gather usually. I enjoy it with my family really with my kids, with my wife, we take them to outings.’ (Middle-aged man, São Paulo, Brazil)¹

As we show in this study, this ongoing linguistic change in Portuguese has not only reached the southernmost tip of Brazilian territory, but has entered Uruguayan Portuguese, a variety of Portuguese spoken in northern Uruguay by Portuguese–Spanish bilinguals. We base our findings on quantification of the *a gente/nós* variable in subject position in sociolinguistic interviews conducted in two border communities: Aceguá, Brazil, and Aceguá, Uruguay, by the first author (Pacheco 2014). A comparison of the distribution of this variable in both dialects points to similar trends in both communities, indicating that Spanish–Portuguese bilinguals in Uruguay have assimilated Brazilian linguistic innovations despite long-term contact with Spanish.

Portuguese in Uruguay

Since colonial times, Portuguese has been spoken in Uruguay alongside Spanish, the national language. The presence of Portuguese in Uruguayan territory resulted from a long period of disputes during colonial times between the Portuguese and Spanish crowns over a vast, open territory along the northern border of Uruguay that was sparsely populated by Portuguese-speaking settlers. Uruguayan independence in 1828 did not have immediate repercussions in the region, and not until the second half of the nineteenth century did the central government take measures to Hispanicize the north and rid the region of Portuguese and Brazilian influence through the foundation of border towns and the establishment of Spanish-only public schools. As Elizaincín (1992, p. 158) describes, the area around Aceguá was especially deserted, remaining unpopulated or barely populated until the nineteenth century. Not until 1852 were markers installed in the area to demarcate the national border (Pedemonte 1985). Then, in 1862, the Uruguayan government named the land Juncal Pacheco. By that time, Portuguese had been established as the local language on both sides of the entire Uruguayan–Brazilian border.

Due to prolonged contact with Spanish, Uruguayan Portuguese is characterized by a heavy presence of Spanish loanwords and Spanish–Portuguese code-switching. In addition, the occurrence of vernacular Portuguese morphosyntactic and phonological variants and archaic words from rural Portuguese indicates that Uruguayan Portuguese has undergone fewer diachronic changes than Brazilian Portuguese (Elizaincín et al. 1987; Carvalho 2016). In fact, Elizaincín et al. (1987, p. 85) documented only the presence of lexical *a gente* in Uruguayan Portuguese, plus consistent use of *nós* as the first-person plural pronoun. As we demonstrate, since Elizaincín and colleagues' research, Uruguayan Portuguese has incorporated the pronominal *a gente* in its grammar. The presence of this innovative variant in Uruguayan Portuguese represents convergence toward Brazilian Portuguese and divergence from Spanish (given that Spanish categorically uses *nosotros* as the first-person plural pronoun and preserves the purely lexical meaning of its cognate *la gente*). Example (2) illustrates the coexistence of lexical *la gente* and the *nosotros* conjugation (*acentuamos*) in Uruguayan border Spanish:

- (2) Ahora trato, lógico, de corregirme porque como todos se ríen principalmente la gente del sur porque dicen los riverenses acentuamos las “eses” y las “uves”.
 ‘Now I try of course, to correct myself, because since everyone laughs, especially people from the South, because they say (Ø-we) Riverans stress the ‘s’ and the ‘v’.’
 (Middle-aged man, Rivera, Uruguay)²

This study advances a long history of studies on Uruguayan Portuguese that began with Rona's seminal work in 1965 (Rona 1965). Later, Hensey conducted a series of sociolinguistic analyses of the phonological variables of Uruguayan Portuguese (Hensey 1972, 1982; among others), and Elizaincín and associates published multiple studies on morphosyntactic and lexical characteristics from the perspective of dialectology (e.g., Elizaincín et al. 1987; Elizaincín 1992). Carvalho's (1998) doctoral dissertation led to follow-up variationist studies on phonological (Carvalho 2003, 2004; Garrido Meirelles 2009; Castañeda 2011, 2016; Córdoba 2013), morphosyntactic (Carvalho and Child 2011; Pacheco 2013, 2014; Carvalho and Bessett 2015; Carvalho 2016, 2021; Pacheco 2017a, 2017b, 2017c, 2018, 2020), and discourse variables (Carvalho and Kern 2019). This research produced quantitative evidence that, despite prolonged bilingualism along the Uruguay–Brazil border, Spanish and Portuguese have retained distinct variable grammars, albeit with extensive presence of lexical borrowings and conversational code-switching. However, most of what is known about Uruguayan Portuguese has been based on data collected on the Uruguayan side, mainly in the border town of Rivera.

To address this gap, the present study analyzes the use of pronominal *a gente* among Spanish–Portuguese bilinguals in Aceguá, Uruguay, and among Portuguese monolinguals in Aceguá, Brazil, so as to enable direct comparisons of cross-national varieties of Portuguese that are in daily contact. The comparison of a bilingual and a monolingual variety

will allow us to identify convergent and divergent behavior, an essential element in the analysis of how language contact may drive language change. In comparative sociolinguistic studies (Tagliamonte 2013), it is vital that data collection and analysis of both dialects are handled in the same manner and that the varieties under examination serve as appropriate reference languages (Poplack and Levey 2010). First, this method enables us to identify to what extent *a gente* has spread to the Brazilian southern border, rather than comparing Uruguayan Portuguese with varieties of Brazilian Portuguese spoken far from the border, in regions where this linguistic innovation is widespread and well documented. Secondly, it allows us to investigate whether this innovation has crossed the border and influenced Uruguayan Portuguese, which would represent a linguistic change in light of Elizaincín and colleagues' (Elizaincín et al. 1987) finding that it was not present in the 1980s. Importantly, if *a gente* has been incorporated into Uruguayan Portuguese, this would counter the thesis that the categorical use of lexical *la gente* in Spanish without a pronominalized counterpart would hinder this innovation in Portuguese. Finally, it is important to subject both datasets to the same variationist analysis, so that the factors that condition the realization of the variable can be compared cross-dialectally. The objective of such an analysis is to compare variable patterns in each dialect and determine whether Uruguayan Portuguese variable grammar matches the linguistic and social patterns that govern the use of *a gente* in the variety of Brazilian Portuguese spoken across the border. Ultimately, this analysis sheds light on the linguistic and social forces that advance or restrain the spread of a language change across a national border. In addition, by examining a context where cognate languages are in prolonged contact, this analysis tests commonly held assumptions that similar linguistic systems in contact tend to converge, resulting in bilingual dialects that differ significantly from their monolingual counterparts.

The consistent analysis of different datasets to the same analysis and their subsequent comparison is especially important in this case because the variable under study is subject to a wide range of variation in Brazilian Portuguese. Since the 1980s, multiple variationist studies have analyzed the distribution of first-person singular pronouns and found that the rate of *a gente* as opposed to *nós* ranges from 78% (Borges 2004) to 39% (Muniz 2008), depending on the community and the method used for quantification. Comparisons of *a gente* use between national varieties are rare. To our knowledge, they are limited to Vianna (2011) and Rubio (2012), who compared Brazilian and European Portuguese, finding that the innovative form is noticeably and significantly more prevalent in Brazil than in Portugal. In addition, Rubio (2012) reported that, while this variable presents stable variation in Portugal, it shows clear signs of rapid change toward the grammaticalized form and away from *nós* in Brazil. Rubio attributes this difference to the fact that in Portugal, *a gente* is seen as nonstandard and is avoided by women and well-educated speakers. The opposite situation occurs in Brazil, where since this innovation began (as early as the eighteenth century, according to Lopes (2003)), it has spread widely (Omena 1986, 1996a, 1996b) and does not have any stigma attached to it (Zilles 2007, p. 37). Unlike in Portugal, Uruguay Portuguese is spoken in communities that are in daily contact with monolingual speakers of Brazilian Portuguese, providing a favorable context for linguistic convergence, as we explain later.

In this paper, we first explain the speech community where sociolinguistic interviews were collected, the participants' demographic information, and the methods used for data extraction and analysis. We then present our results, including the overall frequency of the variable in terms of communities and individuals and the multivariate analyses for each community. The analysis allows us to compare the variable BP and UP grammars in terms of the factor groups that are selected by each one and the ranking of the factors within these groups, and shows clear continuities across the dialects. Finally, we combine both datasets to test whether community would be chosen as a statistically significant factor due to any significant cross-dialectal distributional differences. We finally conclude that, while *a gente* has crossed the border and reached the UP grammar largely following the grammatical

paths attested in BP, *nós* is still preferred by bilinguals, indicating that this linguistic change is more advanced in BP.

2. Materials and Methods

2.1. The Aceguá Community

Aceguá is a border town of approximately 5000 residents located in the state of Rio Grande do Sul, Brazil. Its twin town of Aceguá, in the Department of Cerro Largo, Uruguay, hosts a smaller population of approximately 1500 (Figure 1).



Figure 1. Location of the twin towns of Aceguá (from Pacheco et al. 2018).

No physical boundaries separate the two communities (Figure 2), and the national border is undetectable except for the linguistic landscape of public signs in Spanish in Uruguay and in Portuguese in Brazil. The movement of people and vehicles is uncontrolled, and residents cross from one side to the other often for work, visits, and shopping.



Figure 2. Open border between Aceguá, Brazil, and Aceguá, Uruguay (photo by Pacheco 2014, p. 41).

2.2. Data Collection

The first author of this paper first visited Aceguá, Brazil, in 2009 and conducted sociolinguistic interviews with local residents. In 2011, she returned to the area and carried out similar sociolinguistic interviews with residents of Aceguá, Uruguay. The participants were identified through the snowball technique, and all promptly agreed to be interviewed. The recorded informal conversations lasted approximately one hour and took place at the participants' homes or in public spaces such as plazas and restaurants. The present analysis is based on tokens extracted from 19 Portuguese speakers from Aceguá, Brazil, and 19 Portuguese speakers from Aceguá, Uruguay, divided into three age groups and two binary genders (Table 1).

Table 1. Participants.

Age Groups	Aceguá, Brazil	Aceguá, Uruguay
15–30 years old	4 men 3 women	3 men 4 women
31–49 years old	2 men 5 women	6 women
More than 50 years old	3 men 2 women	2 men 4 women
TOTAL	19 speakers	19 speakers

Source: Pacheco (2014, p. 149).

2.3. Data Analysis

Once the interviews were transcribed, the first step was to distinguish the pronominal *a gente* from its lexical equivalent (‘the people’, occurring with a non-specific referent). The first token in Example 3, extracted from the corpus, is used with its lexical meaning: “*toda a gente se confunde*” (‘all people get confused’) carries an indeterminate referent and coincides with the Spanish equivalent, *la gente*. We excluded this type of token from our analysis because it does not carry the first-person plural meaning and, therefore, is outside the envelope of variation. The next token, *nós notamos* (‘we notice’), carries the first-person plural morpheme *-mos*, so was included in the analysis and coded as the *nós* form. The following three tokens were included as well—“*a gente criou*” (‘we created’), “*a gente fala*” (‘we speak’), and “*não fala*” (‘[we] don’t speak’)—since all are undoubtedly pronouns with first-person plural referents. Example (3) illustrates what Tagliamonte (2006, p. 168) calls a super-token that clearly shows both variants coexisting in the same conversational turn of a Uruguayan Portuguese speaker.

- (3) Isso aqui, a cultura é mais ou menos a mesma, de toda a gente se confunde. Pra nós, não notamos. . . vocês que vem de longe podem notar a diferença, mas pra nós, a gente criou um dialeto pra falar, a gente fala portunhol, não fala nem espanhol nem português.

‘This here, the culture is more or less the same, people get confused. For us, we don’t notice. . . you who come from far may notice some difference, but for us, we made up a dialect to speak it, we speak Portunhol, [we] don’t speak either Spanish or Portuguese.’ (50-year-old Uruguayan man³)

Once all the verbs with first-person referents had been identified, both in their singular (referring to either expressed and unexpressed *nós* or *a gente*, depending on the previously expressed pronoun) and plural form (referring to *nós*), they were subjected to a multivariable analysis. The decision to include singular forms and further investigate whether the referent was *nós* or *a gente* (and code accordingly) was based on the fact that verbal agreement with *nós* subjects is variable in Brazilian Portuguese (Omena 1996a, 1996b; Naro et al. 1999; Vianna 2011; Vianna and Lopes 2015; Lopes 2003; Zilles 2005; Muniz 2008; Rubio 2012; Mattos 2013, 2017; Foeger 2014; Benfica 2016; Foeger et al. 2017; Naro et al. 2017; Scherre et al. 2018a, 2018b; among others). While *a gente* subjects may also take plural verbal morphemes in some areas of Brazil, this occurred only once in both varieties of Aceguá Portuguese (Pacheco 2014), coinciding with Zilles’s results for Porto Alegre, the largest city in the area (Zilles 2005, p. 36). Given the rarity of verbal *-mos* occurring with *a gente* referents, only singular verbal forms with *a gente* referents were included in the present analysis. Once all tokens had been identified, the first phase of analysis compared the frequency of occurrence of *nós* and *a gente* across the Brazilian and Uruguayan dialects to assess whether this linguistic innovation had reached both communities. Once the presence of *a gente* had been attested on both the Brazilian and Uruguayan sides of the border, our next step was to subject the two datasets to the same multivariate analysis in

order to compare the variable patterns of each dialect and determine the permeability of the grammars.

Four linguistic factor groups and two social factors were included in this study. First, it was important to account for the type of subject; that is, to separately code verbal forms that were accompanied by an expressed subject (either *nós* or *a gente*) from those that were not. Example (4) illustrates the traditional first-person plural conjugation (*temos*, '[we] had') preceded by an explicit pronoun *nós*, followed by another plural token (*temo*) without an expressed subject.

- (4) Nós temos um clima semelhante ao do Rio Grande, pouquinho coisa mais frio. \emptyset temo quatro estações bem definidas.
'We have similar weather compared to Rio Grande, just a little bit colder. [We] have four well-defined seasons.' (50-year-old Uruguayan man)

Example 5, uttered by the same speaker, illustrates the innovative first-person plural form (*fala*, '[we] speak'), first with an explicit pronominalized *a gente* as the subject, followed by repetition of the same verb (*fala*, '[we] speak') without the subject.

- (5) A gente fala portunhol, não fala nem o espanhol nem o português
'We speak portunhol, [we] don't speak either Spanish or Portuguese'. (50-year-old Uruguayan man)

As explained earlier, while plural verbs with unexpressed pronouns do refer to implicit *nós*, a verb that lacks a plural morpheme and has no explicit pronoun could have either *nós* or *a gente* as its referent. Thus, in cases where a singular verb was produced without an expressed subject, it was necessary to look at the preceding clauses to locate an explicit *nós* or *a gente* in order to correctly code the token. Example (6) illustrates the need to consider the preceding clause (*a gente apresenta*, 'we present') in order to properly classify the following verb clause (*libera a mercadoria*, 'approve') as an *a gente* token with an unexpressed subject.

- (6) Então vem o cliente, a gente apresenta a mercadoria, libera a mercadoria, e aí é a aprovação do fiscal. Se ele carimbou tu tá aprovado.
'So the customer comes, we present the product, approve the product, and then it is up to the inspector's approval. If he stamps it, it is approved.' (Adult Brazilian male)

Example (6) leads us to the second linguistic factor considered in the present analysis: persistence of one or the other variant. It is well documented that several morphosyntactic variants, once realized, tend to be repeated, due to a tendency for similar forms to occur together within a stretch of discourse (Poplack 1980; Weiner and Labov 1983; Scherre and Naro 1991; Scherre 1998; Paiva and Scherre 2022) Thus, the first verb in Example (7) (*íamos*, 'went') was coded as first token; and the second (*juntávamos*, 'we gathered') and third (*movimentávamos*, '[we] moved') tokens were both coded as preceded by a *nós* form.

- (7) Era um lixão aquilo. Nós do Rotary íamos lá, juntávamos o lixo, movimentávamos.
'That was a big dump. We, from Rotary, used to go there, gathered the trash, moved around'. (Middle-aged Uruguayan woman)

Example (8) illustrates a series of singular verbs with a clear *a gente* antecedent. The first verb (*chegava*, '[we] arrived') was coded as the first token, while the subsequent verbs (*pagava*, '[we] paid'; *vinha*, '[we] came'; and *trazia*, '[we] brought') were also coded as preceded by *a gente*.

- (8) A gente chegava ali, pagava 50% sobre o valor da mercadoria e vinha embora ou ia para qualquer lugar do Brasil, e o resto trazia.
'We arrived there, paid 50% on the product's value, and came back or went somewhere else in Brazil, and brought the rest.' (Elderly Brazilian man)

The third linguistic independent variable included in the analysis was the verb's tense and phonic salience. These features, when combined, have been proven to predict the choice of *a gente* in Brazilian Portuguese. The underlying hypothesis for why salience is a potential predictor of *a gente* is that the less salient the difference between *nós* vs. *a gente* verbal forms is (*amava/amávamos*; *ama/amamos*), the higher the chance that *a gente* will

be chosen. In contrast, the more salient the difference between the plural and singular variants is (*amou/amamos; foi/fomos*), the higher the chance that *nós* will be chosen. When Naro et al. (1999) studied this variable in Rio de Janeiro, they found that, indeed, phonic salience was the main conditioning factor behind this variable in speakers from older generations. For younger speakers, however, verbal tense played a more significant role. The authors noted a strong interaction between the two variables of phonic salience and verb tense, given that the plural vs. singular opposition is less salient in imperfect forms (*amava/amávamos*) than in preterit forms (*amou/amamos*). More recent studies combine both variables into a single group (e.g., Mattos 2013, 2017; Foeger 2014; Benfica 2016; Foeger et al. 2017; Naro et al. 2017; Scherre et al. 2018a, 2018b). We follow the precedent of these previous studies in combining phonic salience and verbal tense into one variable.

In Brazilian Portuguese, verbal tense impacts the *nós/a gente* variable in several ways (Foeger 2014). For example, *a gente* is more common with imperfect verb forms, in order to avoid placing stress on the antepenultimate syllable (*amava/amávamos*). In addition, the lack of distinction between the present (*amamos*) and past (*amamos*) forms of first-person plural verbs may lead to more uses of *a gente* (*ama*) to avoid ambiguity. Finally, especially among urban speakers, the use of *a gente* has the potential to decrease the realization of verbal forms that lack standard plural agreement with present and imperfect forms, where the referent is *nós* (e.g., *a gente fala* instead of *nós fala*; *a gente falava* instead of *nós falava*). Thus, we follow previous studies (Zilles 2005; Foeger 2014; Scherre et al. 2018b) and include six factors within the verbal tense and phonic salience group, illustrated in Examples (9)–(15) from the least to the most salient forms:

- Gerund or infinitive (least salient)
 - (9) E a gente vendo da nascente lá. . .
'And we looking at the sunrise there.' (Young Uruguayan male)
- Imperfect (least salient singular/plural opposition)
 - (10) Daqui de Aceguá a gente tava em três.
'From here, from Aceguá, we were three.' (Adult Uruguayan male)
- Present, potentially with the same form as the preterit (least salient tense opposition)
 - (11) Porque aqui é muito raro ter um jornal ou algo do Uruguai. Aí nós lemo tudo brasileiro.
'Because here it is rare that there is a paper or something from Uruguay. Thus we read all in Brazilian.' (Young Uruguayan woman)
- Present, with a different form than the preterit (moderately salient opposition)
 - (12) E agora nós já estamos há dez anos aqui, né?
'And now we have been here for ten years, right?' (Elderly Uruguayan male)
- Preterit, potentially with the same form as the present (more salient opposition):
 - (13) A gente comprou faz pouco tempo até uma geladeira brasileira, mas a gente comprou em Melo, no Uruguai. . .
'We bought a short while ago even a Brazilian fridge, but we bought [it] in Melo, in Uruguay.' (Adult Uruguayan Woman)
 - (14) E aí nós conseguimos para a Colônia, campeonato sete, saímo campeão último ano. Claro, muito cansativo! Eu saí daqui até Porto Alegre de carro, né? Terça-feira saímo daqui.
'And then we managed to go to Colonia, the seventh championship, we won the championship in the last year. Of course, very exhausting! I left here to Porto Alegre by car, right? On Tuesday we left.' (Adult Brazilian male)
- Preterit, with a different form than the present (more salient opposition):
 - (15) Aí nós fomos pra churrascaria, era 40.00 por pessoa e um horror a churrascaria.
'Then, we went to the steak house, it was 40.00 per person and a horrible steak house.' (Young Brazilian woman)

Finally, we included in the statistical analysis whether the semantic reference to the grammatical person was specific or generic. Example (16), from Aceguá, Brazil, illustrates the use of *a gente* with a clearly specific reference: the speaker and his family.

- (16) Interviewer: Vocês moram aqui do lado do Brasil?
 ‘Do you (pl.) live here, on the Brazilian side?’
 Interviewee: Do lado brasileiro a gente mora.
 ‘On the Brazilian side we live.’ (Young Brazilian man)

In contrast, Example (17) shows that the same form may carry a generic reference, one that does not specify a particular referent, but alludes instead to a collective, generalizing the first-person plural to include everyone in the region.

- (17) E a gente aqui não tem trânsito né. Hoje mesmo eu saí com um chimarrão, eu vou guiando e tomando chimarrão.
 ‘And we here have no traffic, right. Today I went out with my tea and went driving and drinking my tea.’ (Elderly Uruguayan man)

Because the original meaning of the lexical item *a gente* is a collective ‘the people’, Omena (2003) argues that a generic reference should favor pronominal *a gente*, while a specific reference would favor *nós*. Previous studies (Omena 2003; Neves 2008) have shown that, indeed, generic reference often triggers the use of *a gente*, while specific reference tends towards *nós* in Brazilian Portuguese. In the current analysis, we investigate whether a first-person plural form with a generic reference favors *a gente*, as it does in Brazilian Portuguese, and whether Uruguayan Portuguese follows the same tendency.

Due to the small sample size, age group and binary sex were the only extra-linguistic factors included in the analysis. Age is an especially important factor in a study that investigates the possibility of language change in progress, since a preference for an innovative form among younger generations is usually interpreted as evidence that the innovation is underway in the community. Thus, three age groups were included in the analysis: 15–25 years old, 31–49 years old, and more than 50 years old.

The linguistic interviews of 19 speakers on each side of the border yielded a total of 1002 tokens that were submitted to Rbrul (Johnson 2009), yielding the following results.

3. Results

In Table 2, we compare the distribution of the *nós/a gente* variable on both sides of the border. Note that Uruguayans used *a gente* at a lower rate (29%) than Brazilians, who preferred the grammaticalized form 59% of the time.

Table 2. Overall frequency of *a gente* vs. *nós* in subject position in Aceguá, Brazil, and Aceguá, Uruguay: data from all 38 participants (19 speakers from each country).

Country	A Gente	Nós
Brazil	59% 317/541	41% 224/541
Uruguay	29% 135/461	71% 326/461

Zilles (2005) studied *a gente* vs. *nós* in Porto Alegre, the capital of the Brazilian state of Rio Grande do Sul and the closest metropolis to Aceguá. She found that *a gente* was preferred at a rate of around 70% in all syntactic positions, and at around 58% in subject position (Zilles 2005, p. 40). Given the linguistic prestige of urban centers in Brazil, it is reasonable to assume that the speech of Porto Alegre serves as a linguistic model for the border communities. Thus, Figure 3 compares the rate of *a gente* vs. *nós* in subject position across the three communities: Porto Alegre; Aceguá, Brazil; and Aceguá, Uruguay.

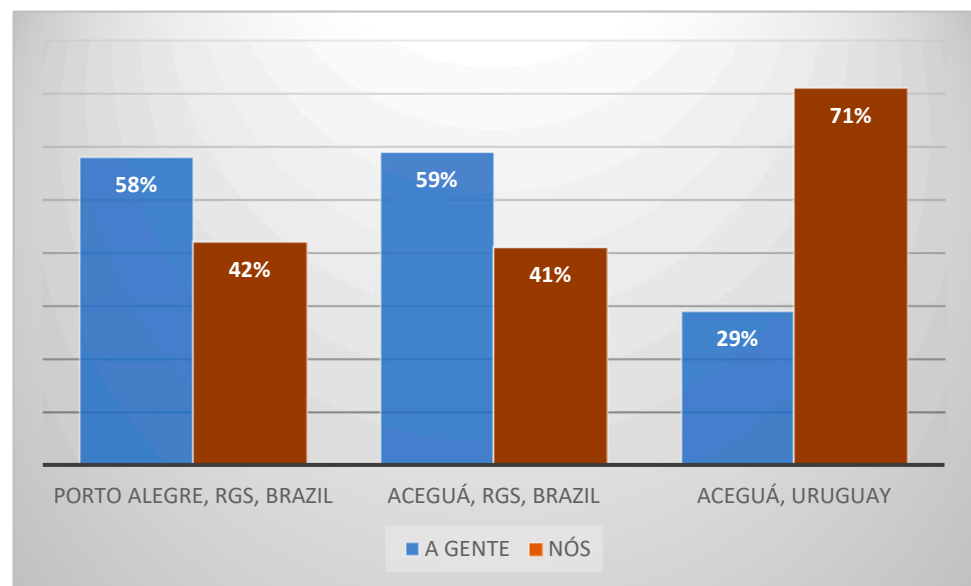


Figure 3. Overall frequencies of a gente and nós in subject position in Porto Alegre, Brazil (N = 1645), compared with Aceguá, Brazil (N = 541) and Aceguá, Uruguay (N = 461).

As Figure 3 shows, the Portuguese spoken in Porto Alegre and Aceguá, Brazil, shows similar rates of the innovative *a gente* pronoun, which has replaced more than half of the *nós* occurrences. Speakers in Aceguá, Uruguay, on the other hand, still prefer *nós*. The slower adoption of *a gente*, an urban linguistic innovation of Brazilian Portuguese, supports Carvalho’s (2016) claim that Uruguayan Portuguese tends to be more conservative and to lag behind language changes occurring in urban Brazilian Portuguese. The oscillation between *nós* and *a gente* in Uruguayan Portuguese is yet another reflection of the urban–rural continuum in which small towns in the area are situated (Carvalho 2003, 2004, 2016).

To account for individual variation, each speaker was coded in the data, allowing for identification of important differences (Figure 4).

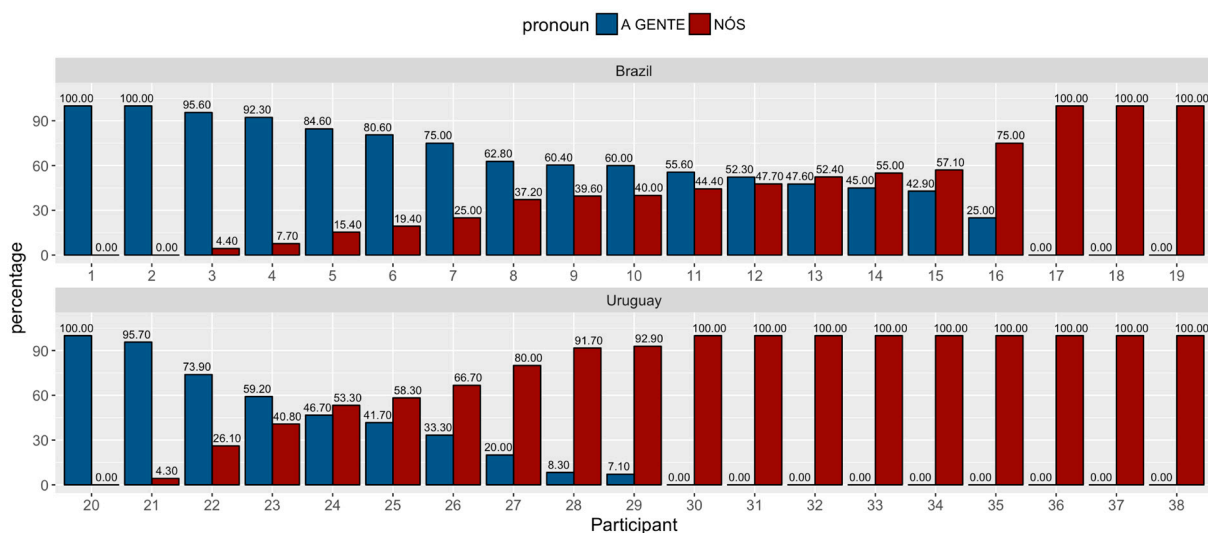


Figure 4. Overall frequency of *a gente* vs. *nós* by individual speaker in Aceguá, Brazil (top), and Aceguá, Uruguay (bottom).

As Figure 4 shows, it is possible to find speakers who used one or the other variant categorically during their sociolinguistic interviews. In Aceguá, Brazil, two speakers used only *a gente*, while three used solely *nós*. In Aceguá, Uruguay, one speaker used only *a gente*, while nine speakers used only *nós*. In order to explore the linguistic and extralinguistic

factors underlying variation, it was necessary to exclude these categorical users from the sample. Therefore, the data on Table 3 excludes three categorical *nós* users from Brazil and nine from Uruguay. The overall frequencies were recalculated, as shown in Table 3.

Table 3. Overall frequency of *a gente* vs. *nós* in subject position in Aceguá, Brazil, and Aceguá, Uruguay: data from 26 participants, (16 from Brazil, 10 from Uruguay).

Country	A Gente	Nós
Brazil	63% 317/500	37% 183/500
Uruguay	51% 135/266	49% 131/266

As Pacheco (2014, p. 248) reported, when only speakers who used both variants are considered, the difference in overall rates between Brazilian and Uruguayan Portuguese are dramatically reduced. Both groups show higher rates of *a gente*: Brazilians' usage increases slightly, from 59% to 63%, while Uruguayans' usage of *a gente* increases more steeply, from 29% to 51% of the time. We have already established that the frequency of *a gente* is much higher in Brazilian than in Uruguayan Portuguese (Table 2), and that more Uruguayan than Brazilian speakers show categorical use of *nós* (Figure 4). This pattern allows us to conclude that *a gente* is more advanced in the Brazilian dialect.

A comparison of the factors that trigger the use of *a gente* in Brazilian Portuguese allows us, first, to explore the linguistic contexts in which the innovative variant has been incorporated and by what social groups. After we establish this baseline, we investigate the linguistic and social factors that condition the use of *a gente* in Uruguayan Portuguese to explore the extent to which that usage mirrors the same linguistic and social factors as for Brazilian Portuguese. To that end, our next step was to submit both datasets containing variation to a multivariate analysis in Rbrul (Johnson 2009). Following the premises of comparative sociolinguistics, we submitted both corpora to the same analysis, which included the type of subject, discourse persistence, tense and phonic salience of the verb, and the specificity of the pronoun, in addition to each speaker's age group and sex (Table 4).

Table 4 shows that five predictors statistically influenced the use of pronominal *a gente* in Brazilian Portuguese (discourse persistence, age, verbal tense and phonic salience, type of subject, and type of reference). However, only three were significant for Uruguayan Portuguese (verbal tense and phonic salience, discourse persistence, and type of subject). This finding suggests that *a gente* is more advanced in Aceguá, Brazil, than in Aceguá, Uruguay, where the incorporation of *a gente* is gradually catching up with the tendencies detected among Portuguese monolinguals.

For both communities, a strong predictor of the use of *a gente* is discourse persistence, also known as priming (see Paiva and Scherre (2022) for a review of the priming effect on variation). As expected, once one *a gente* is produced, the likelihood that this variant will recur is very high, reaching 0.79 for Brazilian border Portuguese and 0.89 for Uruguayan border Portuguese. On the other hand, *a gente* is strongly disfavored when preceded by *nós* in Brazilian border Portuguese (0.30), and even more so in Uruguayan border Portuguese (0.13). The probability weights show that *a gente* is disfavored as a first token in Brazilian border Portuguese (0.37) and slightly disfavored in Uruguayan border Portuguese (0.44). These results provide strong evidence that, once the innovative *a gente* is used, it by itself triggers recurring uses, in line with previous research (cf. Omena 1996a; Zilles 2005; Mendonça 2010; Rubio 2012; Mattos 2013; Foeger 2014; Pacheco 2014; Vianna and Lopes 2015; Souza 2020; among others). Notably, speakers in both communities respond to this factor, similarly showing continuity across variable grammars.

Table 4. Rbrul results of pronominal *a gente* vs. *nós* in subject position in Portuguese from Aceguá, Brazil, and Aceguá, Uruguay (categorical users of *nós* excluded; random effects: individual).

Aceguá, Brazil	N = 500		Overall Proportion = 63%	Input = 0.60
	Fixed Effects:	FW	%	n/N
Discourse persistence	Preceded by <i>a gente</i>	0.79	84%	112/134
	First token	0.37	63%	174/278
	Preceded by <i>nós</i>	0.30	35%	31/88
	Range	49		
	<i>p-value</i>	2.22×10^{-9}		
Age	15–25	0.75	84%	113/135
	31–49	0.42	62%	139/224
	>50	0.31	46%	65/141
	Range	44		
	<i>p-value</i>	0.0171		
Verbal tense and phonic salience	Present, potentially with same form as preterit	0.73	82%	141/171
	Gerund or infinitive	0.60	61%	11/18
	Imperfect	0.56	52%	29/56
	Present, with different form than preterit	0.44	57%	103/180
	Preterit, with different form than present	0.36	47%	18/38
	Preterit, potentially with same form as present	0.30	40%	15/37
	Range	43		
	<i>p-value</i>	2.56×10^{-5}		
Type of subject	Expressed	0.65	67%	266/397
	Unexpressed	0.35	50%	51/103
	Range	30		
	<i>p-value</i>	8.08×10^{-6}		
Type of reference	Generic	0.58	68%	189/276
	Specific	0.42	57%	128/224
	Range	16		
	<i>p-value</i>	0.0154		
Sex	Female	[0.48]	66%	177/268
	Male	[0.52]	60%	140/232
	<i>p-value</i>	0.693 (>0.05)		

Log.likelihood: -250.076

Table 4. Cont.

Aceguá, Uruguay		N = 266	Overall proportion = 51%	Input = 0.36	
		Fixed Effects:	FW	%	n/N
Verbal tense and phonic salience	Present, potentially with same form as preterit		0.87	85%	57/67
	Gerund or infinitive		0.70	71%	5/7
	Preterit, with different form than present		0.65	50%	12/24
	Present, with different form than preterit		0.43	41%	47/114
	Preterit, potentially with the same form as present		0.31	30%	9/30
	Imperfect		0.09	25%	6/24
	Range		78		
	p-value		6.89×10^{-9}		
Discourse persistence	Preceded by <i>a gente</i>		0.89	85%	62/73
	First token		0.44	49%	64/130
	Preceded by <i>nós</i>		0.13	14%	9/63
	Range		76		
	p-value		1.19×10^{-10}		
Type of subject	Expressed		0.76	61%	110/180
	Unexpressed		0.24	29%	25/86
	Range		52		
	p-value		2.17×10^{-6}		
Age	15–25		[0.77]	79%	41/52
	31–49		[0.42]	53%	62/118
	>50		[0.30]	33%	32/96
	p-value		0.273 (>0.05)		
Sex	Female		[0.64]	62%	101/164
	Male		[0.36]	33%	34/102
	p-value		0.304 (>0.05)		
Type of reference	Generic		[0.54]	50%	74/147
	Specific		[0.46]	51%	61/119
	p-value		0.453 (>0.05)		

Log.likelihood: -103.915

The next factor that conditions the use of pronominal *a gente* in Aceguá, Brazil, is the speaker’s age. As Table 4 illustrates, young participants in Brazil clearly prefer the innovative *a gente* (0.75), while elderly speakers tend to maintain *nós* (0.31). This is a typical pattern where age stratification signals an ongoing change led by young speakers. In Uruguay, however, age was not found to be significant, although the distribution of the use of *a gente* follows the same tendency detected on the other side of border: young Uruguayans, similarly to young Brazilians, are the ones leading the linguistic innovation in the community. The lack of significance could be the result of the small size of the sample and the fact that this statistical analysis includes speakers as a random effect. As

more participants are added to future studies and *a gente* becomes more widespread in the community, age may become statistically significant in Aceguá, Uruguay. The next factor Rbrul identified for Brazilian Portuguese was verbal tense and phonic salience, which is the most influential factor in Uruguayan Portuguese. While both dialects are clearly influenced by this variable, there are some differences in the constraint ranking across the varieties, and they are summarized in Table 5.

Table 5. Ranking the impact of tense and phonic salience factors on *a gente* use in Aceguá Portuguese.

Aceguá, Brazil		Aceguá, Uruguay	
Favoring			
Present, potentially with same form as preterit	0.73	Present, potentially with same form as preterit	0.87
Gerund or infinitive	0.60	Gerund or infinitive	0.70
<i>Imperfect</i>	0.56	<i>Preterit, with different form than present</i>	0.65
Disfavoring			
Present, with different form than preterit	0.44	Present, with different form than preterit	0.43
<i>Preterit, with different form than present</i>	0.36	Preterit, potentially with the same form as present	0.31
Preterit, potentially with the same form as present	0.30	<i>Imperfect</i>	0.09

The results in Table 5 show that both dialects respond strongly to the same top two factors within the verbal tense and phonic salience group: present, potentially with the same form as the preterit (less salient opposition), and gerund or infinitive (forms with no person marking), both in italics. This result indicates that *a gente* is entering both dialects through similar linguistic routes, both of which are attested in previous studies of Brazilian Portuguese. Two other linguistic contexts also show a similar impact on both dialects: present with a different form than the preterit (‘*vamos/fomos*’) and preterit potentially with the same form as present (‘*falamos/falamos*’). The present with a different form than the preterit slightly disfavors *a gente* in Brazilian (0.44) and Uruguayan (0.43) varieties of Portuguese, while the preterit, potentially with the same form as the present, strongly disfavors *a gente* in both dialects (0.30 and 0.31, respectively).

Yet, the comparison of factor constraints in Table 5 points to two cross-dialectal discrepancies: the role of the imperfect (‘*falava/falávamos*’) and of the preterit that has a different form from the present (‘*fomos/vamos*’). As previously explained, in a few varieties of Brazilian Portuguese, contexts with imperfect verb forms favor *a gente* as a strategy to avoid words with the stress on the antepenultimate syllable, usually reduced in vernacular speech (Mattos 2013; Benfca 2016; Foeger et al. 2017; Scherre et al. 2018a; among others). This is the case in Aceguá, Brazil, albeit not to a great extent (0.56), while in Uruguayan Portuguese, the opposite tendency is found, where the imperfect strongly disfavors *a gente* (0.31) and favors *nós*. A closer look at the imperfect tokens in both dialects reveals the data shown in Table 6a, which indicates that while proparoxytones are avoided in Aceguá, Brazil (fewer than 10% of the imperfect tokens take plural morphemes), this tendency is not replicated in Aceguá, Uruguay, where 50% of the imperfect tokens take plural morphemes. While this apparent divergence between Brazilian and Uruguayan Portuguese merits future research, it might explain why imperfect tense has minimal impact on Uruguayan Portuguese speakers’ choice of *a gente*.

Table 6. (a). Verbal morphemes used in imperfect verbal forms in Aceguá, Brazil, and Aceguá, Uruguay (without 12 categorial users of *nós*). (b). Verbal morphemes used in preterit, with different form than present in Aceguá, Brazil, and Aceguá, Uruguay (without 12 categorial users of *nós*).

(a)		
Structure	Aceguá, Brazil	Aceguá, Uruguay
<i>a gente</i> + singular morpheme	52% (29/56)	25% (6/24)
<i>nós</i> + singular morpheme	39% (22/56)	25% (6/24)
<i>nós</i> + plural morpheme	9% (5/56)	50% (12/24)
(b)		
Structure	Aceguá, Brazil	Aceguá, Uruguay
<i>a gente</i> + singular morpheme	47% (18/38)	50% (12/24)
<i>nós</i> + singular morpheme	Not attested	Not attested
<i>nós</i> + plural morpheme	53% (20/38)	50% (12/24)

The other factor where the constraining order differs is preterit with a different form than the present (*fomos/vamos*), which disfavors *a gente* in Brazilian Portuguese (0.36) but favors it in Uruguayan Portuguese (0.65). Brazilian Portuguese follows the tendency found in other varieties to use *nós* with plural agreement in forms where this difference is highly salient (e.g., *nós fomos*, ‘we went’). Uruguayan Portuguese, on the hand, tends to use *a gente* with singular agreement, also avoiding verbal non-agreement (e.g., *a gente foi*, ‘we went’). Therefore, both varieties in the samples tend to show standard verbal agreement with both variants (*nós fomos* and *a gente foi*). A further analysis that investigates the details of preterit forms is in order so that the reason why border Uruguayan Portuguese does not follow the general tendency seen in Brazilian Portuguese and other varieties of Brazilian Portuguese is clarified. Nevertheless, the results in Table 6b show that both dialects behave similarly in their preference for standard verbal agreement:

In summary, aside from the discrepancies in the phonic salience and verbal tense group, the cross-dialectal comparison of constraint ranking shows that the grammars are strikingly consistent, especially given that the incorporation of *a gente* into Uruguayan Portuguese is incipient and infrequent, totaling only 29% of the total tokens when categorial users of *nós* are included.

The next factor identified for both communities by Rbrul is the type of subject. In both varieties, an expressed pronominalized *a gente* favors more *a gente* at a probability rate of 0.65 for Brazilian Portuguese and 0.76 for Uruguayan Portuguese, while an expressed pronoun favors *nós*. The fact that the direction of the constraint ranking is identical demonstrates that, like discourse persistence, subject expression influences both dialects in the same way. The results for both factor groups together reveal that, even though *a gente* is more frequently used in Brazilian Portuguese (Tables 2 and 3), the variable grammars behind first-person singular pronouns follow similar patterns, which leads us to conclude that both communities share similar grammars. Notably, the linguistic contexts where this change is being favored match tendencies in other dialects of Brazilian Portuguese, since subject expression and discourse persistence top the factors found to condition *a gente* (e.g., Mattos 2013; Foeger 2014; Benfica 2016).

The final factor identified by Rbrul is type of reference. Speakers in Aceguá, Brazil, follow the tendency found in previous studies of Brazilian Portuguese: generic references slightly favor *a gente* (0.58), presumably a vestige of the semantics from its lexical source, ‘the people’, while a specific reference slightly favors the use of *nós*. Interestingly, the type of reference did not achieve significance in the sample from Uruguayan speakers, even though a similar distribution points to the same tendency. Finally, sex was not significant in either speech community. Although women commonly lead linguistic changes, this tendency was not seen in the present analysis.

In summary, a comparison of the independent variables influencing the realization of *a gente* instead of *nós* in the Portuguese varieties spoken across the Brazilian–Uruguayan border shows that both varieties present very similar variable grammars, illustrating clear continuities across the dialects. On both sides of the national border, *a gente* is triggered by discourse persistence, verbal tense and phonic salience, and type of subject, indicating that this linguistic innovation is spreading into the Portuguese pronominal system following similar linguistic routes. While the factor constraints for discourse persistence and type of subject show that the use of *a gente* trends in the same direction for both dialects (mirroring tendencies found in previous studies of Brazilian Portuguese), the order of constraints within the verbal tense and phonic salience group reveals two differential impacts pertaining to imperfect and preterit with a different form than the present. First, the imperfect favors *a gente* in Brazilian Portuguese but favors *nós* in Uruguayan Portuguese. We interpret this difference as being based on greater use of (and thus less resistance to) proparoxytones such as *falávamos* (‘we spoke’) in Uruguayan Portuguese. The second difference—namely, that preterit, with a different form than the present disfavors *a gente* in border Brazilian Portuguese, unlike in Uruguayan Portuguese and other varieties of Brazilian Portuguese—is harder to interpret and needs further investigation. In addition, two factors explain some of the variance in the Brazilian Aceguá dialect, but not in its Uruguayan counterpart: age and type of reference. We interpret these differences as signaling that the inclusion of *a gente* in Uruguayan Portuguese is more incipient, and we predict that, as the variant spreads more widely, both factors may become relevant to the distribution of *a gente* across the linguistic system and the community.

Lastly, in Table 7, we present a similar multivariate analysis, but one that includes both corpora and adds ‘community’ as an additional factor. Doing so enables us to determine whether community is a significant predictor of the expression of *a gente*. If this is the case, it would provide evidence of significant cross-dialectal differences and lack of convergence between the two grammars. If, on the other hand, community is insignificant, this would be yet another sign that the communities in fact behave as one dialectal area that shares the same variable grammar.

Table 7. Pronominal *a gente* vs. *nós* in subject positions in Portuguese from Aceguá, Brazil, and Uruguay (without 12 categorical users of *nós*; random effects: individual).

Aceguá, Brazil-Uruguay		N = 766	Overall Proportion = 59%	Input = 0.53	
		Fixed Effects:	FW	%	
				n/N	
Discourse persistence	Preceded by <i>a gente</i>		0.84	84%	174/207
	First token		0.38	58%	238/408
	Preceded by <i>nós</i>		0.24	26%	40/151
	Range		60		
	<i>p-value</i>		1.09×10^{-18}		
Age	15–25		0.73	82%	154/187
	31–49		0.43	59%	201/342
	>50		0.32	41%	97/237
	Range		41		
	<i>p-value</i>		0.00637		
Verbal tense and phonic salience	Present, potentially with the same form as preterit		0.76	83%	198/238
	Gerund or infinitive		0.61	60%	15/25
	Imperfect		0.42	44%	35/80
	Present, with different form than preterit		0.43	51%	149/294

Table 7. Cont.

Aceguá, Brazil-Uruguay		N = 766	Overall Proportion = 59%	Input = 0.53
Verbal tense and phonic salience	Preterit, potentially with same form as present	0.29	36%	24/67
	Range	47		
	p-value	4.86×10^{-12}		
Type of subject	Expressed	0.69	65%	376/577
	Unexpressed	0.31	40%	76/189
	Range	38		
	p-value	6.03×10^{-10}		
Type of reference	Generic	0.56	62%	263/423
	Specific	0.44	55%	189/343
	Range	12		
	p-value	0.024		
Community	Brazilian	[0.56]	63%	317/500
	Uruguayan	[0.44]	51%	135/266
	p-value	0.185 (>0.05)		
Sex	Female	[0.50]	64%	278/432
	Male	[0.50]	52%	174/334
	p-value	0.983 (>0.05)		

Log.likelihood: -365.414

Table 7 shows that, when the two communities are merged in the same statistical run, the same linguistic factor groups are selected, and the same ranking of factors is maintained. Discourse persistence, verbal tense and phonic salience, type of subject, and type of reference are significant, similar to the results in Table 4. Age is also significant, as expected if a linguistic change is in progress: the youngest group is most likely to use *a gente*. Importantly, community does not reach significance, countering the hypothesis that the odds of using *a gente* would be lower among bilinguals, despite the difference in overall frequencies. Thus, there is no statistical support for dialect-specific linguistic behavior in the use of pronominalized *a gente* in the larger border area of Aceguá.

4. Conclusions

In summary, following the premises of comparative sociolinguistics, we compared the distribution of the *nós/a gente* variable on both sides of the Brazil–Uruguay border. We found that Uruguayans still prefer *nós* and use *a gente* less often than Brazilians, who prefer the grammaticalized form more than half the time. Yet, it is clear that *a gente* has not only reached speakers in small towns in the south of Brazil, but has also crossed the border in Uruguay and is currently establishing itself alongside *nós* there. This represents a clear development since Elizaincín et al.’s (1987) study, where the authors attested only *nós* in Uruguayan Portuguese. A multivariate analysis shows that *a gente* is entering both communities in largely the same linguistic contexts, which are parallel to tendencies found for other dialects of Brazilian Portuguese, since verbal tense and phonic salience, discourse persistence, type of subject, and type of reference are linguistic constraints commonly found to condition this variable in other areas of Brazil (e.g., Omena 1996a; Lopes 2003; Zilles 2005; Scherre et al. 2018a). In terms of extra-linguistic factors, in both communities, the younger group of speakers shows a tendency to use *a gente*. However, this tendency has not reached statistical significance in Uruguayan Portuguese yet, which could indicate that this change is more incipient, or less advanced, among bilinguals.

In addition, these results support Carvalho’s argument (Carvalho 2003, 2004, 2014, 2016) for dialect leveling in the region due to the recent urbanization of Uruguayan Portuguese. On one hand, the incorporation of *a gente* is less frequent in the speech of Aceguá,

Uruguay, compared to its counterpart in Aceguá, Brazil, in line with previous claims that Uruguayan Portuguese shows conservative traits when compared to Brazilian Portuguese. On the other hand, the variable distribution is replicated in both communities, which indicates the continued sociolectal extension into Uruguayan Portuguese of Brazilian Portuguese features.

Author Contributions: Conceptualization, C.P., A.C. and M.P.S.; data collection, C.P.; data analysis, C.P., A.C. and M.P.S. original draft preparation, C.P., A.C. and M.P.S.; visualization, C.P., A.C. and M.P.S. All authors have read and agreed to the published version of the manuscript.

Funding: Maria Marta Pereira Scherre receives research funds from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of Brasilia, Brazil.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data for this study is protected and unavailable to the public.

Acknowledgments: We thank the participants of this study for offering us their time and sharing their stories with us.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

- ¹ Extracted from *Projeto SP 2010* (SO 2012-045-M41DPO interview).
- ² Extracted from Carvalho (1998) corpus of Rivera Spanish (003 interview).
- ³ All examples henceforth were extracted Pacheco (2014) corpus of Uruguayan Portuguese.

References

- Benfica, Samine de Almeida. 2016. A Concordância verbal na fala de Vitória. Master's dissertation, Universidade Federal do Espírito Santo, Vitória, Spain.
- Borges, Paulo R. S. 2004. A Gramaticalização de "a gente" no português brasileiro. Ph.D. dissertation, Programa de Pós-Graduação em Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, Spain.
- Carvalho, Ana M. 1998. The Social Distribution of Uruguayan Portuguese in a Bilingual Border Town. Ph.D. dissertation, University of California, Berkeley, CA, USA.
- Carvalho, Ana M. 2003. Rumo a uma definição do português uruguaio. *Revista internacional de Linguística Iberoamericana* 2: 125–49.
- Carvalho, Ana M. 2004. I Speak Like the Guys on TV: Palatalization and the Urbanization of Uruguayan Portuguese. *Language Variation and Change* 162: 127–51. [CrossRef]
- Carvalho, Ana M. 2014. Sociolinguistic Continuities in Language Contact Situations: The Case of Portuguese in Contact with Spanish along the Uruguayan-Brazilian Border. In *Portuguese/Spanish Interfaces*. Edited by Patricia Amaral and Ana M. Carvalho. Philadelphia: John Benjamins, pp. 263–94.
- Carvalho, Ana M. 2016. The Analysis of Languages in Contact: A Case Study through a Variationist Lens. *Cadernos de Estudos Linguísticos* 58: 401–24. [CrossRef]
- Carvalho, Ana M. 2021. Differential object marking in monolingual and bilingual Spanish. In *The Routledge Handbook of Variationist Approaches to Spanish*. Edited by Manuel Díaz-Campos. New York: Routledge, pp. 393–410.
- Carvalho, Ana M., and Joseph Kern. 2019. On the Permeability of Tag Questions in the Speech of Spanish-Portuguese Bilinguals. *Pragmatics* 29: 463–92. [CrossRef]
- Carvalho, Ana M., and Michael Child. 2011. Subject Pronoun Expression in a Variety of Spanish in Contact with Portuguese. In *Selected Papers of the Fifth Workshop in Spanish Sociolinguistics*. Edited by Jim Michnowicz. Summerville: Cascadilla Press. Available online: <http://www.lingref.com/cpp/wss/5/paper2502.pdf> (accessed on 2 October 2023).
- Carvalho, Ana M., and Ryan M. Bessett. 2015. Subject Pronoun Expression in Spanish in Contact with Portuguese. In *Subject Pronoun Expression in Spanish: A Cross-Dialectal Perspective*. Edited by Ana M. Carvalho, Rafael Orozco and Naomi R. Shin. Washington: Georgetown University Press, pp. 143–66.
- Castañeda, Rosa M. 2011. Linguistic Variation in a Border Town: Palatalization of Dental Stops and Vowel Nasalization in Rivera. Ph.D. dissertation, University of Florida, Gainesville, FL, USA.
- Castañeda, Rosa M. 2016. The Sociolinguistic Evolution of a Sound Change. *Journal of Portuguese Linguistics* 15: 3. [CrossRef]

- Córdoba, Alexander S. 2013. A Neutralização das vogais postônicas finais do português uruguaio falado na cidade de Tranqueras, Uruguai. Master's dissertation, Universidade Católica de Pelotas, Pelotas, Brazil.
- Elizaincín, Adolfo. 1992. *Dialectos en contacto: Español y portugués en España y América*. Montevideo: Arca.
- Elizaincín, Adolfo, Luis E. Behares, and Graciela Barrios. 1987. *Nós falemo brasileiro. Dialectos portugueses en el Uruguay*. Montevideo: Amesur.
- Foeger, Camila C. 2014. A primeira pessoa do plural no português falado em Santa Leopoldina. Master's dissertation, Universidade Federal do Espírito Santo, Vitória, Spain.
- Foeger, Camila C., Lilian C. Yacovenco, and Maria Marta Pereira Scherre. 2017. A primeira pessoa do plural em Santa Leopoldina/ES: Correlação entre alternância e concordância. *Letrônica* 10: 5–17. [CrossRef]
- Garrido Meirelles, Virginia. 2009. O português da fronteira Uruguai-Brasil. In *Português em contato*. Edited by Ana M. Carvalho. Madrid: Vervuert Iberoamericana, pp. 257–75.
- Hensey, Frederick G. 1972. *The Sociolinguistics of the Brazilian Uruguayan Border*. The Hague: Mouton.
- Hensey, Frederick G. 1982. Spanish, Portuguese and Fronteiriço: Languages in contact in Northern Uruguay. *International Journal of the Sociology of Language* 34: 9–23. [CrossRef]
- Johnson, Daniel E. 2009. Getting off the GoldVarb Standard: Introducing Rbrul for Mixed Effects Variable Rule Analysis. *Language and Linguistics Compass* 3: 359–83. [CrossRef]
- Lopes, Célia R. dos Santos. 2003. *A inserção de 'a gente' no quadro pronominal do Português*. Frankfurt am Main and Madrid: Vervuert/Iberoamericana, vol. 18.
- Mattos, Shirley Eliany Rocha. 2013. Goiás na primeira pessoa do plural. Ph.D. dissertation, Universidade de Brasília, Brasília, Spain.
- Mattos, Shirley Eliany Rocha. 2017. A primeira pessoa do plural na fala de goiás. *Revista (Con) Textos Linguísticos (UFES)* 11: 145–66.
- Mendonça, Alexandre Kronemberger de. 2010. NÓS e A GENTE em Vitória: Análise sociolinguística da fala capixaba. Master's dissertation, Universidade do Espírito Santo, Vitória, Spain.
- Muniz, Cleuza A. G. 2008. Nós e A gente: Traços sociolinguísticos no assentamento. Master's dissertation, Universidade Federal do Mato Grosso, Campo Grande, Spain.
- Naro, Anthony J., Edai Görski, and Eulália Fernandes. 1999. Change without Change. *Language Variation and Change* 11: 197–211. [CrossRef]
- Naro, Anthony J., Maria Marta Pereira Scherre, Camila Foeger, and Samine de Almeida Benfica. 2017. Linguistic and Social Embedding of Variable Concord with 1st Plural Nós 'We' in Brazilian Portuguese. In *Studies on Variation in Portuguese*, 1st ed. Edited by Pilar Barbosa, Maria da Conceição de Paiva and Celeste Rodrigues. Amsterdam: John Benjamins, pp. 219–31.
- Neves, Maria Helena de Moura. 2008. Pronomes. In *Gramática do português culto falado no Brasil*. Edited by Ataliba Teixeira de Castilho, Rodolfo Ilari and Maria Helena de Moura Neves. Campinas: Editores UNICAMP.
- Omena, Nelize Pires de. 1986. A referência variável da primeira pessoa do discurso do plural. In *Relatório final de pesquisa: Projeto Subsídios do Projeto Censo à Educação*. Edited by Anthony J. Naro. Rio de Janeiro: UFRJ, pp. 286–319.
- Omena, Nelize Pires de. 1996a. A referência à primeira pessoa do discurso no plural. In *Padrões sociolinguísticos: Análises de fenômenos variáveis do português falado na cidade do Rio de Janeiro*. Edited by Giselle Machline de Oliveira e Silva and Maria Marta Pereira Scherre. Rio de Janeiro: Tempo Brasileiro, pp. 183–215.
- Omena, Nelize Pires de. 1996b. As influências sociais na variação entre nós e a gente na função de sujeito. In *Padrões sociolinguísticos: Análises de fenômenos variáveis do português falado na cidade do Rio de Janeiro*. Edited by Giselle Machline de Oliveira e Silva and Maria Marta Pereira Scherre. Rio de Janeiro: Tempo Brasileiro, pp. 309–23.
- Omena, Nelize Pires de. 2003. A referência à primeira pessoa do plural: Variação ou mudança? In *Mudança linguística em tempo real*. Edited by Maria da Conceição de Paiva and Maria Eugênia Lamoglia. Rio de Janeiro: Contra Capa Livraria, pp. 63–80.
- Pacheco, Cíntia da Silva. 2013. Primeiras reflexões sobre o português fronteiriço de Aceguá. In *Variação linguística: Contato de Línguas e Educação*. Edited by Caroline Rodrigues Cardoso, Maria Marta Pereira Scherre and Heloísa Maria Moreira Lima-Salles. Cíntia Pacheco. Contribuições do III encontro do grupo de estudos avançados de sociolinguística da Universidade de Brasília. Campinas: Pontes Editores, pp. 187–207.
- Pacheco, Cíntia da Silva. 2014. Alternância NÓS e A GENTE no português brasileiro e no português uruguaio da fronteira Brasil-Uruguai (Aceguá). Ph.D. dissertation, Universidade de Brasília, Brasília, Spain.
- Pacheco, Cíntia da Silva. 2017a. Como definir o falar da fronteira Brasil-Uruguai? *(Con)textos linguísticos* 11: 9–26.
- Pacheco, Cíntia da Silva. 2017b. Contato linguístico na fronteira de Aceguá (Brasil-Uruguai). *Papia* 27: 109–33.
- Pacheco, Cíntia da Silva. 2017c. Identidade sociolinguística na fronteira de Aceguá (Brasil-Uruguai). *Revista de Estudos da Linguagem* 25: 276–304. [CrossRef]
- Pacheco, Cíntia da Silva. 2018. A diacronia e a sincronia dos pronomes de primeira pessoa do plural NÓS e A GENTE no português brasileiro e no português uruguaio. *Revista de Estudos da Linguagem* 1: 1–33. [CrossRef]
- Pacheco, Cíntia da Silva. 2020. Contextualização sócio-histórica da fronteira Brasil-Uruguai. *Revista de Estudos e Pesquisas Sobre as Américas* 14: 257–93.
- Pacheco, Cíntia da Silva, Ana M. Carvalho, and Maria Marta Pereira Scherre. 2018. When a new pronoun crosses the border: The spread of a gente on the Brazilian-Uruguayan frontier. In *47 New Ways of Analyzing Variation—NWAV*. New York: New York University.

- Paiva, Maria da Conceição, and Maria Marta Pereira Scherre. 2022. Revisitando o efeito da repetição na variação linguística. In *Contribuições para a linguística brasileira: Uma homenagem a Dinah Callou*, 1st ed. Edited by Josane Moreira de Oliveira, Jacyra Andrade Mota and Regiane Coelho Pereira Reis. Campo Grande: Editora da UFMS, vol. 1, pp. 259–77. Available online: <https://repositorio.ufms.br/handle/123456789/5811> (accessed on 30 November 2023).
- Pedemonte, J. C. 1985. Assembléia Geral. *Diario El País*.
- Poplack, Shana. 1980. The Notion of the Plural in Puerto Rico Spanish: Competing Constraints on /s/ Deletion. In *Locating Language in Time and Space*. Edited by William Labov. Philadelphia: University of Pennsylvania Press, pp. 55–67.
- Poplack, Shana, and Stephen Levey. 2010. Contact-Induced Grammatical Change: A Cautionary Tale. In *Language and Space: An International Handbook of Linguistic Variation*. Edited by Peter Auer and Jürgen Erich Schmidt. Berlin: De Gruyter, pp. 391–419.
- Rona, José P. 1965. *El dialecto “fronterizo” del norte del Uruguay*. Montevideo: Adolfo Linardi.
- Rubio, Cassio F. 2012. Padrões de concordância verbal e de alternância pronominal no português brasileiro e europeu: Estudo sociolinguístico comparativo. Ph.D. dissertation, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, Spain.
- Scherre, Maria Marta Pereira. 1998. Paralelismo linguístico. *Revista de Estudos da Linguagem* 7: 29–59. Available online: <http://periodicos.letras.ufmg.br/index.php/relin/article/view/2293/2242> (accessed on 25 December 2023). [CrossRef]
- Scherre, Maria Marta Pereira, and Anthony J. Naro. 1991. Marking in Discourse: “Birds of a Feather”. *Language Variation and Change* 3: 23–32. [CrossRef]
- Scherre, Maria Marta Pereira, Anthony J. Naro, and Lilian C. Yacovenco. 2018a. Nós e a gente em quatro amostras do português brasileiro: revisitando a escalada da saliência fônica. *Diadorim* 20: 428–57. [CrossRef]
- Scherre, Maria Marta Pereira, Lilian C. Yacovenco, and Anthony J. Naro. 2018b. Nós e a gente no português brasileiro: Concordâncias e discordâncias. *Estudos de Linguística Galega* 1: 13–27. [CrossRef]
- Souza, Maria Helena Menezes. 2020. A variação nós e a gente na posição de sujeito na comunidade quilombola Serra das Viúvas/Água Branca—AL. Master’s dissertation, Universidade Federal de Alagoas, Maceió, Spain.
- Tagliamonte, Sali A. 2006. *Analysing Sociolinguistic Variation*. Cambridge, UK: Cambridge University Press.
- Tagliamonte, Sali A. 2013. Comparative sociolinguistics. In *The Handbook of Language Variation and Change*. Edited by J. K. Chambers and Natalie Schilling. New York: John Wiley & Sons, pp. 128–56.
- Vianna, Juliana Barbosa de Segadas. 2011. Semelhanças e diferenças na implementação de a gente em variedades do português. Ph.D. dissertation, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Spain.
- Vianna, Juliana Barbosa de Segadas, and Célia R. dos Santos Lopes. 2015. Variação dos pronomes “nós” e “a gente”. In *Mapeamento sociolinguístico do português brasileiro*. Edited by Marco Martins and Jussara Abraçado. São Paulo: Contexto, pp. 109–31.
- Weiner, E. Judith, and William Labov. 1983. Constraints on the Agentless Passive. *Journal of Linguistics* 19: 29–58. [CrossRef]
- Zilles, Ana M. S. 2005. The Development of a New Pronoun: The Linguistic and Social Embedding of A Gente in Brazilian Portuguese. *Language Variation and Change* 17: 19–53. [CrossRef]
- Zilles, Ana M. S. 2007. O que a fala e a escrita nos dizem sobre a avaliação social do uso de a gente? *Letras de Hoje* 42: 27–44.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Reanalyzing Variable Agreement with *tu* Using an Online Megacorporpus of Brazilian Portuguese

Scott A. Schwenter *, Lauren Miranda, Ileana Pérez and Victoria Cataloni

Department of Spanish and Portuguese, The Ohio State University, Columbus, OH 43210, USA; miranda.137@osu.edu (L.M.); perez.641@osu.edu (I.P.); cataloni.1@osu.edu (V.C.)

* Correspondence: schwenter.1@osu.edu

Abstract: We reanalyze the phenomenon of verbal (non)agreement with the 2SG *tu* in a megacorporpus of Brazilian Portuguese compiled from the web. Unlike previous research, which has analyzed sociolinguistic interview data and regional differences, we examine these data with a focus on the internal linguistic factors that constrain the variability. Our analysis of 4860 tokens of *tu* + verb reveals that non-agreement with the 3SG verb form is by far the most common pattern, 2SG agreement being relatively infrequent. Individual verb lexemes show highly distinct rates of (non)agreement. In addition, the specific tense/aspect/mood forms and main/auxiliary status are likewise significant factors affecting the variation. We conclude that future studies of this phenomenon should not ignore these internal linguistic factors. We situate our study within a group of other recent studies in Romance linguistics, which have found that individual verbal and constructional patterns can have diverse effects on morphosyntactic variation.

Keywords: Brazilian Portuguese; pronouns; variation; online corpus data; quantitative analysis

Citation: Schwenter, Scott A., Lauren Miranda, Ileana Pérez, and Victoria Cataloni. 2024. Reanalyzing Variable Agreement with *tu* Using an Online Megacorporpus of Brazilian Portuguese. *Languages* 9: 197. <https://doi.org/10.3390/languages9060197>

Academic Editor: Patrícia Amaral

Received: 4 November 2023

Revised: 15 March 2024

Accepted: 19 May 2024

Published: 28 May 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction: The Phenomenon

While Brazilian Portuguese (BP) is widely known for its widespread use of *você* to denote second-person singular (2SG) reference (Faraco 1996), there are regions of the country that still make use of the pronoun *tu*, either exclusively or in conjunction with *você*, in order to effect 2SG reference. Probably most famous is the use of *tu* in the South of Brazil, especially in the state of Rio Grande do Sul, where *você* is rarely used, neither as a subject nor a direct object (cf. Schwenter et al. 2018). In other regions, for example, in the North (e.g., Pará) and some of the Northeast states (e.g., Maranhão, Ceará) of Brazil, the two pronouns are used to differing extents. The variation is even found in the Southeast dialect of Rio de Janeiro, which Paredes Silva (2003) has characterized as a “return” of *tu* to the Carioca dialect.

Beyond the presence or absence of the pronoun *tu* in Brazil, there is also variability in BP regarding verbal agreement with *tu* between the normative and historical 2SG paradigm found in European Portuguese (where *tu* occurs invariably with 2SG agreement) and the “mixed” system that pairs *tu* with 3SG agreement. The latter is the pattern found in BP with the subject *você*, which derives from an originally third-person possessive NP used for 2SG reference (*vossa mercê* ‘your mercy’), which evolved to be the most widespread 2SG pronoun in BP. Indeed, most grammatical descriptions of BP (e.g., Perini 2002; Kato et al. 2022) overlook this variation completely and simply state that 2SG *tu* is found invariably with 3SG agreement in Brazil. A recent overview article on variable agreement in BP, both nominal and verbal (Mendes and Oushiro 2015), likewise makes no mention of the variation in agreement found with *tu*. More specialized studies (e.g., Scherre et al. 2015) have actually revealed considerable nuance in the situation, however, and identified regional variation in the patterns of (non)agreement. This variable phenomenon is evidenced in the following examples of present indicative (1), simple past (2), imperfect (3), and future subjunctive (4). The (a) versions in each case illustrate the normative 2SG conjugation (as found in European

Portuguese), while the (b) versions show the common BP pattern of non-agreement between *tu* and the 3SG verb conjugation.

(1a)	Tu	és (2SG)	
	You	be.PRES:2SG	
	'You are'		
(1b)	Tu	é (3SG)	
	you	be.PRES:3SG	
	'You are'		
(2a)	Tu	comeste (2SG)	
	you	eat.PRET:2SG	
	'You ate'		
(2b)	Tu	comeu	
	you	eat.PRET:3SG	
	'You ate'		
(3a)	Tu	escrevias (2SG)	
	you	write.IMPF:2SG	
	'You wrote/were writing'		
(3b)	Tu	escrevia (3SG)	
	you	write.IMPF:3SG	
	'You wrote/were writing'		
(4a)	(Se)	tu	achares (2SG)
	if	you	believe.FUTSUBJ.2SG
	'(If) you believe'		
(4b)	(Se)	tu	achar (3SG)
	if	you	believe.FUTSUBJ.3SG
	'(If) you believe'		

This variation between the 2SG and 3SG verb forms has traditionally been treated as a dialectal phenomenon, though all regions where *tu* is found seem to show at least a minimal amount of 2SG agreement. One issue with this research is that the data sets that have been analyzed in the many regions where *tu* is found are not uniform in size or in the methods of collection used, making it difficult to draw valid comparisons. One goal of this paper is to examine a much larger data set that allows for greater lexical and constructional variety, and also for the random sampling of the data. We hope that this will provide at least a partial model for future studies of the phenomenon, no matter what the source of the data or its manner of collection.

As stated above, previous studies that have investigated (non)agreement with *tu* have mainly concentrated on the regional distribution and frequencies of the variation between 2SG and 3SG agreement patterns. Rather large discrepancies in (non)agreement rates have been found between these studies, even in studies of the same region or city. Thus, to take one example from a city often cited as having a high rate of 2SG agreement, Loregian (1996) found 39% agreement in sociolinguistic interview data from Florianópolis and a similar rate (43%) in her own follow-up study (Loregian-Penkal 2004). However, a more recent study by Davet and Campos-Antoniassi (2014), using a different corpus of interviews, found only 14% 2SG agreement in Florianópolis. While the latter authors consider the possibility of change in progress (which we deem unlikely in corpora separated by only 20 years), another possibility could be that the varying results are due to interspeaker differences in the populations surveyed or are due to internal linguistic factors, such as the particular verbs that occur in the corpus, the tense/aspect/mood forms in which they appear, their functions as main verbs or auxiliaries, etc.

Beyond the geographical variability of (non)agreement, some have considered education levels (*nível de escolarização*) as an independent variable, with some researchers finding increasing amounts of 2SG agreement as education increases (e.g., Davet and Campos-Antoniassi 2014 in Florianópolis) but others finding higher levels of this agreement among speakers with lower education levels in other localities (e.g., Loregian-Penkal 2004 in Porto Alegre). The spoken versus written mode has also been analyzed in several studies, with

written language showing more 2SG agreement overall than spoken, but once again with widely varying rates depending on the study (Guimarães 1979; Loregian-Penkal 2004). However, in none of these prior studies has the effect of individual speaker variation been taken into account, and it is well known that individual speech patterns can heavily skew data sets. Indeed, when individual speakers are identifiable in the data, they should be treated as random effects in any variationist study (Baayen 2008; Johnson 2009); this has not been the case in prior studies of 2SG agreement in BP.

In their summary of the prior research on 2SG pronoun variation and (non)agreement in BP, Scherre et al. (2015; also Scherre and Duarte 2016) argue for a classification of six different pronominal systems for 2SG reference. They base the classification of these six systems on the following four factors, which can have positive or negative values. First, there is the possible presence of *tu* in the system, since there are major dialects or dialect regions (e.g., São Paulo, Belo Horizonte) in which *tu* does not play a role. Second, there is the frequency of *tu* in the system, which can range from rather sporadic in the dialects of Rio de Janeiro to the nearly exclusive usage of *tu* in the dialect of Porto Alegre. The third factor is the presence or absence of the canonical verb inflection *-s*, and the fourth factor is the average rate of overt 2SG agreement. The third factor is very limited in its application, since not all of the verb conjugations with *tu* are marked by *-s*. For instance, the simple past form is marked by *-ste*, as in *falaste* or *comeste*.

Scherre et al. (2015) go on to state that a “remarkable feature of the system” is non-overt agreement: “the lack of the overt 2nd person singular agreement mark *-s*”, which is not associated with any particular social stigma, a position corroborated by Souza and Chaves’ (2015) study of speaker evaluations of (non)agreement in Florianópolis. However, two of the factors that Scherre et al. (2015) use to distinguish 2SG pronominal systems are precisely related to 2SG verbal inflection, i.e., the presence or absence of 2SG agreement and the average rate of overt agreement with the pronoun *tu*. Thus, it would be more accurate, in our view, to simply state that 3SG agreement is always the most frequent variant, no matter what region is analyzed, and the rates of 2SG agreement with *tu* can vary by region.

The updated survey of 2SG pronoun research presented in Scherre et al. (2020) reduces the variation in (non)agreement considerably to the North region of Brazil (Pará, Maranhão), with low rates of agreement (<25%) in the region. The South region, specifically in Santa Catarina, shows the highest rates of agreement (just over 25%), while agreement in Rio Grande do Sul is sporadic, and lower than 5%.¹

Scherre et al. (2020, p. 274) conclude that more work is needed on the geographic spread of *tu* and of the patterns of (non)agreement in Brazil: “[P]ara que tenhamos um mapa ainda mais próximo da realidade, são necessárias e urgentes mais pesquisas no vasto território brasileiro, com o controle de, pelo menos, cinco possibilidades disponíveis no português brasileiro: *você, ocê, cê, tu* com concordância e *tu* sem concordância, com o controle rigoroso dos contextos sintáticos e das nuances interacionais” [‘For us to have a map that is closer to reality, more research is necessary and urgent in the vast Brazilian territory, with control of, at least, five possibilities available in Brazilian Portuguese: *você, ocê, cê, tu* with agreement and *tu* without agreement, with rigorous control of syntactic contexts and interactional nuances’—our translation]. While we of course recognize that more work along these lines is always useful to clarify the empirical reality of Brazilian Portuguese, it must be pointed out that the map created by Scherre et al. (2020) is the fruit of nearly 60 studies (Scherre et al. 2020, p. 270). In this paper, we present a distinct approach with a new data source that may help further clarify this variation, and especially the variation in verbal (non)agreement with the subject *tu*. We are pessimistic about the possibility of interview methods allowing for the “rigorous control of syntactic contexts”, as in the quote above (Scherre et al. 2020, p. 274). It is for this reason that we decided to shift the empirical focus in this study away from sociolinguistic interviews to a megacorporus of Brazilian Portuguese web data, which will permit us to examine the variation in question with greater control of several internal linguistic factors, which we enumerate below.

In the remainder of this paper, we present the results of the corpus analysis of verbal (non)agreement with *tu* in a random sample taken from the web-based megacorporus used for this study. As alluded to above, we are interested in whether there are differing patterns according to the verb lexeme, verb form (tense/aspect/mood), lexical vs. auxiliary verb, phonic salience (determined by the morphophonological differences between verb forms), frequency, etc. The next section presents our methods in more detail, including the independent variables that we coded for. Section 3 presents both the descriptive statistics of our data (Section 3.1) and the inferential statistical analysis (Section 3.2). We discuss the broader contributions of our study in Section 4.

2. Methods

To collect data for the study, we utilized the Brazilian National Domain (.br) of the Portuguese Web 2018 (ptTenTen18) corpus from the Sketch Engine family of corpora (<http://www.sketchengine.eu>; accessed on 10 April 2023; see Kilgarriff et al. 2014). Sketch Engine was developed by Lexical Computing CZ and functions as a corpus manager, where users can analyze authentic texts of billions of words, known as their text corpora. The Brazilian National Domain corpus alone contains 4.7 billion words from diverse online sources. Due to that diversity, we do not claim that the data analyzed are necessarily representative of either spoken or written BP. However, the data are representative of Brazilian Portuguese as it is found on the web, where both spoken and written texts are available. Moreover, we believe that the constraints we uncover in the analysis to come likely resemble those found in BP more generally, and these constraints can be uncovered more easily by using a diverse online megacorporus. For this study, we opted to use Sketch Engine because it allowed us to look at a broader variety of data from distinct registers, topics, and styles, and also examine the lexical and constructional differences between verbs and different verb forms, facets which previous studies using sociolinguistic interviews have not included in their analysis.

Using the data from the ptTenTen18 corpus, we aim to determine what linguistic factors license 2SG verbal agreement or non-agreement with the pronoun *tu* in these corpus data. Following research on the effect of verbal lexemes and their frequency on variable phenomena in other Romance varieties (e.g., Poplack et al. 2018), as well as the intuitions of several native BP speakers who speak a *tu*-dominant dialect, we hypothesize that individual verbs will show distinct rates of second singular (non)agreement. We also hypothesize, following Naro (1981) among others, that phonic salience will play a critical role in the rates of 2SG agreement and that there will be more agreement when there is higher phonic salience between the 2SG and 3SG variants, i.e., when there are greater morphophonological differences between the two forms.

To select the verbs for our analysis, we used a random number generator to select 10 of the 20 most frequent verbs in the corpus. These verbs were *ser*, 'to be', *estar*, 'be', *ir*, 'go', *ter*, 'have', *ficar* 'stay', *poder* 'to be able to', *achar*, 'think', *falar*, 'speak', *gostar* 'like', and *escrever*, 'write.' The other 10 verbs that we included in our analysis were randomly selected from the top 1000 most frequent verbs, excluding the top 20 verbs since these had already been included in the first round of verb selection. The less-frequent verbs included *abrir*, 'open', *aparecer*, 'appear', *chamar*, 'call', *crer*, 'believe', *escolher*, 'choose', *lembrar*, 'remember', *mudar*, 'change', *preferir*, 'prefer', and *sair*, 'to go out'. The reasoning behind this method was to attempt to include verbs of varying frequency among those used in the analysis, but at the same time knowing that less frequent verbs were likely to have lower numbers of usable tokens. As we detail below in Section 3.1, we ended up separating two distinct forms of one of these verbs (*estar*) in our analysis due to broadly diverging patterns of (non)agreement between these forms.

For each of the 20 verbs, we used Sketch Engine's built-in random sample generator to select 100 tokens of the verbs for four different finite forms (see below for explanation). We therefore attempted to collect at least 400 tokens (100 per finite form) for each of the 20 verbs. Verbs that were less frequent in the corpus did not always meet the 100-token

goal; therefore, for those verbs, we included all the tokens available in the corpus for each verb tense. In Sketch Engine's search interface, we used the corpus query language (CQL) to search for examples with the explicit co-occurrence of the subject pronoun *tu*; see the sample search results in Figure 1. Using the CQL, we controlled for the (non)agreement of the pronoun but not necessarily the use of the variant forms in general. This was due to the fact that the corpus does not always provide sufficient accessible context to determine the reference of third-person singular forms that, without a co-occurring subject pronoun, are ambiguous between 2SG (*você/tu/o senhor/a senhora*) and 3SG (*ele/ela*) reference. Note also that we found no cases of the inverse of what we were investigating, i.e., there are no examples in the corpus of a 2SG verb form co-occurring with a subject pronoun that normatively takes 3SG agreement. Thus, while both *tu falas* 'you speak' and *tu fala* occur, as well as *você fala*, the non-agreeing form **você falas* does not occur in the data (nor are we aware of it being used in BP).



Figure 1. Example search result from Sketch Engine.

For each of the tokens, we coded our dependent variable (agreement or non-agreement with the overt second-person singular pronoun *tu*) and seven independent variables, chosen on the basis of findings from prior studies on agreement in BP (e.g., Scherre et al. 2007) and also based on our hypotheses about what factors might constrain the patterns of (non)agreement. The dependent variable contrasted the normative form, which has agreement corresponding to the 2SG *tu*, with the non-agreeing 3SG verb form that Brazilian speakers often use instead of the normative 2SG variant. For our independent variables (see Table 1), we coded the verb lexeme in the form of the infinitive for each of the 20 verbs included in the study. We also included low, mid, and high phonic salience, which is the degree of phonological difference between the two possible forms (to be elucidated further in Section 3), an important variable that has been used in other studies on (non)agreement in BP since the seminal work of Naro (1981). In addition, we coded for frequency, specifically the frequency per million words of the verb lexeme in the corpus, hypothesizing that frequency could affect rates of (non)agreement. In the coding scheme, we also distinguished between main and auxiliary verb function, which we hypothesized could show differing rates of agreement—specifically, lower rates for auxiliaries and higher rates for main verbs, since the former tend to be more phonologically reduced and semantically bleached than the latter. We coded as well for the particular finite form of the verb, specifically the present, preterit, imperfect and future subjunctive forms for each verb, since these allowed us to determine different degrees of phonic salience. Finally, we included coding for the polarity of the sentence, i.e., affirmative or negative, as well as intervening words, that is, whether and how many words occurred between the pronoun *tu* and the finite verb in each token, which is another factor often included in studies of (non)agreement. Given the paucity of tokens with more than one intervening element between the subject pronoun and verb, in the end we only compared the binary contrast between zero and one intervening word. Collinearity between polarity and intervening words can arise if a negative word occurring

in the preverbal position between the subject pronoun *tu* and the verb is also counted as an intervening word. As a result, we chose not to include such cases of negation as an intervening word, thus guaranteeing the independence of the two categories. Intervening words in the data were therefore mainly limited to adverbs such as *já* ‘already’, *sempre* ‘always’, or clitic pronouns such as the (direct or indirect) object *me* ‘you’ or reflexive *te* ‘you’ (as in *tu te chama[s]* ‘you call yourself’).

Table 1. Independent variables and their values.

Independent Variables	Values
Verb Lexeme	verb (labeled by infinitive)
Phonic Salience	low, mid, high
Verb Lemma Frequency	tokens per million
Verb Type	main or auxiliary
Intervening Words	0 or 1
Polarity	affirmative, negative
Verb Form	present, preterit, imperfect, future subjunctive

3. Results

3.1. Descriptive Statistics

This section presents a description, using rates of (non)agreement, of the independent variables found to be significant in our inferential statistical analysis (see Section 3.2). In total, we extracted and coded 4860 tokens of *tu* with a conjugated verb from the Portuguese Web 2018 corpus. We performed a descriptive and inferential statistical analysis of these tokens in R (R Core Team 2023). The overall results of (non)agreement in this data set were 1514 cases of 2SG agreement (31.2%) versus 3346 cases (68.8%) of 3SG agreement (i.e., non-agreement with 2SG *tu*). This distribution is shown graphically in Figure 2.

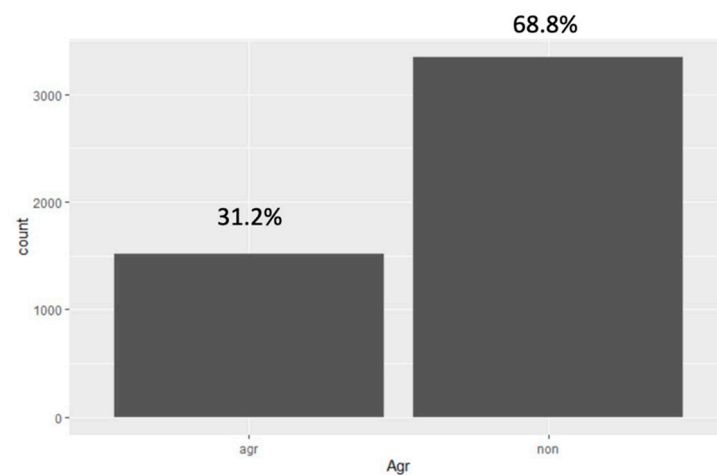


Figure 2. Overall distribution of (non)agreement in data set.

There were notable differences between the verbs under consideration. These are seen in Table 2, where each verb lexeme is presented, ordered by its frequency per million in the corpus, followed by its rates of agreement and non-agreement with the subject *tu* in our data. The relative frequency here stands for the rate per million words of occurrences of the verbal lexeme in the .br domain of the ptTenTen18 corpus. As can be seen, there is a wide range of variability by verb: the reduced form *tar* (<*estar*; see below for more discussion of this form) is the verb that shows the lowest rate of agreement at only 1.9%, in stark contrast to its unreduced counterpart *estar*, whose agreement rate of 63.2% is surpassed only by the verb *crer*, a lexeme which, in our data, is heavily restricted to religious contexts (BP uses the verbs *acreditar* and *achar* more commonly for ‘to believe’), at 65.3%.

Table 2. Overall rates of (non)agreement, by verb, ordered by frequency per million.

	<i>ser</i>	<i>estar</i>	<i>tar</i>	<i>poder</i>	<i>ter</i>	<i>fazer</i>	<i>ir</i>	<i>ficar</i>	<i>falar</i>	<i>achar</i>	<i>chamar</i>
Freq/million	17,426.63	4007.25	N/A	3217.78	2503.25	2258.97	1720.42	1212.86	606.29	412.89	377.94
Agr	150 (44.0%)	199 (63.2%)	4 (1.9%)	95 (36.3%)	109 (29.9%)	103 (28.7%)	95 (23.1%)	48 (19.1%)	64 (24.7%)	88 (18.6%)	88 (55.3%)
Non-Agr	191 (56.0%)	116 (36.8%)	204 (98.1%)	167 (63.7%)	255 (70.1%)	256 (71.3%)	317 (76.9%)	203 (80.9%)	195 (75.3%)	385 (81.4%)	71 (44.7%)

	<i>gostar</i>	<i>sair</i>	<i>lembrar</i>	<i>escrever</i>	<i>abrir</i>	<i>mudar</i>	<i>escolher</i>	<i>aparecer</i>	<i>preferir</i>	<i>crer</i>
Freq/million	330.87	323.6	265.65	236.05	220.9	205.93	198.46	189.27	93.25	92.05
Agr	51 (18.8%)	48 (19.1%)	51 (34%)	108 (48.9%)	15 (19.7%)	69 (39.7%)	49 (40.5%)	7 (10.3%)	19 (26.8%)	64 (65.3%)
Non-Agr	221 (81.2%)	203 (80.9%)	99 (66%)	113 (51.1%)	61 (80.3%)	105 (60.3%)	72 (50.5%)	61 (89.7%)	52 (73.2%)	34 (34.7%)

A visualization of these results is given in Figure 3, ordered from the greatest rates of 2SG non-agreement on the left to the greatest rates of 2SG agreement on the right. We see that only three verbs (*chamar*, *crer*, *estar* [excluding *tar*]) have rates of 2SG agreement over 50%. Most of the other verbs are well under that percentage and nearly half of them are at or under a rate of 25% agreement. Again, however, there is considerable variability among the verbs analyzed, and their behavior is far from uniform with respect to (non)agreement. This result suggests, therefore, that sociolinguistic studies on 2SG agreement that utilize interview data to determine rates of (non)agreement with *tu* could be heavily skewed by individual verb frequency, as well as the potential (lack of) diversity of verbs in the data analyzed. To take an obvious hypothetical example, a data set of informal BP conversation that includes copious amounts of a highly frequent verb lexeme in the 2SG like *ser*, but few others, would most likely have a very low overall rate of 2SG agreement, but this overall rate for the region or dialect in question could be artificially suppressed precisely due to the high frequency of *ser* in the data.

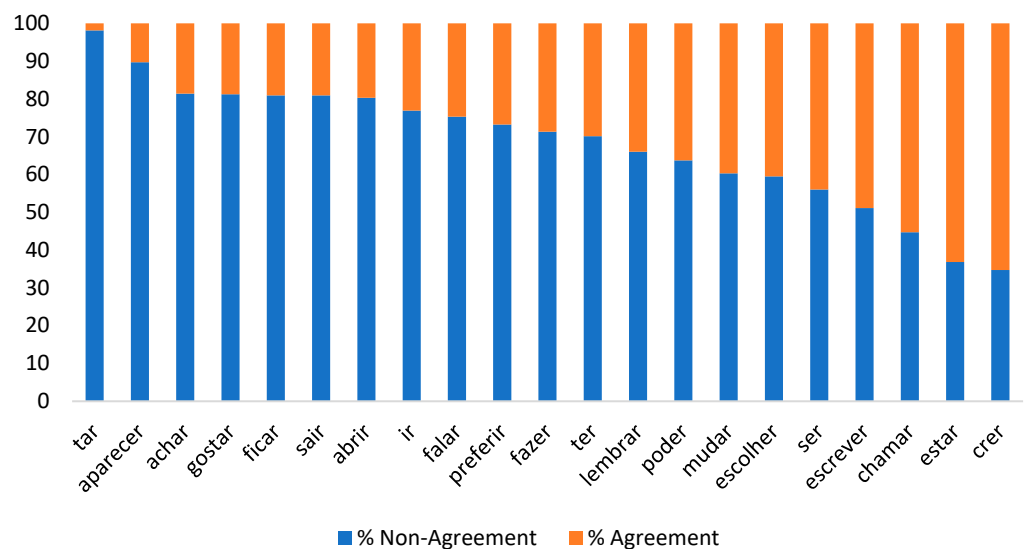


Figure 3. Rates of (non)agreement by verb.

Indeed, one of the most curious cases in our data comes from the stark contrast between the copular verb *estar* ‘to be’, originally included in the random sample of 20 verbs to be analyzed from the corpus, and its reduced variant *tar*, which turned out to be extremely frequent in our searches for the distinct forms of *estar*. In BP, *estar* is often reduced in speech and informal writing (e.g., on social media or informal chats between family or friends such as on WhatsApp) via the deletion of the initial syllable, as shown in the examples that follow.

- (5) Eu estou > Eu tô
'I am'
- (6) Tu estava(s) > Tu tava(s)
'You were'
- (7) Ela esteve > Ela teve
'She was'

When we analyzed these two verbs separately with respect to (non)agreement, we encountered a large discrepancy: 2SG agreement was the majority variant in the case of unreduced *estar* (63.2%) but there was virtually no such agreement (4/208 tokens or 1.9%) for reduced *tar*. See the full results in Table 3.

Table 3. Rates of (non)agreement by *estar* vs. *tar*.

	<i>Estar</i>	<i>Tar</i>
Agreement	199 (63.2%)	4 (1.9%)
Non-agreement	116 (36.8%)	204 (98.1%)

This contrast is, we believe, reflective of the normative view of agreement, which is mirrored by the use of the full, normative, bisyllabic (or more in some tense/aspect/mood conjugations) verb form. We should note, however, that the reduced form created several issues for our corpus searches, since some of the reduced forms deriving from *tar* become homophonous with (and are also homographs of) forms of completely different verbs. Thus, for example, the preterit *tu teve* 'you were' with a nonagreeing 3SG form (instead of 2SG *tu tiveste*) could have been a reduced form of *estar* (*teve* < *esteve*), or the 3SG preterit form of the verb *ter* 'to have'. A preliminary search of these preterit forms revealed that their tagging in the corpus was inconsistent and therefore we chose to exclude them from the analysis. For this reason, the number of tokens for *estar* (*n* = 315) and those for *tar* (*n* = 208) are not equivalent, and not directly comparable for each of their tense/aspect forms.

The results for the four different tense/aspect forms that were extracted from the corpus and analyzed for each of the 21 verbs are provided in Table 4. The rates of agreement follow the hierarchy Imperfect > Present > Past > Future Subjunctive. As we will discuss immediately below, this particular independent variable is highly correlated with that of phonic salience. In our inferential analysis (see Section 3.2 below), it was not advisable to include both of these factors due to their collinearity. Once again, the discrepancy in the totals for the different forms is due to the paucity of some forms (see e.g., future subjunctive) in collocation with *tu* in the corpus, a problem that does not arise for the more abundant present tense forms.

Table 4. Rates of (non)agreement by verb form.

	Present	Past	Imperfect	Future Subjunctive
Agreement	644 (32.2%)	384 (28.4%)	303 (37.7%)	183 (26.2%)
Non-agreement	1359 (67.8%)	970 (71.6%)	501 (62.3%)	516 (73.8%)

The results for the factor phonic salience (PS) are presented in Table 5. As mentioned in Section 2, the different levels of PS corresponded to different verb forms, here exemplified with the competing forms of the verb *falar* 'to talk, say'.

Table 5. Rates of (non)agreement by phonic salience.

	Low	Mid	High
Agreement	928 (34.3%)	202 (25.2%)	384 (28.4%)
Non-agreement	1776 (65.7%)	600 (74.8%)	970 (71.6%)

- Low Phonic Salience: Present (*tu fala/falas*) and Imperfect (*tu falava/falavas*)

- Mid Phonic Saliency: Future Subjunctive (*tu falar/falares*)
- High Phonic Saliency: Preterit (*tu falou/falaste*)

These distinctions were made based on the degree of difference between the 2SG and 3SG forms for each verb form. In the case of low PS, the only difference between the two forms is the addition of the morpheme *-s* in the 2SG form, while the 3SG form lacks that morpheme. For mid PS, the difference is found not only in the addition of the morpheme *-es* to make the 2SG form, but also in the additional syllable created by that morpheme when compared to the 3SG form. Finally, for high PS, there is a much greater change when comparing the past tense morphemes, *-ou* (or *-eu/-iu*) for 3SG versus *-aste* (or *-este/-iste*) for 2SG, in addition to the extra syllable that the 2SG morpheme adds to the word.

The overall rates for (non)agreement depending on Phonic Saliency show that low PS forms show higher rates of 2SG agreement than mid or high PS, a result that is contrary to prior results in the literature for 3PL verb forms (Naro 1981, among many others), where high PS forms showed higher rates of agreement than lower PS forms. We interpret this result as owing to the relative ease of converting the low PS form from 3SG to 2SG, and speakers’ recognition of *-s* as the normative marker of 2SG agreement for *tu*. While it may be true that Brazilians learn (but possibly forget) the full paradigm of 2SG in the formal educational system, the overall frequency of the low PS forms most likely gives speakers more familiarity with them than the preterit or future subjunctive forms (which, in the case of the latter, are considered to be more pedantic, according to several native speakers of BP *tu* dialects that we asked).

The next set of descriptive results we present here concerns the distinction between the main and auxiliary uses of verbs. As noted in grammaticalization studies by researchers such as Bybee et al. (1994) and Heine (1993), the reduction in semantic content in the diachronic shift from main (lexical) to auxiliary verb is paralleled by a reduction in the formal properties and other features, such as agreement, that are characteristic of lexical verbs. In the case of 2SG agreement, we hypothesized that auxiliary verbs would show lower rates of agreement than their main verb counterparts, since the principal lexical content is not conveyed by the auxiliary but by the main verb. As Tables 6 and 7 show, this hypothesis is true not only for the full data set (Table 6), but is also more clearly true for those verbs that have both main and auxiliary verb uses (Table 7), such as *estar*, *ter*, and *ir*.

Table 6. Rates of (non)agreement for main vs. auxiliary verb uses, full data set.

	Main	Auxiliary
Agreement	1271 (32.0%)	243 (27.3%)
Non-agreement	2699 (68.0%)	647 (72.7%)

Table 7. Rates of (non)agreement for main vs. auxiliary verb uses, only verbs with both uses.

	Main	Auxiliary
Agreement	390 (31.5%)	152 (23.8%)
Non-agreement	848 (68.5%)	487 (76.2%)

The factor Intervention distinguished between tokens where no words intervened between the subject pronoun *tu* and the following verb, and tokens where one word intervened between the subject and verb. As noted above, there were extremely few tokens containing more than one element intervening between the subject and the verb, and we made the resulting decision to limit this factor to a binary comparison. As can be seen in Table 8, there was considerably more agreement in the case of one intervening element than in the case of no intervening elements, which leads us to believe that there may be a constructional effect of subject pronoun *tu* + 3SG verb collocations (e.g., *tu come* ‘you eat’), which is not as robust when intervening elements such as temporal adverbs or clitic pronouns intervene between the subject and the verb (e.g., *tu sempre comes* ‘you always

eat’). Due to space limitations, we do not discuss this possibility further here, but it is undoubtedly an important issue for future research.

Table 8. Rates of (non)agreement by intervening elements (0 vs. 1).

	0 Intervening Elements	1 Intervening Element
Agreement	1299 (29.7%)	215 (45.5%)
Non-agreement	3088 (70.3%)	258 (54.5%)

Regarding the factor Polarity, there was slightly more non-agreement found in negative sentences (71.2%) than in affirmatives (68%), as seen in Table 9. The differences in the (non)agreement rates for Intervention in Table 8 versus the rates for Polarity in Table 9 provide further justification for our decision to separate these two factors in the analysis.

Table 9. Rates of (non)agreement by Polarity.

	Affirmative	Negative
Agreement	1399 (32.0%)	115 (28.8%)
Non-agreement	3062 (68.0%)	284 (71.2%)

The last set of descriptive statistics that we present in this section relates to the corpus frequency of the verbal lexemes included in our study in combination with the 2SG pronoun *tu*. Again, there were vast discrepancies in the frequencies per million words of the different verb forms analyzed, as seen above in Table 2, ranging from 17,426.63/million in the case of *ser* to only 92.05/million for *crer*, with the median frequency being 330.87/million, indicating the considerable rightward skew of verb frequency. Therefore, we opted to normalize the verb frequencies by log-transforming them in order to meet the necessary assumptions needed for valid statistical analysis. After the log transformation, a conditional inference tree was used to examine the effects of frequency on the data set overall, and it was determined that the verbs could be split into two groups, namely high and low frequency, where the high group included essentially those verbs at or above the median frequency and the low group those below. The overall rates of (non)agreement by high/low log frequency are given in Table 10; as can be seen, the rates were nevertheless very similar in both the high and low-frequency groups.

Table 10. Rates of (non)agreement by binary log frequency.

	High	Low
Agreement	803 (32.0%)	711 (30.3%)
Non-agreement	1709 (68.0%)	1637 (69.7%)

Summarizing this section, the descriptive results we have presented show the effects of different linguistic constraints on (non)agreement with *tu* in our data. There are clear patterns to this variation, but perhaps most importantly, we have corroborated our principal hypothesis, which is that individual verbs would display distinct rates of (non)agreement with *tu*. Some verbs have relatively high rates of 2SG agreement, but most do not, leading to an overall rate of agreement of around 30% for the random sample we analyzed. We turn now to the presentation and discussion of the results of our inferential statistical analysis.

3.2. Inferential Statistical Analysis

A mixed-effects logistic regression analysis in R was performed based on the output of a random forest and also by carrying out the step function in order to determine the variables with the greatest potential effect on the dependent variable. The random forest showed that Verb accounts for a substantial amount of the variation, and therefore should

be treated as a random effect (cf. Tagliamonte and Baayen 2012). Random effects are typically independent variables whose individual values cannot be exhausted in the data set, such as speaker in studies where a subset of speakers in a given community is included (Johnson 2009). In our case, it was clearly not possible to include all Portuguese verbs in the analysis, and as shown above, the verbs that we did include in our random sample vary greatly in terms of both their frequency per million and rates of (non)agreement. For these reasons, we included Verb as a random effect. In conjunction with the use of the step function to determine significant factors, we used the random forest to build a set of explanatory models using the step-up method, which included both the factors described above as fixed effects and Verb as a random effect. We then compared those models using the ANOVA function in R to determine the best-fit model for the data set, and checked the interactions using a conditional inference tree and by running additional regression models with interaction terms. Although the variable Form initially appeared to be a potentially significant factor based on the random forest calculation, it was ultimately excluded from our models due to its high degree of collinearity with PS (low PS = present and imperfect; mid PS = future subjunctive; high PS = preterit) and also due to the fact that PS appeared to have stronger effects on the variation, as indicated by the random forest and the results of the step function. The stronger effects of PS were also corroborated by a lower AIC value for the regression model that included PS compared to the model that included Form, thus indicating that the model better fit the data.

The best-fit regression model appears in Table 11, which summarizes the significance of each factor included in this model. To orient readers to our analysis, the values in the Estimate column refer to the likelihood of non-agreement with *tu*. A positive estimate value indicates higher rates of non-agreement (ergo, lower rates of agreement) between *tu* and its corresponding verb. A negative estimate indicates higher rates of agreement (ergo, lower rates of non-agreement).

Table 11. Best-fit logistic regression model output for verbal (non)agreement with 2SG *tu* (* = $p < 0.05$).

	Estimate	Std. Error	Z-Value	p-Value
(Intercept)	0.17242	0.26360	4.944	<0.001
PS (High)	0.33152	0.08051	4.118	<0.001 *
PS (Mid)	0.70859	0.10377	6.829	<0.001 *
Intervening (1 element)	-0.54970	0.11013	-4.991	<0.001 *
Main.aux (Main)	-0.44254	0.13753	-3.218	0.00129 *

For additional orientation, because we used contrast coding for the logistic regression, it should be noted that one variant of each fixed effect is omitted from Table 11 (e.g., PS (low) in the case of phonic salience). The likelihood of the fixed effects variants listed in the table affecting verbal (non)agreement is calculated in comparison to the variant *not* listed in the table. Thus, verbs with a high PS and mid PS both show significantly more non-agreement than verbs with a low PS. Tokens with one intervening element are compared to those with no intervening elements, and there is a statistically significant difference between these two values, such that the former show more agreement. Lastly, main verbs had significantly higher rates of agreement than auxiliary verbs.

To explore potential interactions in our data, we created conditional inference trees (cf. Tagliamonte and Baayen 2012), as shown in Figure 4, which includes only the significant predictors from the logistic regression illustrated in Table 11. As can be seen, there are no interactions between other factors and tokens with one intervening element ($n = 473$), which show significantly higher rates of agreement than the rest of the data set with no intervening material. A low degree of phonic salience interacts significantly with the Main vs. Aux status, such that auxiliary verbs show lower rates of 2SG agreement than main verbs. There is no interaction, however, between high/mid phonic salience and the distinction between Main vs. Aux status.²

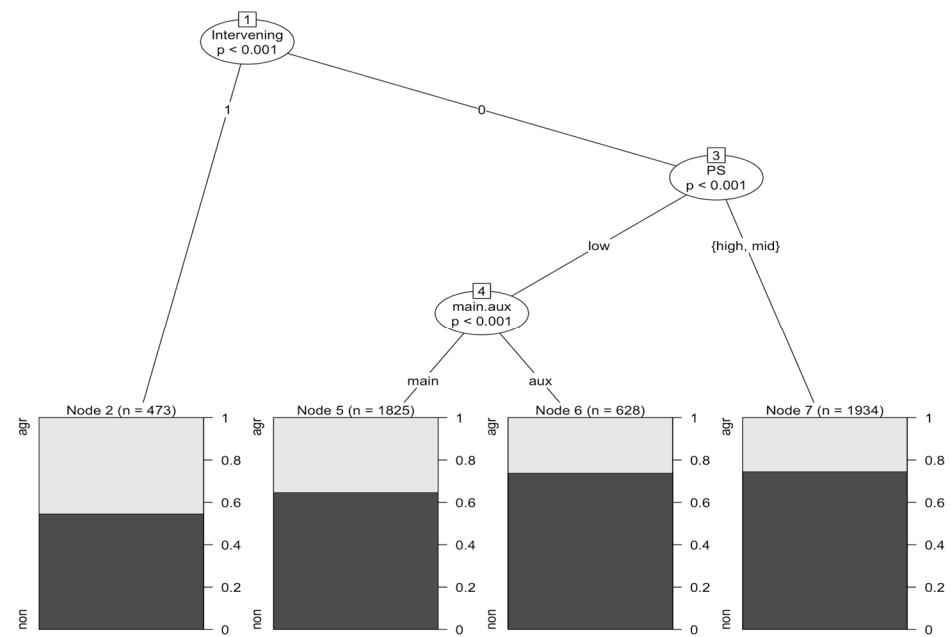


Figure 4. Conditional inference tree showing interactions between significant factors.

In this section, we have shown, using inferential statistical analysis, that (non)agreement with *tu* in our data is significantly constrained by several linguistic factors, the strongest of which, by far, is the individual verb lexeme in question. However, when we consider verb lexeme as a random effect in our models, other factors emerge as significant predictors of the variation: phonic salience, main vs. auxiliary verb status, and the presence/absence of intervening elements.

4. Discussion and Conclusions

In this article, we have shown, using a large random sample of naturally occurring data extracted from online sources in the Portuguese Web 2018 corpus, that 2SG agreement with the pronoun *tu* in this megacorporus is largely a lexically regulated phenomenon in BP, even though the overall rate of 2SG agreement is low (around 30%). Some verbs in our data are highly resistant to 2SG agreement, as shown for the particular case of *tar* (1.9%), the reduced version of the verb *estar* ‘to be’, which is a form typical of spoken BP or of written language that attempts to represent or mimic spoken styles. The unreduced version of this verb has a completely different behavior with regard to 2SG agreement (63.2%), insofar as it is typical of writing and other kinds of more careful linguistic production. This distinction, we believe, will be relevant to future studies using sociolinguistic interviews as their data source. Other verbs, such as *crer* ‘to believe’, which is highly restricted to religious contexts in our data and thus can also be assumed to be more careful in production, likewise show higher levels of agreement (65.3%). The rest of the verbs in our sample show rates ranging from a low of 10.3% in the case of *aparecer* ‘to appear’ to a high of 55.3% in that of *chamar* ‘to call’.

In addition to the individual verb rates, there are also significant constructional effects seen in different tense/aspect forms in our data set. While previous research (e.g., Naro 1981) found important differences in third-person plural forms with respect to greater phonic salience, in the case of the second-person forms analyzed here, agreement is more common in the forms that only require *-s* to make the 2SG, i.e., present indicative and imperfect indicative forms. This pattern contrasts with the findings (Naro 1981) for third-person plural forms where the more distinct forms in the paradigm were those that tended to conserve agreement more than forms that only, e.g., required the nasalization of a final vowel to mark plurality.

The main vs. auxiliary verb status is also a significant predictor of agreement in our data for verbs that show both types of uses, such that auxiliaries display significantly less agreement with *tu* than main verbs. This reflects the greater degree of grammaticalization of the auxiliary uses of the verbs in question, which leads to a further reduction in their phonological content. An important conclusion we draw from these findings is that future studies of variable 2SG agreement in BP, no matter what the source of the data, must take both lexical and constructional factors into account for a more complete explanation of the variation.

Lastly, we found significant effects for Intervention, the variation between a total lack of intervening elements between *tu* and the following verb and one element intervening between *tu* and the verb. When an element intervenes, there is significantly more 2SG agreement than when nothing intervenes, thus suggesting that there may be additional constructional effects of *tu* + verb-3SG that are reduced when an intervening element is present. This contrasting pattern is worthy of further research.

From a theoretical perspective, these results buttress the status of BP as a language “in which the primary function of personal pronouns is carried out by independent personal pronouns that occur as arguments” instead of by affixes that appear on the verb (Bhat 2004, p. 15). As is well known, BP has reduced the paradigm of verbal morphology greatly when compared to European Portuguese, such that the main distinction is now between 1SG forms and all the other person/number combinations (Azevedo 2005; Kato et al. 2022; Perini 2002). The overwhelming tendency for 2SG pronoun *tu* to co-occur with 3SG morphology (the same morphology that co-occurs with the 2SG pronoun *você* throughout Brazil) is another indication of this reduction in the complexity of the verbal conjugation paradigm and the severing of the link between pronouns and their erstwhile verbal morphology. This reduction in the verbal paradigm has also been tied to the growing obligatoriness of subject pronouns in BP (Tarallo 1996), which in that sense seems to be following the same path as French and English, two languages with near-obligatory subjects and a highly reduced verbal morphology. These languages also have a highly fixed SVX order, which is yet another characteristic that increases their resemblance to BP (Silva 2001).

This paper also contributes to recent work across languages showing that many variable phenomena show broad variation across different lexical types (such as verbs) or constructions (such as distinct tense/aspect/mood forms or main vs. auxiliary verb uses). For Romance languages, this has been shown most prominently for the choice of the indicative vs. subjunctive mood (Poplack et al. 2018), which is best considered a case of the lexical routinization of certain main clause verbal governors. In turn, individual languages or dialects (Schwenter and Hoff 2020 for Spanish) differ in their degree of conventionalizing these patterns of mood choice. Likewise, in Spanish, the choice between the past subjunctive forms in *-ra* and *-se* is also heavily restricted to a handful of the most frequent verbs in the language (Rosemeyer and Schwenter 2019). For Portuguese, the variability between alternate forms of past participles (*participios duplos*) for the same verbs (e.g., *pagado* vs. *pago* ‘paid’) has also been shown to be an overwhelmingly lexically regulated phenomenon. Schwenter et al. (2019) found that of 584 irregular participles in a corpus of over 1000 participles from both Brazilian and European Portuguese, three verbs (*pagar* ‘to pay’, *ganhar* ‘to win’, and *gastar* ‘to spend’) accounted for 64% (377/584) of all the irregular participles in their data set. Similar lexically and construction-specific findings for a much larger set of BP data were reported more recently by Dickinson (2022, 2024), who also found parallel patterns for Spanish. Our analysis in this paper advances evidence that 2SG (non)agreement should be included in this growing body of variable phenomena conditioned by verbal lexemes.

In our own future research, we plan to analyze, in more detail, the effects of topic and register in the updated Portuguese Web 2020 corpus, since we have already seen that there are clear differences to be found in religious contexts (reflected in our data by *crer* ‘to believe’), and the updated corpus now provides more options for selecting data by register, topic, or style. In addition, we hypothesize that persistence (aka priming) likely also has

strong effects in the data, which we did not code for in this study. These effects would be expected, given the low overall frequency of 2SG agreement, and in view of the general patterns of persistence in other low-frequency variants that are seemingly on their way to obsolescence (Rosemeyer and Schwenter 2019).³

Author Contributions: Conceptualization, S.A.S.; methodology, S.A.S. and L.M.; software, L.M.; validation, I.P. and V.C.; formal analysis, L.M.; investigation, V.C.; resources, I.P.; data curation, S.A.S. and L.M.; writing—original draft preparation, All; writing—review and editing, All; visualization, L.M. and I.P.; supervision, S.A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Interested parties can contact the first-named author via e-mail to receive a spreadsheet with the coded data set.

Acknowledgments: We thank Kendra Dickinson, Martín Fuchs, Mark Hoff, and two anonymous reviewers for their excellent comments, which improved this paper considerably, as well as the audiences at OSUCHILL 2023 and HLS 2023 at BYU for their useful feedback. We are also grateful to the other participants in Schwenter’s Portuguese 5611 class in Spring 2023 for helping with the initial stages of the project.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

- ¹ This summary of the (non)agreement rates across Brazil is consonant with folk ideas about where agreement with *tu* is found, especially with regard to the phenomenon in Maranhão, which is often considered by laypersons to be the Brazilian state where the “best” Portuguese is spoken (Bagno 2009).
- ² We created an additional conditional inference tree testing interaction between the significant factors seen in Figure 4 as well as Log Frequency (binary division) and Polarity. It appeared there was a boosting effect of Frequency on Polarity, such that only those verbs with high frequency showed a sensitivity to Polarity (cf. Erker and Guy 2012). There appeared to be a further interaction effect of Frequency and main vs. auxiliary verb, but given the low number of low frequency auxiliary tokens with high/mid PS ($n = 11$), this potential interaction appeared dubious. Further testing of these potential interactions in a regression model, however, revealed that they were not statistically significant.
- ³ It is not possible to access a large amount of the prior (or following) context in the Portuguese Web corpus (2018 or 2020 versions). However, there is a finite set of characters available for analysis before each target token, and this would provide an inherent limit on the distance between the prime (e.g., a prior token with 2SG agreement) and the target (a following token with 2SG agreement). While this limit on context is not ideal, it would at least offer a basis for consistent analysis across the full set of data (cf. Rosemeyer and Schwenter 2019).

References

- Azevedo, Milton M. 2005. *Portuguese: A Linguistic Introduction*. Cambridge: Cambridge University Press.
- Baayen, R. Harald. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge: Cambridge University Press.
- Bagno, Marcos. 2009. *Não é errado falar assim!* São Paulo: Parábola.
- Bhat, D. N. Shankara. 2004. *Pronouns*. Oxford: Oxford University Press.
- Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar*. Chicago: University of Chicago Press.
- Davet, Julie, and Paula Isaias Campos-Antoniassi. 2014. Variação na concordância verbal de segunda pessoa do singular—Um estudo de fala florianopolitana. *UFSC Working Papers em Linguística* 15: 95–111. [CrossRef]
- Dickinson, Kendra V. 2022. Past Participles in Spanish and Brazilian Portuguese: A Usage-Based Approach to Grammatical and Social Variation. Ph.D. dissertation, The Ohio State University, Columbus, OH, USA.
- Dickinson, Kendra V. 2024. Regularization and innovation: A usage-based approach to past participle variation in Brazilian Portuguese. *Languages* 9: 52. [CrossRef]
- Erker, Danny, and Gregory R. Guy. 2012. The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88: 526–57. [CrossRef]
- Faraco, Carlos. 1996. O tratamento “você” em português: Uma abordagem histórica. *Fragmenta* 13: 51–82. [CrossRef]
- Guimarães, Ana Maria Mattos. 1979. A Ocorrência de 2ª Pessoa: Estudo Comparativo Sobre o Uso de tu e Você na Linguagem Escrita. Ph.D. dissertation, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil.
- Heine, Bernd. 1993. *Auxiliaries: Cognitive Forces and Grammaticalization*. Oxford: Oxford University Press.

- Johnson, Daniel Ezra. 2009. Getting off the GoldVarb standard: Introducing RBrul for mixed-effects variable rule analysis. *Language and Linguistics Compass* 3: 359–83. [CrossRef]
- Kato, Mary, Ana Maria Martins, and Jairo Nunes. 2022. *Português Brasileiro e Português Europeu: Sintaxe Comparada*. São Paulo: Editora Contexto.
- Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý, and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1: 7–36. [CrossRef]
- Loregian, Loremi. 1996. *Concordância Verbal Com o Pronome tu na Fala do Sul do Brasil*. Master's thesis, Universidade Federal de Santa Catarina, Florianópolis, Brazil.
- Loregian-Penkall, Loremi. 2004. (Re)análise da Referência de Segunda Pessoa na Fala da Região Sul. Ph.D. dissertation, Universidade Federal do Paraná, Curitiba, Brazil.
- Mendes, Ronald Beline, and Lívia Oushiro. 2015. Variable number agreement in Brazilian Portuguese: An overview. *Language and Linguistics Compass* 9: 358–68. [CrossRef]
- Naro, Anthony Julius. 1981. The social and structural dimensions of a syntactic change. *Language* 57: 63–98. [CrossRef]
- Paredes Silva, Vera Lúcia. 2003. O retorno do *tu* ao falar carioca. In *Português Brasileiro: Contato Linguístico, Heterogeneidade e História*. Edited by Cláudia Roncarati and Jussara Abraçado. Rio de Janeiro: 7Letras, pp. 160–69.
- Perini, Mário. 2002. *Modern Portuguese Grammar*. New Haven: Yale University Press.
- Poplack, Shana, Rena Torres Cacoullos, Nathalie Dion, Rosane Berlinck, Salvatore Digesto, Dora LaCasse, and Jonathan Steuck. 2018. Variation and grammaticalization in Romance: A cross-linguistic study of the subjunctive. In *Manuals in Linguistics: Romance Sociolinguistics*. Edited by Wendy Ayres-Bennett and Janet Carruthers. Berlin and Boston: de Gruyter, pp. 217–52.
- R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available online: <https://www.R-project.org> (accessed on 14 March 2024).
- Rosemeyer, Malte, and Scott A. Schwenter. 2019. Entrenchment and persistence in language change: The Spanish past subjunctive. *Corpus Linguistics and Linguistic Theory* 15: 167–204. [CrossRef]
- Scherre, Maria Marta Peireira, and Maria Eugênia Lammoglia Duarte. 2016. Main current processes of morphosyntactic variation. In *The Handbook of Portuguese Linguistics*. Edited by W. Leo Wetzels, João Costa and Sergio Menuzzi. New York: Wiley and Sons, pp. 526–43.
- Scherre, Maria Marta Peireira, Anthony Julius Naro, and Carolines Rodrigues Cardoso. 2007. O papel do tipo de verbo na concordância verbal no português brasileiro. *DELTA* 23: 283–317. [CrossRef]
- Scherre, Maria Marta Peireira, Carolina Queiroz Andrade, and Rafael de Castro Catão. 2020. Redesenhando o mapa dos pronomes *tu/você/cê/ocê* no português brasileiro falado. In *Conquistas e Desafios dos Estudos Linguísticos na Contemporaneidade: Trabalhos do V Congresso Nacional de Estudos Linguísticos—V CONEL*. Edited by Pedro Henrique Witsch, Lucyenne Matos da Costa Vieira-Machado, Cláudia Jotto Kawachi Furlan and Mayara de Oliveira Nogueira. Porto Alegre: Editora Fi, pp. 270–76.
- Scherre, Maria Marta Peireira, Edilene Patrícia Dias, Carolina Queiroz Andrade, and Germano Ferreira Martins. 2015. Variação dos pronomes “*tu*” e “*você*”. In *Mapeamento Sociolinguístico do Português Brasileiro*. Edited by Marco Antonio Martins and Jussara Abraçado. São Paulo: Contexto, pp. 133–72.
- Schwenter, Scott A., and Mark Hoff. 2020. Cross-dialectal productivity of the Spanish subjunctive in nominal clause complements. In *Variation and Evolution. Aspects of Language Contact and Contrast across the Spanish-Speaking World*. Edited by Sandro Sessarego, Juan J. Colomina-Almiñana and Adrián Rodríguez-Riccelli. Amsterdam: John Benjamins, pp. 11–31.
- Schwenter, Scott A., Mark Hoff, Eleni Christodoulelis, Chelsea Pflum, and Ashlee Dauphinais. 2019. Variable past participles in Portuguese perfect constructions. *Language Variation and Change* 31: 69–89. [CrossRef]
- Schwenter, Scott A., Mark Hoff, Kendra V. Dickinson, Justin Bland, and Luana Lamberti. 2018. Experimental evidence for 2SG direct object pronoun preferences in Brazilian Portuguese. *Revista Lingüística* 14: 259–90. [CrossRef]
- Silva, Gláucia V. 2001. *Word Order in Brazilian Portuguese*. Berlin: Mouton de Gruyter.
- Souza, Christiane Maria Nunes de, and Raquel Gomes Chaves. 2015. A avaliação da concordância verbal com o pronome *tu* em Florianópolis. *UFSC Working Papers em Linguística* 16: 170–89. [CrossRef]
- Tagliamonte, Sali, and R. Harald Baayen. 2012. Models, forests, and trees of York English: *Was/were* variation as a case study of statistical practice. *Language Variation and Change* 24: 135–78. [CrossRef]
- Tarallo, Fernando. 1996. Turning different at the turn of the century: 19th century Brazilian Portuguese. In *Towards a Social Science of Language: Papers in Honor of William Labov*. Edited by Gregory R. Guy, Crawford Feagin, Deborah Schiffrin and John Baugh. Amsterdam: John Benjamins, vol. 1, pp. 199–220.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Regularization and Innovation: A Usage-Based Approach to Past Participle Variation in Brazilian Portuguese

Kendra V. Dickinson

Department of Spanish and Portuguese, Rutgers University, New Brunswick, NJ 08901, USA;
kendra.dickinson@rutgers.edu

Abstract: This project explores the synchronic variation of participle forms in Brazilian Portuguese (BP). Despite general systematicity, the language maintains many historically irregular participles, which often compete with regularized variants. The language has also developed innovative participles, which tend to exist in variation with regular forms. Adopting a usage-based framework, the study examines how analogical processes affect persistent irregular participles and short-form forms in BP, emphasizing the role of grammatical context and frequency. Data are drawn from the Portuguese Web 2011 corpus, including 12 verbs with long-form Latinate irregulars ($n = 4800$) and 8 verbs with short-form forms ($n = 3200$). The results show that long-form Latinate irregulars are more common as adjectives and with the verb *estar*, while regularized forms are prevalent with *ser* and in perfect constructions. Conversely, short-form participles occur least frequently in perfect constructions, showing a tendency towards the maintenance of regularity in this context. Additionally, verbs that occur more often in perfect constructions are most resistant to innovation. These findings indicate that perfect constructions play a dual role in *promoting* and *preserving* regularity in BP and shed light on how grammar–internal relationships and contexts of occurrence play a role in language variation and change.

Keywords: language variation and change; usage-based grammar; linguistic analogy

Citation: Dickinson, Kendra V. 2024. Regularization and Innovation: A Usage-Based Approach to Past Participle Variation in Brazilian Portuguese. *Languages* 9: 52. <https://doi.org/10.3390/languages9020052>

Academic Editor: Patrícia Amaral

Received: 14 October 2023

Revised: 9 January 2024

Accepted: 12 January 2024

Published: 30 January 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

This project investigates synchronic variation in Brazilian Portuguese (BP) past participle forms. Specifically, this analysis focuses on the variation of these forms as adjectives modifying nouns (1), with the verb *estar* ('to be') (2), with the verb *ser* ('to be') (often used in the passive voice)¹ (3), and in perfect constructions with *ter* ('to have') (4). Throughout, the term 'participle' will be used as an umbrella term describing both true past participles in perfect constructions and participle-derived adjectival and predicative forms. Despite the general regularity of the formation of BP past participles through the addition of *-ado/-ido* to the verb root according to inflection class, there are two sets of verbs that do not conform to this model, namely, verbs with irregular long-form Latinate² participles, and verbs that possess innovative, often termed 'short-form' participles. In both cases, divergent participle forms often exist in variation with forms that follow the regular template. This is exemplified in examples (1)–(4) for the verb *imprimir* ('to print'), which can be expressed both as long-form Latinate irregular *impresso* and regularized *imprimido*.

- (1) O livro impresso/imprimido terá características únicas.
the book printed.M.SG have.3SG.FUT characteristics unique.F.PL
'The printed book will have unique characteristics.'
- (2) O documento está impresso/imprimido em papel grosso.
The document be.3SG.PRES printed.M.SG on paper thick
'The document is printed on thick paper.'

- (3) O livro foi impresso/imprimido por uma só editorial.
 The book be.3SG.PRET printed.M.SG by one only publisher
 ‘The book was published by only one publisher.’
- (4) Ele já tinha impresso/imprimido o documento.
 He already have.3SG.IMP printed.M.SG the document
 ‘He had already printed the document.’

Similarly, innovative short-form participles in BP, which are often homophonous with other verb forms, most often first-person singular of the present indicative, exhibit alternation with regular forms. Example (5) illustrates the alternation between regular *ganhado* and short-form *ganho* for the verb *ganhar* (‘to earn/win’) in an adjectival context, though these forms can also be used in the same grammatical contexts as shown in (1)–(4). The innovative form *ganho* is identical to the first-person singular present indicative verb form meaning ‘I win’.

- (5) Eles recebem uma moeda para cada ponto ganhado/ganho.
 They receive.3PL.PRES a coin for every point won.M.SG
 ‘They receive a coin for every point won.’

Historically, participles in Latin and Old Portuguese showed greater irregularity than synchronic forms (Laurent 1999). Many previously irregular participles have been regularized in Modern BP, which has been described as a process of analogy and leveling towards increased systematicity in the participle paradigm (Chagas de Souza 2011; Laurent 1999). Significant research has been undertaken with regard to contemporary BP past participle variation (Chagas de Souza 2011; Huber 1933; Lobato 1999; Miara and Coelho 2015; Schwenter et al. 2019; Villalva and Jardim 2018; Queriquelli 2018; *inter alia*), though quite often with focus on innovative forms. Despite their mention in previous work, less is known about verbs whose long-form Latinate irregular participles persist synchronically, such as *morto* (‘dead’) for *morrer* (‘to die’), and their potential for regularization (cf. *morrido*, ‘dead’) (Laurent 1999, p. 74). In focusing on both regularization and innovation, the present analysis shows that these processes share some of the same underlying mechanisms. Specifically, this work applies usage-based frameworks of linguistic analysis to the investigation of synchronic participle variation in BP and demonstrates that this variation is at least in part governed by analogical processes that are conditioned by grammatical context, contextual probabilities, and frequency.

2. Background

2.1. Previous Work on Portuguese Past Participles

Historical variation in Portuguese past participles has continued from Old Portuguese to Modern Portuguese. Attestations of variation between standard and regularized participles in Portuguese can be found, going as far back as the 15th century. Examples of this are shown in (6) and (7), where the regularized *abrido* is used in place of irregular *aberto* from the verb *abrir* (‘to open’), and *morrido* is used instead of the irregular form *morto* for the verb *morrer* (‘to die’), respectively. Both examples were extracted from the *Genre/Historical Corpus do português* (Davies 2004).

- (6) E sera abrido a vos.
 and be.3SG.FUT open.PART to you
 ‘And it will be opened to you.’
Livro de vita Christi (1446)
- (7) Tinha morrido o Rey do Egypto.
 have.3SG.IMP die.PART the King of Egypt.
 ‘The king of Egypt had died.’
Promptuario historico II, Frei Manoel da Mealhada (1760)

Despite historical attestations of regularized *abrido* and *morrido*, neither underwent a full process of regularization. However, there are many verbs in Portuguese that have been fully regularized in the language, for which the regularized variants are the dominant variant and are, in many cases, used categorically. The changes that Portuguese past participles have experienced have been described in three phases (Chagas de Souza 2007). In the first phase, irregular forms from Latin existed in doublets, competing with their regularized counterparts. Examples of this are shown in Table 1.

Table 1. Sample of long-form Latinate double participles in Portuguese.

Verb	Meaning	Long-Form Latinate Irregular	Regularized Form
<i>aceitar</i>	‘to accept’	<i>aceito</i> ¹	<i>aceitado</i>
<i>expulsar</i>	‘to expel’	<i>expulso</i>	<i>expulsado</i>
<i>expresser</i>	‘to express’	<i>expresso</i>	<i>expressado</i>

¹ *Aceitar* has another participle variant, *aceite*. This is an innovative form used primarily in European Portuguese. It is not included in the table because the table is meant to show examples of participles with irregularity inherited from Latin rather than those forms that arose via analogy. *Aceito* is the form with inherited irregularity, while *aceite* is thought to have arisen via analogy to other adjectival forms ending in *-e* (Williams 2016).

In the second phase, Chagas de Souza (2007) suggests that homophony between persistent irregular forms and first-person singular present indicative verb conjugations allowed for the analogical extension of this pattern to other verbs in the first conjugation class (*-ar*) that did not originally exhibit this pattern. For example, the irregular participle for the verb *aceitar* (‘to accept’) is *aceito*, which is identical to the conjugated verb of the first-person singular of the present indicative. Table 2 shows some examples of verbs in the first conjugation to which this pattern was extended.

Table 2. Sample of double past participles for *-ar* verbs in Portuguese.

Verb	Meaning	Regular Participle	1SG	Innovative Participle
<i>pagar</i>	‘to pay’	<i>pagado</i>	<i>pago</i>	<i>pago</i>
<i>chegar</i>	‘to arrive’	<i>chegado</i>	<i>chego</i>	<i>chego</i>
<i>falar</i>	‘to talk’	<i>falado</i>	<i>falo</i>	<i>falo</i>

In a diachronic analysis of ‘double participles’ in Portuguese, Hricsina (2019) finds that these pairs began to be used around the 14th century and began to increase significantly during the 15th century. It is important to note that, in Modern Portuguese, some of the innovative participles in the first conjugation (*-ar*) are considered to be standard, such as *pago* (cf. *pagado*, ‘to pay’), while others (e.g., *chego* ‘to arrive’) are not (Chagas de Souza 2007).

Finally, in the third phase, this pattern expands by analogy to other conjugation classes (i.e., *-er*, *-ir*). Examples of this extension are shown in Table 3. Again, it is critical to note that most of the innovative participles in these conjugation classes are not considered standard in Modern Portuguese (Chagas de Souza 2007).

Table 3. Sample of double past participles for *-er/-ir* verbs in Portuguese.

Verb	Meaning	Regular Participle	1SG	Innovative Participle
<i>trazer</i>	‘to bring’	<i>trazido</i>	<i>trago</i>	<i>trago</i>
<i>pedir</i>	‘to arrive’	<i>pedido</i>	<i>peço</i>	<i>peço</i>
<i>vender</i>	‘to sell’	<i>vendido</i>	<i>vendo</i>	<i>vendo</i>

These last two stages have brought about an abundance of competing participles forms in Modern Portuguese, which are often referred to as *participios duplos* ('double participles') or 'short-form' participles (Lobato 1999). In these pairs, one participle generally follows the regular pattern of adding *-ado* or *-ido* to the infinitival stem. The other participle is identical to the first person singular present indicative verb conjugation (Perini 2002). Examples of these double participles are shown in Table 4.

Table 4. Examples of regular and innovative double participles in Portuguese.

Verb	Stem	Standard Participle	Innovative Participle
<i>pagar</i>	pag-	<i>pagado</i>	<i>pago</i>
<i>vender</i>	vend-	<i>vendido</i>	<i>vendo</i>
<i>falar</i>	fal-	<i>falado</i>	<i>falo</i>

Much of the synchronic variation exhibited in Portuguese participles can be attributed to various stages of the processes described by Chagas de Souza (2007). Traditional grammars have suggested that the distinction between regular and short-form participles in Portuguese is a question of syntactic structures, namely, that regular participles are used in compound tenses, and the innovative analogical participles are used as adjectival complements (Cunha and Cintra 2016). More specifically, it has been suggested that regular participles are used with the auxiliary verb *ter* ('to have') and that short-form participles are generally used with the copular verb *ser* ('to be') or without an auxiliary (Perini 2002, p. 153). While some traditional grammars have acknowledged that there are exceptions to this pattern (cf. Perini 2002, p. 154) and that there are some participles that are used across different types of grammatical constructions, more recent research has shown that the situation is far more complex.

Previous work has shown that the selection between the two forms is most strongly determined on a verb-by-verb basis (Schwenter et al. 2019), indicating the role of specific verb lexemes and their frequency and uses. Furthermore, authors such as Hricsina (2019, p. 89) suggest that this lexical determination of which participle types are used in different grammatical constructions is a feature of a change in progress. Similarly, Chagas de Souza (2011) proposes that the lexical determination of past participle selection is part of a directional syncretism toward the first-person singular of the present indicative.

In their corpus analysis of Portuguese participles, Schwenter et al. (2019) found that for participle pairs that include a regular form and an irregular form based on the first-person singular of the present indicative, the irregular form was used 69% of the time. Conversely, for participle pairs composed of a regular form and an irregular form of some other type (e.g., long-form Latinate irregulars), the irregular form was used just 32% of the time. These findings reveal that long-form Latinate irregular participle forms are more often dispreferred as compared to their regularized counterparts in Portuguese, and that for verb lexemes that have regular and short-form participles, the latter are preferred. The authors conclude, therefore, that not only do grammatical function and verb lexeme play a role in participle selection, but the morphological shape of the participle itself is an important predictor of regular vs. irregular usage rates.

2.2. Usage-Based Grammar and Linguistic Analogy

Critical to the present work are explanatory frameworks for the participle patterns observed in BP. One such framework is Usage-based Theory. This approach to understanding linguistic variation views grammar as the cognitive organization of individuals' experience with language (Bybee 2006), with some scholars going so far as to assert that grammar is an emergent phenomenon that is continually reshaped by its specific contexts of use (Hopper 1987). From the most general standpoint, our cognitive representations of language forms, their meanings, and the contexts they occur in are constructed out of our encounters with the forms in use. Contrary to generative theoretical approaches that

focus on abstractions of language to the exclusion of how language is used in practice, usage-based approaches focus on how individuals' cumulative experiences with language shape variation and change. In this way, language forms are conventionalized to varying degrees out of our experiences with them, meaning that form, frequency, and grammatical context play critical roles in cognitive representations of language.

In usage-based approaches, a critical mechanism of change is linguistic analogy. This refers to the process by which irregular forms, or forms that are perceived as irregular, are remade in the shape of other linguistic forms that are morphologically predictable. This process has been described as being related to child language acquisition, wherein children produce new forms by accessing similar forms in the lexicon and applying the same pattern to the new items (MacWhinney 1978) and has long been thought to be a critical process in language change (Anttila 1977; Blevins and Blevins 2009; Hock 1991). Importantly, historical changes in Portuguese towards greater participle systematicity have been attributed to the mechanism of linguistic analogy (Laurent 1999; Chagas de Souza 2007).

The present work adopts the notion that this type of linguistic analogy is part of a cognitive process in which speakers transfer knowledge from one set of linguistic forms to another based on perceived similarities between them. This type of leveling is more likely to occur among closely related forms (Bybee 2010; Blevins and Blevins 2009). In the case of Portuguese participles, analogy to regular participle formation within the verbs of the same conjugation class is far more likely to occur than analogy to verbs of a different conjugation class or to verb forms that are not past participles. For example, the historically irregular Latinate participle *nato* for the verb *nascer* ('to be born') is thought to have regularized to *nascido*, based on the application of the morphologically predictable rule applied to other verbs in the same conjugation class, which involves the addition of *-ido* to the verbal stem. Additionally, analogical processes tend to affect paradigms that are less frequent and permit more alternations in those that are more frequent (Bybee 2010), which can be attributed to the strength of the mental representations of more frequent paradigms for speakers of the language (Bybee 1985). This mechanism of language variation and change is a critical explanatory framework for understanding not only historical change of Portuguese participles but also synchronic variation. On the one hand, these processes of analogical change affected many historically irregular participles from Latin to Portuguese. Still, not all irregular past participles were regularized in this way, resulting in verbs that either fully maintain long-form Latinate irregular participles or verbs that exhibit synchronic variation between long-form Latinate irregulars and regularized variants. On the other hand, verbs that have developed innovative participle forms in Portuguese have been described as undergoing analogical processes to first-person singular verb forms, driven by the existence of this historical overlap for other verbs. Importantly, the present analysis addresses the extent to which these analogical processes are synchronically applied to both persistent long-form Latinate irregular participles and to the extent of use of short-form participle forms in BP.

2.3. Frequency and Probability

A central tenet of usage-based approaches to grammar is the important role of frequency in linguistic patterns. Linguistic frequency refers to the number of times that a particular language unit occurs in a given corpus or data set and has been used to approximate the number of times speakers are likely to experience specific linguistic forms. Following usage-based approaches, specific instances of grammatical elements are stored in memory, and their frequency of use has an impact on their mental representations (Bybee 2006, 2010). Linguistic elements that speakers more frequently encounter are said to have higher degrees of representation in the mental lexicon (Bybee 1995), making them more resistant to change and more likely to serve as the basis for the development of new forms (Bybee 2003; Bybee and Beckner 2010; Langacker 1987).

When considering frequency and analogy as mechanisms of language variation and change, the relevance of the linguistic contexts in which variable forms occur cannot be ignored.

In lexical diffusion models of sound change, it has long been considered that specific words undergo change, which can then gradually propagate to other words that share similar sounds or phonological patterns before potentially spreading throughout the lexicon. Importantly, these models have further developed to include a consideration of the influence of contextual factors on the rates of linguistic change. For example, the frequency with which a word occurs in a linguistic context that promotes a particular sound change can impact the rate at which the word undergoes said change. In a critical work on this topic, Bybee (2002) finds that patterns of /t,d/ deletion in American English occurred more readily in words that occurred more often before vowels than in words before consonants, demonstrating that the linguistics contexts of use of a given word can affect the rate of sound change.

While these sorts of models have been more frequently applied to sound-based language change, they can also be applied to morphosyntactic phenomena. In his analysis of lexical decision models, Baayen (2010) finds that overall word frequency predicts only a small portion of the variance observed in lexical decision tasks. Instead, he finds that the probabilities associated with local syntactic and morphological co-occurrences, referred to as *contextual probabilities*, are a far more powerful predictor of lexical decisions.

Relatedly, Schmid (2015) proposes that the syntagmatic associations between linguistic forms can be strengthened in a number of ways. First, the repetition of linguistic sequences strengthens the syntagmatic associations between their form and meaning. Second, the repetition of similar linguistic elements under similar contextual circumstances strengthens the relationships between them and facilitates their activation in comparable linguistic environments, termed *contextual entrenchment*. Most crucially to the present analysis, Schmid proposes that this process leads to increased routinization of linguistic forms through analogy as a cognitive process, which allows for the identification of the shared role that different elements play in a given linguistic context. In other words, the strength of association between co-occurring linguistic forms facilitates the development of schematic constructions in our mental grammar. These concepts of contextual probabilities and contextual entrenchment fit into previously established frameworks of linguistic analogy and regularization. In particular, linguistic analogy reforms verbal paradigms in such a way that functionally similar elements also become similar in form, referred to as paradigmatic iconicity (Croft 2003, 2012). These frameworks are critical to the analysis that follows, which demonstrates the important role of contextual relationships in processes of regularization and innovation in BP past participle forms.

3. Materials and Methods

3.1. Corpus Description

Data are drawn from Sketch Engine's 2011 Portuguese Web Corpus (ptTenTen11), which contains 4.6 billion words (Kilgarriff et al. 2014). The corpus creators used the software Heritrix (Mohr et al. 2004) to identify and download text from the Portuguese-speaking Internet and the software program FreeLing (Padró et al. 2010) to clean and tag the data. The corpus includes Internet texts from throughout the Portuguese-speaking world. For the purpose of this analysis, the selection was restricted to Brazil only.³

3.2. Data Collection Procedures

For the data in the present analysis, I collected two different data sets: one including verbs that included irregular long-form Latinate past participles and one that included verbs with short-form past participles, following similar procedures. For the former, I examined a total of 46 verbs in Portuguese that have irregular long-form Latinate participle variants that persist synchronically to varying degrees, based on previous descriptions of irregular past participles in Portuguese (Chagas de Souza 2011; Laurent 1999; Perini 2002; Schwenter et al. 2019). I used the Corpus Query Language (CQL) of the ptTenTen11 corpus to simultaneously search for all morphological variants of a past participle type for each verb. For example, for the verb *imprimir* ('to print'), I searched [word = 'imprimido | imprimidos | imprimida | imprimidas'] to collect regularized tokens,

and [word = 'impresso | impressos | impressa | impressas'] to collect long-form Latinate irregulars. Some verbs required additional query specification due to overlap with words from other lexical categories. For example, the masculine forms of the irregular participle of the verb *morrer* ('to die'), *morto(s)*, are homophonous with the noun meaning 'dead person/people', and the masculine singular form of the irregular participle for the verb *fixar* ('to attach') is also a noun meaning 'landline'. In cases like this, I included additional specifications in the CQL to exclude words tagged as nouns.

Regarding the verbs with short-form variants, I followed a similar procedure. However, as the short-form participle forms are homophonous with other verb forms, I added additional restrictions to the query, specifying exclusion of certain verb conjugations. For example, the short-form forms for the verb *pagar* ('to pay') overlap not only with the noun *pago(s)* ('payment(s)') but also with *pago* ('I pay'), *pagas* ('you paga'), *paga* ('you/he/she pays'). I searched for [word = 'pagado | pagado | pagada | pagadas' and tag! = 'N.* | V.IP123'] to collect regular tokens, and [word = 'pago | pagos | paga | pagas tag! = 'N.* | V.IP123'] to collect a sample of the short-form variant. Any nominal or conjugated verbal uses that were not eliminated following this procedure were excluded during the coding phase. For both data sets, I also recorded the total number of tokens of each participle type in the corpus and determined which verbs demonstrated sufficient variation for inclusion in inferential analysis as had been established previously via power analysis, which in the present case was 200 tokens per verb per participle type.

For each verb that had enough attestations in the corpus for each participle type, I collected a balanced random sample of tokens using Sketch Engine's built-in random sampling feature. This tool works by randomly selecting the number of lines indicated by the user from all parts of the corpus. According to Sketch Engine's documentation, the tool is designed to reduce the number of lines while at the same time maintaining the representativeness of the sample. For verbs with high token counts of both participle types, I used the built-in random sampling tool in the corpus to collect 400 tokens per type, for a total of 800 per verb. From these 400 tokens, I collected and coded the first 200 eligible tokens per verb for grammatical context (adjective modifying a noun, perfect construction with *ter*⁴, with *estar* or with *ser*). I collected a total of 200 tokens per type for each verb form, for a total of 400 tokens per verb. For the data set, which included long-form Latinate irregulars and regularizations, this procedure produced a total of 4800 tokens, or 400 tokens, for each of the 12 verbs that showed variation and had sufficient tokens for statistical inference. For the data set, including regular forms and short-form variants, this procedure resulted in a total of 3200 or 400 tokens for each of the 8 verbs analyzed.

3.3. Statistical Methods

For both of the data sets presented in this paper, I calculated the overall frequency of each verb lexeme using Sketch Engine's built-in tool, which provides frequency per million words. I then calculated the ratios between particular types in order to capture their relative frequencies. For the subset of verbs containing long-form Latinate irregular participles, the relative frequencies range from 1.60×10^{-5} to 35,920.1 (median = 1.23×10^{-2}), indicating a massive right skew for this measure, where values from zero to 1 indicate that the irregular value is more frequent, and values above 1 indicate greater relative frequency of the regularized variant. For the subset of verbs containing short-form participle forms, the relative frequencies ranged from 0.0016 to 125.45 (median = 0.154), also indicating right-skew, though not as extreme as in the first subset of the data. Therefore, for both subsets of the data, I log-transformed the relative frequency values to meet the necessary assumptions for statistical analysis. I then employed a two-tailed *t*-test for each data set to analyze the relationship between overall verb frequency and the relative frequencies of participle types.⁵

Additionally, in order to examine the potential role of lexeme frequency, I searched for each verb in both samples in the corpus, which returns a calculation of the number of occurrences of that word per million words in the corpus. For the sample containing long-form Latinate irregulars, verb lexeme frequencies ranged from 0.7 to 261.9 (median = 16.7).

For the short-form data set, lexeme frequencies ranged from 20.43 to 481.75 (median = 93.5). I, therefore, log-transformed these measures for both data sets for similar reasons to those described above. Finally, in order to investigate the potential role of contextual frequencies in patterns of regularization and innovation, I also followed similar procedures to calculate the frequencies of each of the verbs in each of the grammatical contexts under consideration in these analyses.

Statistical analysis was performed using R (R Core Team 2022). For each separate data set, I generated random forests (Liaw and Wiener 2002), which uses series of decision trees to classify the relative strength of each predictor value with regard to the dependent variable. Based on this output, together with theoretically informed predictions, I created a series of nested models using the lme4 package (Bates et al. 2015), considering grammatical context, verb lexeme frequency, and verb lexeme frequency in each of the grammatical contexts⁶ under consideration as potential predictors. For the sets of models for each grouping of the data, I included a random intercept for the verb. Though further random effects are desirable to create the optimal model, additional specification was not possible due to convergence issues despite optimization. Future analyses with additional data will attempt to resolve this issue. Finally, I tested the goodness of fit for each of the models using the anova() function in R, a likelihood ratio test (Fox and Weisberg 2018) to select the optimal model for each data.

4. Results and Discussion

4.1. Long-Form Latinate Irregulars

4.1.1. Description of the Data

Of the 38 verbs examined, 2 showed no use of irregular forms (*devolver*, ‘to return’; *dissolver*, ‘to dissolve’), 1 showed no evidence of regularized forms (*predizer*, ‘to predict’), and 2 showed categorical usage patterns across grammatical contexts. Specifically, there was a clean split for the verbs *fixar* (‘to possess’) and *juntar* (‘to join’), for which there were no instantiations of long-form Latinate irregulars *fixo* or *junto* in perfect constructions. These verbs were excluded from further analysis, leaving a total of 35 verbs that had attested use of both irregular and regularized participles that were used in overlapping grammatical contexts. Of these verbs, *benzer* (‘to bless’), *corromper* (‘to corrupt’), *desenvolver* (‘to develop’), *despertar* (‘to wake up’), *eleger* (‘to choose’), *envolver* (‘to include, to contain’), *fritar* (‘to fry’), *imprimir* (‘to print’), *morrer* (‘to die’), *revolver* (‘to revolve’), *romper* (‘to break’), *soltar* (‘to release, to detach’) showed sufficient variation for statistical inference, as established by power analysis. See Appendix A for the full list of verbs, token counts, and usage distributions.

For these 12 verbs, I collected a total of 4800 past participle tokens, which included a total of 2671 tokens as adjectives, 280 tokens with *estar*, 1198 tokens with *ser*, and 652 tokens in perfect constructions. The overall results show that the irregular participle variants occurred more frequently as adjectives and with the verb *estar*, while the regularized forms occurred more frequently with *ser* and in perfect constructions. Specifically, 66.8% ($n = 1783$) of adjectival tokens and 66.1% ($n = 185$) of all tokens with *estar* in the sample occurred with the irregular participle variant. Conversely, 64% ($n = 767$) of tokens with *ser* occurred with the regularized variant, and perfect constructions nearly categorically (99.7%, $n = 649$) use the regularized form, as shown in Figure 1.

However, there was still variation by verb lexeme. Figure 2a,b shows the distributions of irregular and regularized participle types across grammatical functions for each verb. These figures show a clear trend of near-categorical use of the regularized variant in perfect constructions across all verbs. Additionally, for the majority of verbs, adjectival tokens and tokens with *estar* showed the lowest rate of regularized forms across all grammatical contexts. However, there are a few exceptions to general trends in these data. For example, while the majority of the verbs in the sample show higher rates of regularized forms with the verb *ser*, the verbs *envolver*, *morrer*, and *soltar* all show preference for the irregular participle in this grammatical context. Despite some idiosyncrasies by verb, there appears to be general consistency in preference for the regularized variant in perfect constructions and a preference for irregular variants as adjectives and with *estar*.

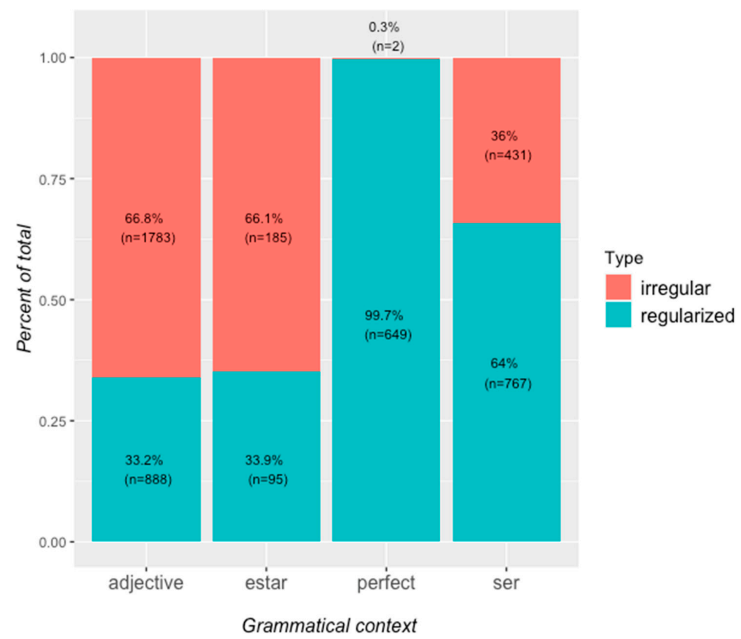


Figure 1. Participle types by grammatical context for long-form Latinate irregulars.

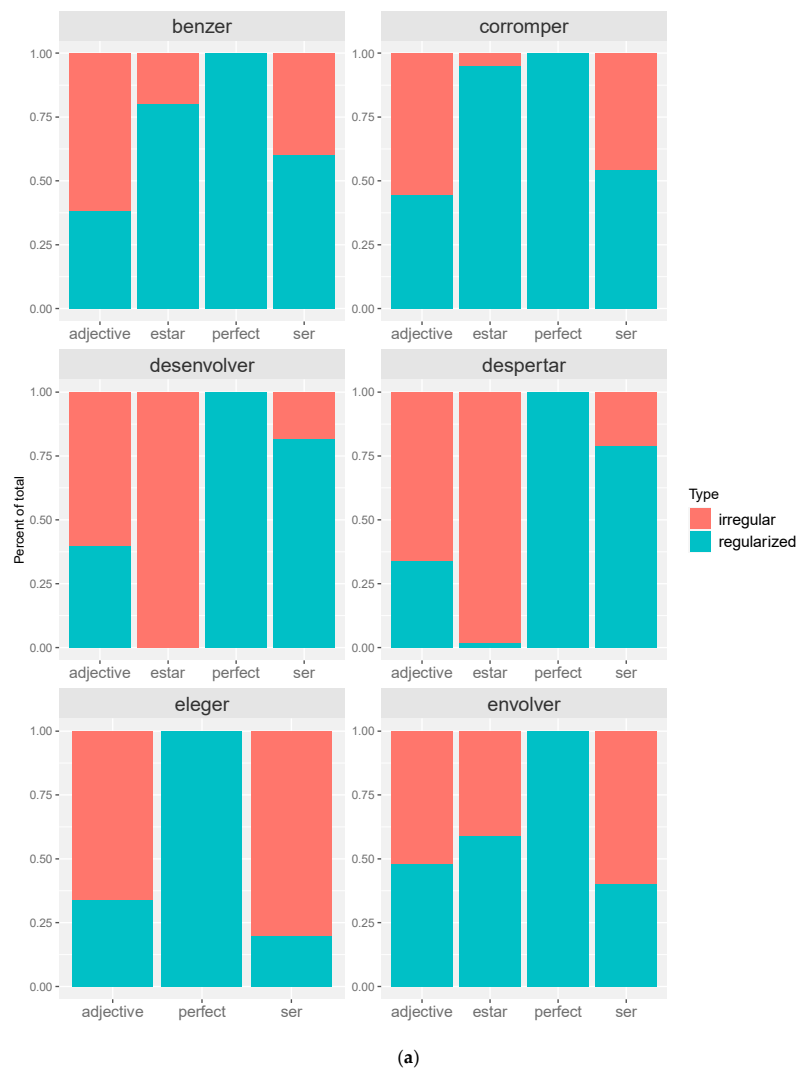


Figure 2. Cont.

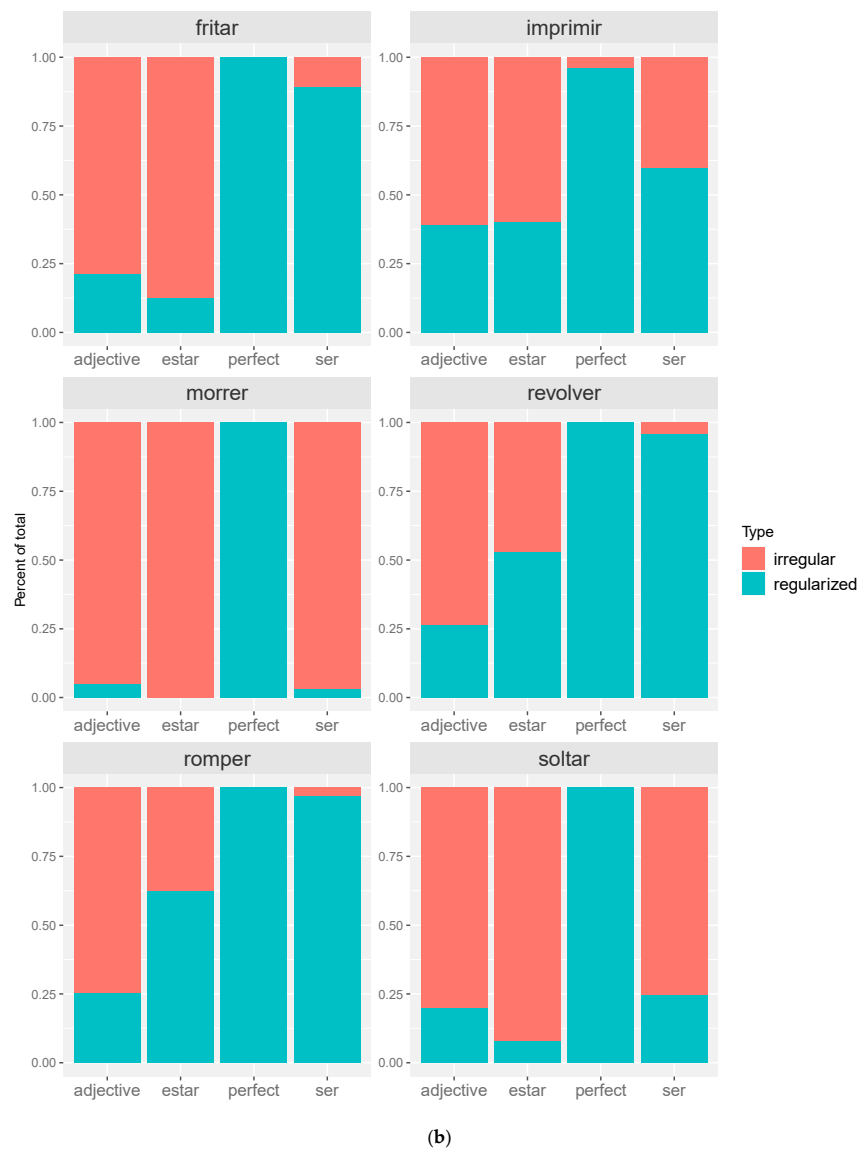


Figure 2. (a) Distribution of participle types by context for verbs with long-form Latinate irregulars. (b) Distribution of participle types by context for verbs with long-form Latinate irregulars.

4.1.2. Inferential Analysis of Long-Form Latinate Irregulars

Following the described statistical procedure, the best-fit model for the data identified grammatical context as the sole main effect. The output of this regression model is shown in Table 5.

Table 5. Best fit logistic regression model output for BP long-form Latinate irregulars.

	Factor Value	Estimate	Std. Error	z-Value	p-Value
Grammatical Context Reference level = adjective	Intercept	−0.96060	0.25364	−3.787	0.000152
	estar	0.08089	0.13890	0.582	0.5603
	perfect	8.37183	0.75619	11.071	<2 × 10 ^{−16}
	ser	1.59433	0.07885	20.220	<2 × 10 ^{−16}

The output of this best-fit model reveals that in these data, regularized forms are statistically significantly more likely to occur with *ser* and in perfect constructions than as adjectives and with *estar*. Conversely, the irregular variants are more likely to occur as

adjectives and with *estar* than are regularized variants. Overall, these results indicate the critical role of grammatical context in the variation between long-form Latinate irregular participle forms and their regularized counterparts.

4.2. Short-Form Participles

4.2.1. Description of the Data

As previously noted, Brazilian Portuguese also possesses short-form participle forms that are often homophonous with other verb forms and, in particular, with the first-person singular of the present indicative. In this section, I describe and analyze a subset of these verbs with short-form past participles in order to compare and contrast their patterning with the participle types described in the previous section. Since the development of innovative participle forms in Brazilian Portuguese is a change in progress (Chagas de Souza 2007; Hricsina 2019), the group of verbs to which this change applies is ever-changing. Therefore, due to the limitations of the present analysis, 10 verbs that have been described as having short-form participle forms (Chagas de Souza 2007; Hricsina 2019; Perini 2002; Schwenter et al. 2019) were selected for analysis.⁷

The verbs examined in the present analysis are *chegar* ('to arrive'), *empregar* ('to employ'), *entregar* ('to turn in'), *ganhar* ('to win, to earn'), *gastar* ('to spend'), *limpar* ('to clean'), *pagar* ('to pay'), *pedir* ('to ask for'), *pegar* ('to catch, to take'), and *trazer* ('to bring'). Though all of the verbs examined showed variable use across grammatical contexts, two verbs did not have sufficient short-form tokens in the corpus for inferential analysis (*pedir*, 'to ask for'; *trazer*, 'to bring'). These verbs, reported in Table 6, were excluded from further analysis, leaving a total of eight verbs that showed sufficient variation for statistical analysis. While only the masculine singular participle form is listed in the table for the ease of the reader, the totals include all gender and number forms.

Table 6. List of BP verbs with short-form participles in the present analysis.

Verb	Lexeme Freq./Million	Regular Form	Total Regular	% Regular	Short-Form	Total Short-Form	% Short-Form
limpar	20.43	limpado	1099	0.80%	limpo	137,864	99.20%
empregar	36.25	empregado	102,684	92.80%	emprego	7989	7.20%
gastar	51.33	gastado	1292	2.50%	gasto	51,352	97.50%
entregar	89.95	entregado	197,788	98.90%	entregue	2150	1.10%
pegar	97.04	pegado	21,587	70.70%	pego	8943	29.30%
pagar	140.17	pagado	1336	1.20%	pago	107,008	98.80%
pedir	207	pedido	25,102	99.80%	peço	57	0.20%
trazer	231.4	trazido	66,442	99.80%	trago	104	0.10%
ganhar	280.65	ganhado	10,346	81.30%	ganho	2373	18.70%
chegar	481.75	chegado	55,500	97.70%	chego	1323	2.30%

I also calculated the relative frequencies of short-form to regular participles for each verb in this data set. The results of a two-tailed *t*-test show a significant relationship between verb frequency and the relative frequency of regularized participles for these 10 verbs included in the present analysis ($t = 3.8899$, $df = 9$, p -value = 0.003675). Additionally, Pearson's correlation coefficient for these measures shows a moderate correlation between verb frequency and relative participle frequency ($r = -0.6$). This relationship indicates that, in general, more frequent verb lexemes show fewer instantiations of short-form participles than less frequent verbs in this data set. Nevertheless, this tendency is not absolute. While there is a general tendency for lower frequency verbs such as *limpar* and *gastar* to have high rates of innovation, *empregar* is a notable exception. Additionally, while higher frequency verbs *chegar* and *trazer* show lower rates of innovation, *ganhar* shows 18.70% regularization despite a higher frequency. This indicates that lexeme frequency is not the sole predictor of participle innovation. Other factors, such as the form of the short-form participle, the

relationship between specific participles and other grammatical categories (e.g., nominal *gasto(s)* and *emprego(s)* from *gastar* and *empregar*, respectively), and more may play a role in the overall rates of innovation and will be explored in future analyses.

However, a stronger trend is shown regarding the overall rate of occurrence in perfect constructions and the degree of regularization for these verbs. Specifically, verbs in this sample that tend to occur more frequently in perfect constructions in the corpus overall also tended to show lower rates of innovation. While this analysis could be made more robust by the addition of more verbs, the results of a two-tailed *t*-test show a significant relationship between the number of tokens in perfect constructions in the corpus and the relative rate of innovation for these 10 verbs ($t = 2.5076$, $df = 9$, $p\text{-value} = 0.03344$). Additionally, Pearson's correlation coefficient for these measures shows a small negative correlation ($r = -0.32$).

For the eight verbs that had sufficient tokens for analysis, I collected a total of 3200 past participle tokens. The overall results show that both short-form and regular participle types are used variably across all grammatical contexts. However, these data show the highest preference for the regular form in perfect constructions, specifically, 75.8% ($n = 1151$), whereas the frequency of regular forms in the other grammatical contexts under consideration ranges from 26.2 to 27.2%, as shown in Figure 3. Recall that because this is a balanced random sample, the overall prevalence of regularized forms may appear inflated. Nevertheless, Figure 3 illustrates the distribution of regular versus short-form forms across the grammatical contexts in the present data set, most notably highlighting the relative frequency of occurrence of each participle type across grammatical contexts.

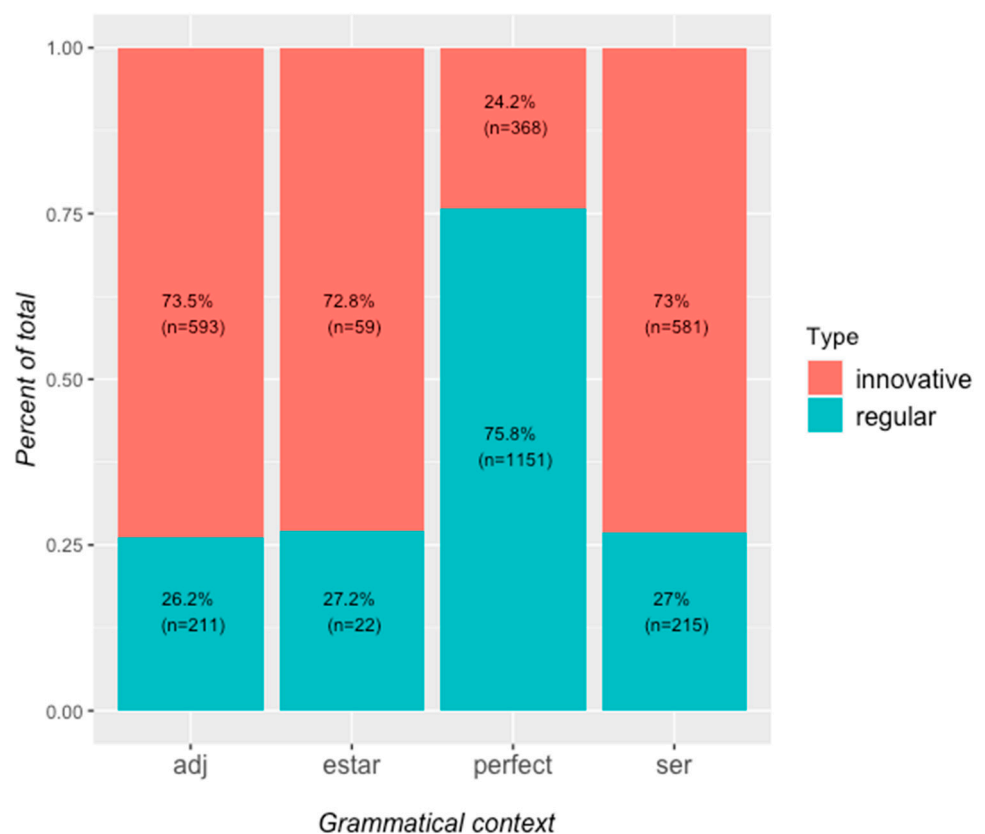


Figure 3. Participle types by grammatical context for verbs with short-form participles.

However, as in the previous data set, there was still variation between verbs. Figure 4 shows the distributions of regular and short-form participle types across grammatical functions for each verb. Interestingly, the most notable pattern in these data is the preference for the regular participle variant in perfect constructions, which is fairly consistent across verbs. Only 2/8 verbs, namely *chegar* ('to arrive') and *empregar* ('to employ'), do not show a preference for the regular participle form in perfect constructions. Regarding adjectival

contexts, all verbs except for *chegar* show a preference for the short-form variant. In constructions with *estar*, all verbs show greater rates of the short-form variant, except for *pagar* ('to pay') and *pegar* ('to catch/get'). With *ser*, there is a great deal of variation between verbs, with overall rates of short-form vs. regular participles in this context showing the greatest variation between verbs of any of the grammatical contexts. Generally, most verbs show a preference for the short-form variant, but *chegar* ('to arrive') and *limpar* ('to clean') show more instantiations of the regular form. These differences between verbs suggest that while grammatical context plays an important role in this variation, there are undoubtedly other factors that influence these patterns. Future analysis will endeavor to examine more verbs in order to uncover additional motivating reasons for this variation.

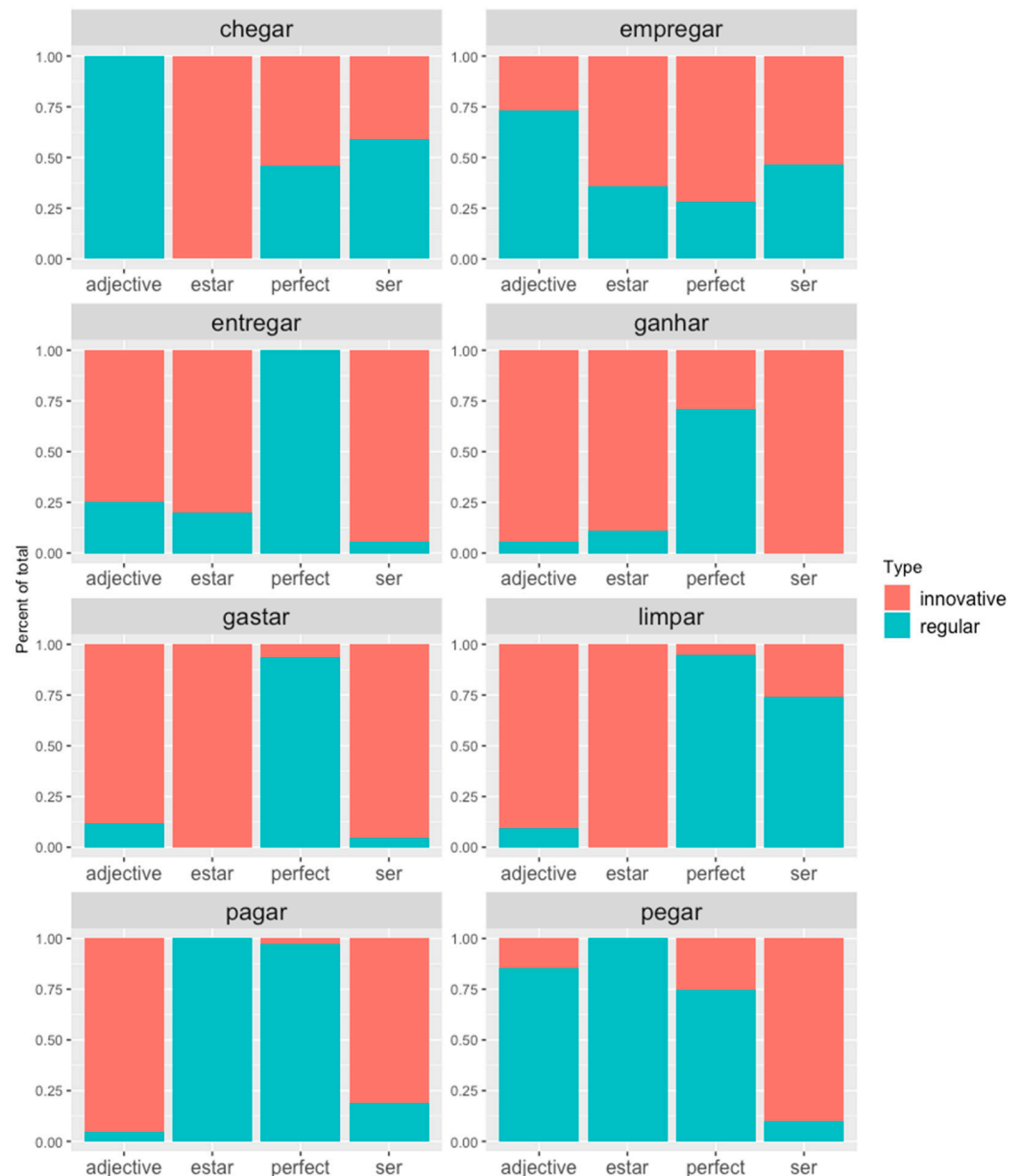


Figure 4. Distribution of participle types by context for verbs with short form.

4.2.2. Inferential Analysis of Short-Form Participles

The results from long-form Latinate participles, as well as the descriptive statistics provided in Section 4.2.1, highlight the potential for perfect constructions to be considered conditioning contexts for regular participles in BP. Therefore, the construction of models for the present data considered overall lexeme frequency as well as the contextual frequency of

verb lexeme in perfect constructions as potential predictors of the use of regular participle forms. Following the described statistical procedure, the best-fit model for the data included grammatical context as a significant main effect and a significant interaction between grammatical context and verb frequency in perfect constructions⁸. The output of this model is shown in Table 7.

Table 7. Best fit logistic regression model output for short-form forms.

	Factor Value	Estimate	Std. Error	z	p-Value
	<i>Intercept</i>	0.3451	0.0845	4.084	<0.001
Grammatical Context Reference level = adjective	estar	0.43139	0.26649	1.619	0.105
	perfect	−0.71328	0.09933	−7.181	<0.001
	ser	−0.08327	0.12968	−0.642	0.521
Grammatical Context * Freq. in Perfect Constructions	estar * freq.	−0.77875	0.55394	−1.0406	0.16
	perfect * freq.	−1.35474	0.17782	−7.619	<0.001
	ser * freq.	−2.70295	0.25577	−10.568	<0.001

The output of this best-fit model highlights some compelling patterns related to short-form past participles by grammatical context. First, the output of this model shows that grammatical context is a statistically significant predictor of participle type. Specifically, short-form past participles are significantly *less* likely to occur in perfect constructions than they are to occur in all other grammatical contexts. Where the findings described in Section 4.1 showed that perfect constructions promote the regularization of long-form Latinate irregular past participles, these results show that perfect constructions also serve as conserving environments for regular participles that compete with short-form forms. This finding is yet another reinforcement of the strength of perfect constructions in BP as conditioning environments for regularity in the form of *-ado/-ido* participles.

Second, the significant interaction between grammatical context and verb frequency in perfect constructions further emphasizes this conditioning. The output of this model shows that verbs that occur in perfect constructions, most frequently in the corpus as a whole, are more likely to show participle innovation as adjectives than in other grammatical contexts. This relationship is visualized in Figure 5. The *x*-axis shows the rate at which the verb lexeme associated with a given token occurs in perfect constructions in the corpus overall. In other words, this axis indicates the rate of co-occurrence of a participle with auxiliary *ter* or *haver*, with increasing co-occurrence from left to right. The *y*-axis shows the predicted probability of a short-form participle based on the best-fit model for these data.

Examining this figure, we can see the effects of the significant interaction in the logistic regression model. First, this shows that the perfect constructions and contexts with *estar* are generally consistent in terms of the rates of short-form participles, independent of how often a particular verb lexeme occurs in perfect constructions in the corpus as a whole. Second, this figure shows the verbs that tend to occur *more* frequently in perfect constructions in the corpus were more likely to show short-form participles used as adjectives in the present data set. Third, *ser* shows an inverse relationship with adjectival contents, whereby verbs that occur *more* frequently in perfect constructions in the corpus as a whole are *less* likely to be realized as innovative variants with *ser*. These results show that although lexemes that are most strongly associated with perfect constructions in the corpus as a whole exhibit participle innovation, these short-form participles tend to be restricted to adjectival contexts.

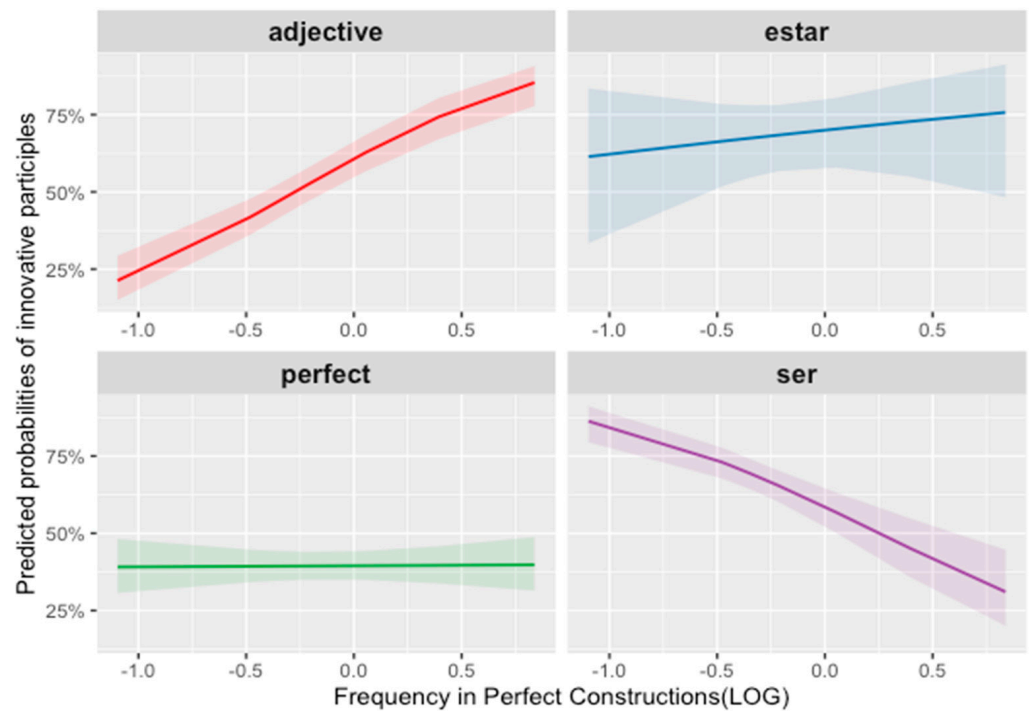


Figure 5. Predicted probabilities of short-form participles.

4.3. Discussion

Two distinct trends emerge from these data for BP past participles. First, the participles for verbs with one long-form Latinate irregulars are conditioned by grammatical context, with the strongest conditioning environment for regularization being perfect constructions. Second, the degree of participle innovation is a product of both grammatical context and contextual frequency in perfect constructions. The common thread in both of these findings is that independent of the data set; there is a strong tendency towards the regular *-ado/-ido* participle forms in perfect constructions. In the case of long-form Latinate irregulars, this means that instantiations of regularized variants are more likely to appear in perfect constructions. In the case of short-form participles, this means that innovations are more likely to appear as adjectives and with *ser* or *estar*, while participles in perfect constructions are more likely to remain regular.

These trends can at least partially be explained by the paradigmatic regularity. Specifically, in the BP subset of the ptTenTen11 corpus, there are a total of 4.1 million instantiations of *ter* + past participle. Of these, 3.6 million (87.9%) end in *-ado/-ido*, providing evidence for a strong lexical association between perfect constructions and regularly formed past participles. Furthermore, of the 230.4 million adjectival tokens of Brazilian Portuguese in the PtTenTen11 corpus, only 1.1 million (0.5%) end in *-ado/-ido*. This comparison is illustrated in Figure 6. These findings also indicate that the contextual probability of a regular past participle after an instance of an auxiliary verb⁹ is also nearly 9 to 1 in BP.

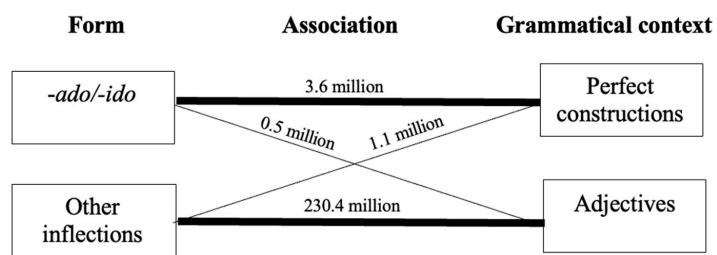


Figure 6. Form–function relationships of BP participle forms.

These findings align with previous work indicating that contextual probabilities are important predictors of lexical decisions (Baayen 2010) and provide additional evidence for the role of the strength of association between form and context in the development of grammatical constructions (Bybee 2013; Schmid 2015). The overall frequency with which auxiliary verbs and regular past participles are accessed together in BP, as illustrated by their high rate of co-occurrence in the corpus, heightens the possibility of analogical regularization due to the pressures of paradigmatic iconicity (Croft 2003, 2012) and contextual entrenchment (Schmid 2015).

These findings build on those laid out in previous work. First, these results align well with previous findings that regular past participles in Brazilian Portuguese are used in compound tenses, where innovative forms are used as adjectival complements (Cunha and Cintra 2016; Lobato 1999; Perini 2002; Villalva and Jardim 2018). Furthermore, these findings build on those of Schwenter et al. (2019), which showed that irregularity in Brazilian Portuguese past participles is highly lexically dependent. While a general correlation can be observed between the number of tokens in perfect constructions and the rate of innovation, there is still variation that cannot be explained via this mechanism. Observing Figure 1 above, the verb *chegar* shows the third-lowest rate of innovation in the data set but shows the overall highest number of tokens in the sample. While this could be a result of the overall frequency of *chegar*, it still indicates that there is some degree to which this variation is lexically constrained.

5. Conclusions

An important contribution of this analysis is the comparative examination of long-form Latinate irregulars and innovations rather than their treatment as a monolithic set of ‘irregulars’. This more fine-grained analysis shows that these two types of participles have overlapping yet distinct constraints on their use, which is critical for understanding how grammar–internal relationships between forms can influence use. Specifically, these findings highlight two distinct yet related effects of perfect constructions in BP. Namely, perfect constructions serve as a conditioning context *promoting* regularization in the case of long-form Latinate irregulars and serve as a *conserving mechanism* for regularity in the case of innovation, which can be attributed to the strength of association and frequency of occurrence between regular participle forms and perfect constructions. This analysis shows that these two types of participles have overlapping yet distinct constraints on their use, which is critical for understanding how grammar–internal relationships between forms can influence their usage patterns and highlights the role of the relationship between forms and contexts in language change. The present analyses not only provide insight into the synchronic variation of past participles in BP but also contribute to our understanding of the usage-based mechanisms of language variation and change.

Funding: This research received no external funding.

Institutional Review Board Statement: Ethical review and approval were waived for this study because the data collection methodology was based on the observation of public behavior. The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects (Federal Code of Regulations Title 45, Subchapter A, Part 26, section (d)(2)(i)).

Informed Consent Statement: Informed consent was waived for this study because the data collection methodology was based on the observation of public behavior. The information obtained is recorded by the investigator in such a manner that the identity of the human subjects cannot readily be ascertained, directly or through identifiers linked to the subjects (Federal Code of Regulations Title 45, Subchapter A, Part 26, section (d)(2)(i)).

Data Availability Statement: Data can be made available upon request.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A

Table A1. Full list of verbs considered for analysis, including token counts and usage distributions.

Verb	Irregular Form	Total Irregular	Regularized Form	Total Regularized	Use across Contexts
abrir	<i>aberto</i>	447,921	<i>abrido</i>	111	variable
absolver	<i>absolto</i>	31	<i>absolvido</i>	10,014	variable
bendizer	<i>bendito</i>	10,081	<i>bendizido</i>	1	variable
benzer	<i>bento</i>	2237	<i>benzido</i>	442	variable
cobrir	<i>coberto</i>	81,254	<i>cobrido</i>	41	variable
contradizer	<i>contradito</i>	553	<i>contradizido</i>	3	variable
corromper	<i>corrupto</i>	25,754	<i>corrompido</i>	8798	variable
descobrer	<i>descoberto</i>	61,744	<i>descobrido</i>	86	variable
descrever	<i>descrito</i>	111,788	<i>descrevido</i>	35	variable
desenvolver	<i>desenvolto</i>	685	<i>desenvolvido</i>	517,521	variable
desfazer	<i>desfeito</i>	8896	<i>desfazido</i>	1	variable
despertar	<i>desperto</i>	12,746	<i>despertado</i>	10,699	variable
devolver	<i>devolto</i>	0	<i>devolvido</i>	30,740	<i>no irreg.</i>
dissolver	<i>dissolto</i>	0	<i>dissolvido</i>	12,128	<i>no irreg.</i>
dizer	<i>dito</i>	217,670	<i>dizido</i>	33	variable
eleger	<i>eleito</i>	200,553	<i>elegido</i>	1072	variable
envolver	<i>envolto</i>	9514	<i>envolvido</i>	324,280	variable
escrever	<i>escrito</i>	425,412	<i>escrevido</i>	105	variable
extinguir	<i>extinto</i>	54,329	<i>extinguido</i>	667	variable
fazer	<i>feito</i>	1,997,297	<i>fazido</i>	32	variable
fixar	<i>fixo</i>	125,943	<i>fixado</i>	77,117	<i>categorical</i>
fritar	<i>frito</i>	17,060	<i>fritado</i>	544	variable
imprimir	<i>impresso</i>	96,505	<i>imprimido</i>	1654	variable
inscrever	<i>inscrito</i>	139,499	<i>inscrevido</i>	6	variable
juntar	<i>junto</i>	63,831	<i>juntado</i>	7834	<i>categorical</i>
maldizer	<i>maldito</i>	17,585	<i>maldizido</i>	3	variable
morrer	<i>morto</i>	183,516	<i>morrido</i>	15,348	variable
possuir	<i>posseso</i>	11	<i>possuído</i>	6481	variable
predizer	<i>predito</i>	1640	<i>predizido</i>	0	<i>no reg.</i>
prover	<i>provisto</i>	24	<i>provido</i>	16,180	variable
recobrir	<i>recoberto</i>	5247	<i>recobrido</i>	2	variable
resolver	<i>resolto</i>	6	<i>resolvido</i>	91,165	variable
reverter	<i>revolto</i>	6431	<i>revolvido</i>	595	variable
romper	<i>roto</i>	3749	<i>rompido</i>	10,053	variable
satisfazer	<i>satisfeito</i>	103,785	<i>satisfazido</i>	2	variable
soltar	<i>solto</i>	59,925	<i>soltado</i>	1145	variable
transcrever	<i>transcrito</i>	12,898	<i>transcrito</i>	2	variable
ver	<i>veido</i>	28	<i>visto</i>	445,431	variable
voltar	<i>volto</i>	9	<i>voltado</i>	323,281	variable

Notes

- ¹ Importantly, uses of participle forms with *ser* may correspond to different structures. Specifically, they can be used adjectivally (e.g., *As comidas são feitas de milho* ‘The foods are made with corn’), as well as in passive constructions (e.g., *A comida foi feita pela cozinheira* ‘The food was prepared by the cook’). Though there are potentially important semantic differences in these uses, the present analysis groups them together as uses with *ser*, in order to provide initial analysis of the role of grammatical context. Nevertheless, future analysis will endeavor to analyze potential distinctions between the two usage types.
- ² In the present work, ‘Latinated’ is used as an umbrella term to refer to participles whose irregularity can be traced to influence from forms in Classical Latin. These are to be distinguished from other irregular participle forms which arose as innovations via other historical processes.
- ³ Interestingly, Schwenter et al. (2019) found greater use of short-form forms in European Portuguese as compared to Brazilian Portuguese. Though of great interest, this is beyond the scope of the present analysis.
- ⁴ There are different grammatical constructions with *ter* and participle forms. For example, the participle form can be part of a perfect construction like the *Preterito Perfeito Composto* (*tenho preparado a comida* ‘I have prepared the food’), or function as an adjective as in *tenho a comida preparada* (‘I have a prepared food’). In the present data, instances of the first case were coded as perfect constructions with *ter*, while instances of the second were coded as adjectives modifying nouns.
- ⁵ The corpus data for long-form Latinated irregulars does not indicate a linear relationship between verb frequency and relative frequency of regularized past participles. Unexpectedly, results of a two-tailed *t*-test show a significant direct positive relationship between verb frequency and the relative frequency of regularized participles for these 35 verbs that show variation ($t = 6.151$, $df = 34$, $p\text{-value} = 5.487 \times 10^{-7}$). However, the Pearson’s correlation coefficient for these measures, which indicates the strength of association between the two continuous variables, shows only a very slight, if not negligible positive correlation ($r = 0.01$).
- ⁶ Though verb lexeme frequency in all grammatical contexts under analysis was calculated and taken into consideration in all statistical models, a significant contextual effect was seen only with regard to frequency in perfect constructions for short-form participles.
- ⁷ It is important to acknowledge that the data that comprise this corpus are from 2011, and it is likely that there have been changes since that time, in particular, on social media and other informal written genres. After the analysis of these data, Sketch Engine published a newer Portuguese copora (ptTenTen18 and ptTenTen20). This work represents only a preliminary view of this type of variation in Brazilian Portuguese past participles in the ptTenTen11 corpus and will continue to be developed.
- ⁸ Overall lexeme frequency was also considered as a potential predictor of degree of participle innovation, but was not found to be a statistically significant independent variable for these data and was not included in the best-fit model.
- ⁹ Due to the limitations on the current data set, teasing apart potential differences in participle selection between *ter* and *haver* was not possible, and most of the tokens occur with the former. Future research will endeavor to explore the degree to which *haver* participates in regular perfect constructions as compared to *ter*.

References

- Anttila, Raimo. 1977. *Analogy*. New York: Mouton Publishers.
- Baayen, R. Harold. 2010. Demythologizing the word frequency effect: A discriminative learning perspective. *The Mental Lexicon* 5: 436–61. [CrossRef]
- Bates, Douglas, Martin Maechler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software* 67: 1–48. [CrossRef]
- Blevins, James P., and Juliette Blevins. 2009. Introduction: Analogy in grammar. In *Analogy in Grammar: Form and Acquisition*. Edited by James P. Blevins and Juliette Blevins. Oxford: Oxford University Press, pp. 1–12.
- Bybee, Joan. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam: John Benjamins.
- Bybee, Joan. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes* 10: 425–55. [CrossRef]
- Bybee, Joan. 2002. Word frequency and context of use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change* 14: 261–90. [CrossRef]
- Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In *The Handbook of Historical Linguistics*. Edited by Brian D. Joseph and Richard D. Janda. Oxford: Blackwell, pp. 602–23.
- Bybee, Joan. 2006. From usage to grammar: The mind’s response to repetition. *Language* 82: 711–33. [CrossRef]
- Bybee, Joan. 2010. *Language, Usage and Cognition*. Cambridge: Cambridge University Press.
- Bybee, Joan. 2013. Usage-based theory and exemplar representations of constructions. In *The Oxford Handbook of Construction Grammar*. Edited by Thomas Hoffman and Graeme Trousdale. Oxford: Oxford University Press, pp. 49–69.
- Bybee, Joan, and Clay Beckner. 2010. Usage-based theory. In *The Oxford Handbook of Linguistic Analysis*. Edited by Bernd Heine and Heiko Narrog. Oxford: Oxford University Press, pp. 953–80.
- Chagas de Souza, Paulo. 2007. Athematic participles in Brazilian Portuguese: A syncretism in the making. *Acta Linguistica Hungarica* 54: 119–28. [CrossRef]
- Chagas de Souza, Paulo. 2011. Participios atemáticos no PB: Um processo paradigmático. *ReVEL* 5: 176–85.
- Croft, William. 2003. *Typology and Universals*, 2nd ed. Cambridge: Cambridge University Press.
- Croft, William. 2012. *Verbs: Aspect and Causal Structure*. Oxford: Oxford University Press.

- Cunha, Celso, and Lindley Cintra. 2016. *Nova Gramática do Português Contemporâneo*. Rio de Janeiro: Lexicon Editora Digital.
- Davies, Mark. 2004. Corpus do Português: Genre/Historical (Corpus). Available online: <https://www.corpusdoportugues.org/hist-gen/> (accessed on 1 January 2022).
- Fox, John, and Sanford Weisberg. 2018. *An R Companion to Applied Regression*. California: Sage Publications.
- Hock, Hans Henrich. 1991. *Principles of Historical Linguistics*. Berlin: Mouton de Gruyter.
- Hopper, Paul J. 1987. Emergent grammar. *Annual Meeting of the Berkeley Linguistics Society* 13: 139–57. [CrossRef]
- Hricsina, Jan. 2019. The Periphrasis Formed by the Verb *ir* + Infinitive in the History of the Portuguese language. *Etudes Romanes de Brno* 40: 81. [CrossRef]
- Huber, Joseph. 1933. *Altportugiesisches Elementarbuch*. Heidelberg: Carl Winters Universitätsbuchhandlung, (Portuguese translation Gramática do Português Antigo. Translated by Maria Manuela Delille. Lisboa: Fundação Calouste Gulbenkian, 1986).
- Kilgarriff, Adam, Miloš Jakubíček, Jan Pomikalek, Tony Berber Sardinha, and Pete Whitelock. 2014. PtTenTen: A corpus for Portuguese lexicography. In *Working with Portuguese Corpora*. Bloomsbury: Bloomsbury Academic, pp. 111–30.
- Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar: Theoretical Prerequisites*. Stanford: Stanford University Press.
- Laurent, Richard. 1999. *Past Participles from Latin to Romance*. Berkeley: University of California Press.
- Liaw, Andy, and Matthew Wiener. 2002. Classification and Regression by randomForest. *R News* 2: 18–22. Available online: <https://CRAN.R-project.org/doc/Rnews/> (accessed on 1 January 2022).
- Lobato, Lucia. 1999. Sobre a forma do participio do português e o estatuto dos traços formais. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada* 15: 113–40. [CrossRef]
- MacWhinney, Brian. 1978. The acquisition of morphophonology. *Monographs of the Society for Research in Child Development* 43: 1–123. [CrossRef]
- Miara, Fernanda Lima Jardim, and Izete Lehmkuhl Coelho. 2015. Participios duplos: Norma, avaliação e uso escrito. *Cadernos de Letras da UFF* 25: 217–37. [CrossRef]
- Mohr, Gordon, Michael Stack, Igor Rnitovic, Dan Avery, and Michele Kimpton. 2004. An Introduction to Heritrix. Paper presented at 4th International Web Archiving Workshop, Bath, UK, September 16; pp. 109–15.
- Padró, Lluís, Miquel Collado, Samuel Reese, Marina Lloberes, and Irene Castellon. 2010. Freeling 2.1: Five years of open-source language processing tools. Paper presented at 7th Language Resources and Evaluation Conference, Valletta, Malta, May 17–23.
- Perini, Mário. 2002. *Modern Portuguese: A Reference Grammar*. New Haven: Yale University Press.
- R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Software. Vienna: R Foundation for Statistical Computing. Available online: <http://www.R-project.org/> (accessed on 1 January 2022).
- Schmid, Hans-Jörg. 2015. A blueprint of the entrenchment-and-conventionalization model. *Yearbook of the German Cognitive Linguistics Association* 3: 3–26. [CrossRef]
- Schwenter, Scott A., Mark Hoff, Eleni Christodulelis, Chelsea Pflum, and Ashlee Dauphinais. 2019. Variable past participles in Portuguese perfect constructions. *Language Variation and Change* 31: 69–89. [CrossRef]
- Villalva, Alina, and Fernanda Jardim. 2018. Participios atemáticos no Português: Tipologia, distribuição e avaliação. *Estudios de Lingüística del Español* 39: 279–303. [CrossRef]
- Queriquelli, Luiz Henrique Milani. 2018. Os participios rizotônicos emergentes no português brasileiro e sua gênese histórica. *Filologia e Linguística Portuguesa* 20: 47–59. [CrossRef]
- Williams, Edwin B. 2016. *From Latin to Portuguese: Historical Phonology and Morphology of the Portuguese Language*. Philadelphia: University of Pennsylvania Press.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Fostes tu?: Analogical Change in European Portuguese and the Case of the Second Person Singular in the Simple Past (Indicative)

Ana Guilherme

Centro Linguística, Universidade NOVA de Lisboa, 1099-032 Lisbon, Portugal; aguilherme@fcsh.unl.pt

Abstract: This paper sets out to study the second person–number marking in the (indicative) simple past in the history of European Portuguese, with a particular focus on morphological innovations such as *fostes tu*, which are considered deviant. These innovations, according to some brief descriptions in the literature (cf. Piel 1989; Williams 1994), are considered a case of morphological change by analogy; however, it remains to be determined whether it is a case of analogical extension or, possibly, leveling that would have resulted in syncretism. Based on data retrieved from private letters from the 16th to the 20th century by almost illiterate authors, we will argue that this innovation results from an analogical extension, motivated by morphological and pragmatic factors.

Keywords: (indicative) simple past; morphological change; analogy; private letters; historical pragmatics

1. Introduction

This paper sets out to study the second person marking in the simple past (indicative) in the history of European Portuguese, with a particular focus on morphological innovations such as *fostes tu*, considered deviant as the canonical morpheme is *-ste*, inherited from Latin. These innovations have not received much attention from the scientific community, although there have been appeals for their study, given their relevance as a linguistic phenomenon (Peres and Móia 1995; Rodrigues 2004). The investigation carried out by Guilherme (2021) provides a more recent account of this phenomenon.

For this study, data was retrieved from two *corpora* of private letters written between the 16th century and the 20th century (until 1974). There is, impressionistically, the idea that this non-canonical use of the second person singular simple past is somewhat recent, and is an ongoing change, but its development can be traced back to the 18th century, as the following example reveals, taken from a private letter written by Francisco Henriques, tailor, to his wife, Inácia de Jesus, in 1791:

(1) [E]u qua resebi huma carta tua_{2sg} cesta feira da comana pasada eu tomei bem atenção no que mandastes_{2sg_innovative morpheme} dizer.

(a letter from *PS—Post Scriptum—Digital Archive of the Ordinary Writing in Portugal and Spain in the Modern Era*)

‘I have received your_{2sg} letter last Friday and I paid well attention to what you send_{2sg_innovative morpheme} me to say’

The innovative morpheme *-stes*, instead of the canonical form *-ste*, inherited from Latin, marking person–number in the second person singular (henceforth 2sg) of the simple past (indicative) (*pretérito perfeito simples*), according to some brief descriptions in the literature (cf. Piel [1944] 1989; Williams 1994), is considered a case of morphological change by analogy; however, it remains to be determined whether it is a case of analogical extension or, possibly, leveling that would have resulted in syncretism.

The morphological changes that have occurred in the history of the Portuguese language, as well as in other languages, are often attributed to analogy, explicitly or implicitly.

Citation: Guilherme, Ana. 2024. *Fostes tu?: Analogical Change in European Portuguese and the Case of the Second Person Singular in the Simple Past (Indicative)*. *Languages* 9: 176. <https://doi.org/10.3390/languages9050176>

Academic Editor: Patrícia Amaral

Received: 2 November 2023

Revised: 22 March 2024

Accepted: 22 March 2024

Published: 11 May 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

But this is not a consensual concept and there is a vast amount of literature on the topic. However, it is still a concept recognized by linguists, although it is often used without proper deepening, and there is often no room for clarification on the types or subtypes of analogy that are referred to (Guilherme 2021). More specifically, there are authors who argue that the manuals of historical linguistics of Romance languages and works that address issues of inflectional morphology need, precisely, further clarification on analogy; namely, they must clarify the question of the directionality of analogical changes (Wheeler 2012, p. 1). This paper also aims to contribute to demonstrate the relevance of the concept of analogy in the context of morphological change and to make evident that a deeper comprehension of the concept may help to understand linguistic change.

To fully understand the phenomenon, it also became important to examine the pronominal forms that accompany second persons of the simple past (indicative) (*tu* and *vós*). Therefore, this research involves an interdisciplinary view—historical linguistics and historical pragmatics—and tries to explain this morphological change by considering the following research questions and objectives:

- (i) To contribute to understanding language change and the role of analogy in language change. As mentioned before, this concept has been used in the field to explain change but “The literature on this theme is vast and controversial” (Lahiri 2000, p. 1). Therefore, does this change result from analogical extension or analogical leveling?
- (ii) Considering (i), what role (if any) did the disappearance of *vós* from the courtesy system, in Portuguese, have in the change?
- (iii) How old is the phenomenon?
- (iv) Will the innovative form *-stes* fully replace the canonical *-ste*? In other words, will the change take place in the sense that a significant portion of speakers will use the innovative construction (Fertig 2013)?

This article is organized into the following sections: Section 2 focuses on the concept of analogy in linguistic studies, referring especially to Hock (1991) and, more recently, Fertig (2013). Section 3 traces, in general lines, the diachronic path of the pronominal forms *tu* and *vós* in the courtesy system in European Portuguese. The methodology and *corpus* are then presented in Section 4. The research data and discussion are presented in Section 5. Finally, Section 6 presents the concluding remarks.

2. Analogy in Linguistics: Definition, Processes, and Problematics

2.1. From the Neogrammarian and the Traditional Approach to Other Frameworks

The morphological changes that have happened in the history of the Portuguese language, as well as in the histories of other languages, are often attributed to analogy, either explicitly or implicitly. The neogrammarians were responsible for the relevance they gave to the concept of analogy as a fundamental concept for the understanding of linguistic change, namely morphological change. The neogrammarian school led, above all, to the recognition that analogy, as the basic principle of human language (and cognition), is fundamental to the understanding of linguistic change. It was also in the neogrammarian context that the idea of proportion associated with analogy—Fertig (2013) points out that proportion is equivalent to ‘where there is a rule’—began to be further developed. Also, at this point, the traditional perspective on analogy, heiress of the neogrammarian school, was presented through Jerzy Kuryłowicz (Kuryłowicz [1945] 1995) and Witold Mańczak (1958), who played a very important role in understanding what happens when an analogy occurs. This traditional approach, which arises from the neogrammarian dichotomy of phonetic change versus analogy, emphasizes its irregular character, but has brought significant advances since it was able to identify some directionality in analogical operations. However, certain concepts used, such as basic form, and less marked, among others, are open to debate.

The notion of expansion from one pattern is present in several definitions of analogy that are conceived precisely around the idea of generalization from one grammatical relation to another, from one word or form to another, or even generalization from one set

of expressions to another (Campbell 1998, p. 89). That is, it always involves the notion of similarity and the notion of appropriation of one linguistic form from another, or of one pattern from another, and can be conceived as ‘internal borrowing’: the idea being that in analogical change a language may ‘borrow’ from some of its own patterns to change other patterns (Campbell 1998, p. 90).

Traugott (2011, p. 25), in an attempt to clarify the concept of analogy, advances two definitions: that of “analogy” and “analogization”. The first refers to analogical reasoning, that is, to analogy as a motivation for change, but which does not necessarily lead to change, while the second refers to the mechanism of any change that takes place (Traugott 2011, p. 25).

Fertig (2013, p. 12), in addition, makes new proposals for the definition of analogy as a technical notion and for analogical innovation. They are proposed definitions that reflect the importance of neogrammarian thinking, but simultaneously reflect a less restrictive view of the concept of analogy and a deeper development of the various definitions and approaches that have come before. Thus, (Fertig 2013, p. 12) proposes the following definitions¹:

Analogy₁ (general sense) is the cognitive ability to reason about relationships between elements in one field of knowledge or based on beliefs about another domain. This ability allows us to make predictions/speculations about objects that we do not know based on the knowledge of objects from another domain in which parallel properties are perceived to exist.

- (i) Analogy₂ (technical sense) is the ability of speakers to produce meaningful linguistic forms from patterns of other forms of language.
- (ii) Analogical formation is the linguistic form (morpheme, word, phrase, etc.) produced by analogy₂.
- (iii) Analogical innovation is an analogical formation and/or the product of an associative interference that deviates from standard usage norms.
- (iv) Analogical change occurs when a significant portion of speakers use an innovative form or set of innovations.

With these revised definitions of analogy, Fertig (2013) presents a broader notion of the concept, in the light of Hock’s (2003). Moreover, for this author, analogy₂ is a basic capacity of linguistic production and not a mechanism (or a type) of change on its own—nevertheless, it is a capacity that can influence change. In the literature, there is the recognition that analogy is a basic principle of human cognition; however, when applied to the understanding of human language, the question of whether it belongs to performance or competence remains.

Regular Analogical Processes: Proportional Analogy, Analogical Extension, and Leveling

Classical studies in historical linguistics often describe and distinguish proportional analogy (or four-part analogy) and extension on one hand and leveling on the other hand as the major regular subtypes of analogical processes. Proportional analogy is often illustrated through a quadripartite equation (cf. 1) that mirrors the extension of a pattern of morphological relationship to forms that did not originally exhibit such a pattern (McMahon 1994, p. 71; Hock 1991, p. 171; Hock 2003, p. 441), focusing essentially on morphological aspects:

$$(1) \quad \begin{array}{l} a: a' \\ b: x = b' \end{array}$$

Simple examples of proportional analogy can be found in English verbs, as in *teach: taught/catch:x?* (Trask 1996, p. 106), or the case of the regular plural in English, as in *dog: dogs/cat: x?* (Hock 2003, p. 441). Another example, taken from Campbell (1998, p. 92), can be found in Spanish. In Spanish, the pronoun of a direct object varies in gender, but the pronoun of an indirect object does not (2). In some non-standard varieties, it can be found and is called *laísmo*, which results from a proportional change where the indirect object can also change in terms of gender distinction (3).

- (2) *Lo ví* ('I saw him'), *la ví* (I saw her)
Le di ('I gave him/her something')
- (3) *Lo ví* ('I saw him'): *la ví* ('I saw her')
Lo di ('I gave him'): x?: *La di* ('I gave her something')

It should be noted, however, that not all proportional analogical changes can undoubtedly fit into this mathematical equation, bearing in mind that linguistic change will not always obey principles as logical and homogeneous as those governing mathematical laws: "Not all cases considered proportional analogy can be easily represented in this proportional formula, and some cases not normally thought to be proportional analogical change can be fitted into such a formula." (Campbell 1998, p. 91). Nevertheless, such changes will not cease to be considered analogical as they still expand a pattern of morphological relationships.

According to Hock (1991, p. 173), proportional analogy will be more successful if it generalizes more productive morphological patterns², even though the concept of productivity itself is open to discussion. What seems to be preponderant is that from the moment a pattern becomes more productive, "[...] it is more likely to be generalized by four-part-analogy than other, less productive categories" (Hock 1991, p. 173).

The term analogical extension often arises associated with proportional analogy, and this association often generates confusion as Fertig (2013, p. 48) warns. For instance, McMahon (1994, p. 71) describes the formation of the regular plural in English in -s as a case of analogical extension. The clarification of both concepts depends on the more or less broad perspective that is adopted, and in broader perspectives 'proportional analogy' and 'analogical extension' are synonymous notions. However, the definition of extension most conveyed in the literature is the diffusion of a paradigmatic distinction (morphological or morphophonological) to other items or contexts where it did not exist (cf. Campbell 1998, p. 94; McMahon 1994, p. 70, e.o.). Another example that fits into the extension is the proportional model that concerns the verb *dive*, in dialectal varieties: the regular morpheme of simple past -ed of strong verbs, as *talk: talked*, was extended to that verb, becoming *dive: dived* (Campbell 1998, p. 94).

Leveling eliminates or reduces morphophonological alternations (generated by phonetic change) within a paradigm (Hock 1991, p. 168; Campbell 1998, p. 92, e.o.). What is eliminated are morphological alternations—in the radical (*stem*) or at the level of other morphemes—which are apparently not preponderant in terms of marking differences in meaning or form (Hock 1991, p. 183; Hock and Joseph 1996, p. 155). One of the most cited examples of leveling is the change that affected the Latin noun *honor*. Some noun paradigms alternated between intervocalic /s/ and /r/ in inflectional forms and this alternation was eliminated through leveling (4) (Albright 2005):

(4)	Before leveling	After leveling
Nom. Sg	[hono:s]	[honor]
Gen. Sg	[hono:ris]	[hono:ris]
Acc. Sg	[hono:rem]	[hono:rem]

Regardless of the perspective adopted, what is important to emphasize is that leveling involves paradigms and extension involves patterns. In any analogical (proportional) change, a distinction is eliminated (leveled) and, at the same time, a pattern is extended (Fertig 2013, p. 48, e.o.). With leveling, what is eliminated are morphophonological alternations—in the radical (*stem*) or at the level of other morphemes—which are not preponderant in terms of marking differences in meaning or form (Hock 1991, p. 183; Hock and Joseph 1996, p. 155).

Analogy, either as a mechanism or as a cause of linguistic change, and arguably as a cognitive capacity, is important to understand diachronic phenomena, as several authors have shown (Hock 1991; Hock and Joseph 1996; Fertig 2013; Fisher 2008). The view in linguistic studies, inherited (unreasonably) perhaps from the neogrammarian position (relevance of phonetic change), of analogy as a vague and poorly defined concept, has

been revised to recognize a need for a better understanding of what analogy really is and its relevance to diachronic research. The studies initiated by the works of Kuryłowicz (Kuryłowicz [1945] 1995) and Mańczak (1958), who sought to describe some systematicity and restrictions on the action of analogy, show that persistency on this path—of relating analogy to perspectives that favor the surface structure of forms as an object of analysis (such as the concept of abduction, in the generative literature, or theories about the iconicity of language)—will only help to better understand the role of analogy in linguistic change (McMahon 1994, p. 96).

3. Pragmatics of Forms of Treatment

In general terms, Portuguese inherited the *tu* and *vos* forms of the late Latin politeness system (Cintra 1986; Faraco 1996, et al. ii). Latin had two pronouns for the second discursive person: *tu* for the singular, regardless of the type of relationship between the participants, and *vos* for the plural. However, the fragmentation of the Roman Empire into two parts (eastern and western) led to a restructuring of the Latin courtesy system (Marcotulio 2014; Lara Bermejo 2018, et al. ii). This fragmentation had pragmatic consequences, since the Empire became represented by two emperors, with *vós* being used to refer to a single interlocutor with courteous value as a result of this division.

The Medieval System

During the medieval period, the pronoun *tu* was used in the context of intimacy or from superior to inferior, marking devaluation and showing ‘derogatory’ value (Cintra 1986, p. 68), and this use has remained practically stable in the context of familiarity until contemporary times. *Vós* was used by an interlocutor in situations requiring greater courtesy, for example, when speaking to the king or some element of the nobility, in other asymmetrical relationships (i.e., from children to parents), but also between husband and wife, or even between brothers (Cintra 1986; Marcotulio 2014). Table 1 below illustrates the forms used during the Middle Ages:

Table 1. Pronominal system of medieval treatment in Portuguese (Lara Bermejo and Guilherme 2018).

	T	V
Singular	Tu	Vós
Plural	Vós	Vós

This scenario changed when nominal forms were introduced into Portuguese society: language reflected these social changes from the 14th century onwards, adopting forms such as (among others) *vossa mercê* (‘your mercy’), which is an item that would later grammaticalize to the current pronoun *você*. The already grammaticalized form has been in use since the 17th century and remained as a form of courtesy until the 19th century (Faraco 1996). In the fifteenth century, nominal forms were already preferred, and in the 16th century, *vós* lost its position until it became archaic in the 18th century (Cintra 1986; Faraco 1996). However, it was mainly in the 16th century that Portuguese underwent profound changes in the treatment paradigm (Bechara 1991; Faraco 1996), possibly mirroring the social changes experienced in the country. The introduction of nominal forms required a restructuring of the treatment system with pragmatic and grammatical consequences since *you* refers to the second discursive person, but occurs, nonetheless, with the morphology of third person singular or third person plural, thus making these innovations complex from a grammatical point of view. In general, the pronominal system of Portuguese treatment between the 16th and 18th centuries is summarized in the following Table 2:

Table 2. Portuguese pronominal treatment system between the 16th and 18th centuries.

	T	V
Singular	Tu	A vossa mercê > você
Plural	Vós	As vossas mercês > vocês

For the disappearance of *vós*, as an address form for a singular interlocutor, the literature points to the 18th century as a chronological milestone (Cintra 1986; Faraco 1996). However, its fall began to take shape in the 16th century, when its use expanded to the lower social classes, and then began to be used in less formal contexts. This pronoun “then gains an archaic and somewhat ridiculous taste from the speech of old or provincial people” (Cintra 1986, p. 30). *Vós* persisted only in certain Northern dialects for a plural interlocutor without deferential value in very ceremonial situations or in religious contexts (Cintra 1986; Faraco 1996, et al. ii).

The current pronominal and nominal treatment system is completely distinct from that of the old period described above, as can be seen in Table 3 below.

Table 3. Pronominal and nominal treatment system in contemporary Portuguese, adapted from Cintra (1986).

	Singular	Plural
Formal treatment	O senhor/A senhora	Os senhores/As senhoras
Non intimate	Você	Vocês
Informal treatment	Tu	Vocês

Regarding the pronoun *tu* and the verbal inflection of 2sg in the context of the treatment system (and in the scope of person–number marking), the literature is equally scarce. It is even sparser regarding the innovations analyzed here with the same few being mentioned without further development. Notwithstanding this scenario, some authors briefly refer to the innovative form, namely Leite de Vasconcellos (1911); Williams (1994); and Piel ([1944] 1989), and argue that non-standard cases as *fostes* are the result of an analogy with the morpheme *-s* existing in the 2sg of the remaining verb tenses and moods, except in the simple past indicative and in the affirmative imperative paradigm.

The stability of *tu* throughout history, in the contexts of intimacy and familiarity, although these contexts have been expanded, especially in the 20th century (for example, between parents and children (Bacelar 2020, p. 2709), probably justifies that the research work has focused on other aspects of the complex treatment system in Portuguese³. However, the morphological innovations studied here and the scarcity of research on them highlight the gap that is found in research on this topic.

4. Methodology and corpus

As mentioned in the introduction, data were collected from *corpora* of private Portuguese letters written between the 16th and 20th century (Guilherme 2021). The *corpora* in question are the *PS—Post Scriptum—Digital Archive of the Ordinary Writing in Portugal and Spain in the Modern Era* (henceforth *PS*) and the *Fly—Forgotten Letters, 1900s* (henceforth *Fly*)⁴. Considering the scope of this study, private letters were used as the main source of the data for various reasons. First, they were used because of their communicative context: these documents are the natural setting for the emergence of second-person usages (Jucker and Taavitsainen 2002, p. 9). Moreover, they follow, to a certain extent, a more formulaic discourse; however, the body of the letters “allows more freedom to the writer” (Jucker and Taavitsainen 2002, p. 9), and these letters are written records very close to orality. Therefore, they are valuable sources for studies in diachrony (Marquilhas 2015, p. 222).

Regarding Portuguese documentation, *Post Scriptum* has a little more than 2000 letters, resulting in a total of 992,878 words in Portuguese. Concerning the number of letters *per*

century, see Table 4, which shows the chronological distribution of the letters⁵ that are available online.

Table 4. Distribution of private letters *per* century.

Century	# Letters
16	245
17	547
18	758
19 (up to 1833)	665

It is necessary to highlight and clarify some limitations in this *corpus*. First, it is evident that the amount of documentation is not equitable; the 16th century has fewer documents when compared, for example, with the 18th or 19th centuries. Second, the documentation from the 19th century is limited to the year 1833. Most of these documents were collected from numerous court cases in two types of courts: ecclesiastical courts and a civil court called the Casa da Suplicação⁶. The ecclesiastical courts belonged to the Portuguese Inquisition, which was in force in the territory between 1532 and 1821. The private documents were filed in legal proceedings, either in inquisitorial proceedings or in proceedings of the Casa da Suplicação⁷, because they were seized from the defendants, thus serving as evidence of an alleged crime or some heresy. However, these letters both contained truly fraudulent content (e.g., letters of extortion) or only represent writing that bears witness to the daily life of ordinary people: “[P]eople who were writing on a criminal impulse, as is the case of extortions (...) or were just elaborating on everyday matters.” (Marquilhas 2015, p. 223). In fact, the existence of these letters in these judicial proceedings is the result of pure luck: “[S]ome randomness was indeed involved in specific letters finding their way into the courts’ archives while others did not” (Marquilhas 2015, p. 223). This randomness clearly causes difficulty in locating documents, notoriously for the 16th century (Vaamonde 2018, p. 147).

The documents of the 19th century are essentially more numerous for two reasons: the population was larger than the 16th-century population and the illiteracy rate was lower, resulting from the educational reforms initiated by the Marquis of Pombal (Marquilhas 2015, p. 226). Finally, the documentation only extends until 1833 because this was the last year of operation of the Casa da Suplicação.

The *corpus Fly—Forgotten Letters—Years 1900–1974*⁸ results from the collection, transcription, and editing of 2000 private letters that represent very specific contexts: war, prison, exile, and emigration, mainly because they were contexts that favored communication by letter. The documents come from public and private archives. The total number of words is around 700,000, and the letters are written by 572 different authors, most of whom are men (78%), although women were the most common recipients of the correspondence. Most of the letters are of private content, addressed to relatives, friends, and boyfriends, among others, while a minority (140) are of a more formal nature.

Various types of private relationships between correspondents were selected, and the results obtained are shown in Table 5:

The private relationship most often portrayed is that between friends (81; 27%), followed by letters from a husband to his wife (57, 19%), and letters exchanged between siblings (46, 15%). The category ‘other’ refers to cases where it was not possible to determine the nature of the relationship between participants, for instance, in cases of extortion or in a few cases of a more institutional relationship between interlocutors that escape the sphere of family or friendship.

Table 5. Relationship between the participants of the private correspondence.

Relationship between Participants	# Letters	%
Friends	81	27%
Husband > wife	57	19%
Siblings	46	15%
Lovers	29	10%
Wife > husband	20	7%
Uncle > nephew	15	5%
Father > son	13	4%
Mother > son/ daughter	13	4%
Other	12	4%
Cousins	6	2%
Brother/Sister-in-law	4	1%
Son/ daughter> father	2	1%
Nephew > uncle	0	0%
Total	0	0%

Bearing in mind that the main objective of this research is to understand the development of a specific morphological change—the innovative forms of 2sg as *foſtes*, *diſſeſtes*, *fiſteſtes*, etc.—through a necessary discussion of the concept of analogy in language change, it was fundamental to extract all verbal forms ending in *-ſte* and *-ſtes* (2sg and 2pl, simple past indicative). As mentioned in the introduction, it became relevant to observe the development of the 2pl of the simple past indicative and understand what effect such development might have had in the change in question, if any. For the sample to be as balanced as possible, and considering that some authors wrote several letters, one letter per author was randomly selected. Then, a database and metadata were populated using *FileMaker* software.

5. The Innovative Morpheme *-ſtes*—Description and Discussion

592⁹ examples of verbal forms of the 2sg and 2pl of the simple past (indicative) were extracted, which corresponds to 298 epistolary documents, that is, 298 different authors, as showed in Table 6:

Table 6. Documents used for the study per century.

Century	# Letters
16th	36
17th	28
18th	24
19th	47
20th	163
TOTAL	298

5.1. Second Person Singular (2sg) forms of Simple Past Indicative: Canonical Morpheme *-ſte*

Regarding the 2sg forms of the simple past indicative, that is, forms with the canonical morpheme *-ſte*, there were 276 occurrences in the 298 private correspondence:

- (5) “(..) triste mosa nūqua **tu**_{2sg} naseras **estiveste**_{2ps} pa morer.”
 ‘sad girl **you**_{2sg} will never be born you **were**_{2sg} to die’
 [letter from Jerónima dos Anjos to her sister, Catarina de Paiva, 17th century, PS]

These occurrences are distributed chronologically as indicated below in Table 7 and graph 1, respectively:

Table 7. 2sg forms of simple paste (indicative) (canonical morpheme *-ste*).

2nd Simple Past Indicative <i>-ste</i>	Occurrences	%
16th	5	2%
17th	8	3%
18th	11	4%
19th	22	8%
20th	229	83%
TOTAL	276	100%

With regard to these forms, it appears that their use has increased as time has advanced, which is not unexpected given the historical evolution of forms of treatment (see Section 3), that is, familiar treatment through verbal, pronoun, and second-person nominal forms of the singular became more frequent from the 19th century and then became practically stabilized in the 20th century. As such, the number of occurrences of 2sg forms of the simple past (indicative) rises from the 19th century, but it is in the documentation of the 20th century that we observe an exponential increase—83% of 2sg forms of the simple past (indicative) are found in the twentieth century.

5.2. Second Person Plural (2pl) Forms of Simple Past Indicative

As for the 2pl forms of the simple past indicative, **185** occurrences, shown in Table 8, like the example in (5), were recorded from the documentation analyzed.

- (6) “Bem **sabeis**_{2pl} que me **deixastes**_{2pl} em casa de meu pai.”
 ‘**You**_{2pl} well **know**_{2pl} that you have **left**_{2pl} me in my father’s house’ [letter from Vicência Jorge to her friend, Jerónimo Monteiro, 16th century, PS]

Table 8. Second person plural forms of simple past (indicative) (canonical morpheme *-stes*).

2nd Simple Past Indicative <i>-stes</i>	Occurrences	%
16th	108	58%
17th	50	27%
18th	14	8%
19th	11	6%
20th	1	1%
TOTAL	185	100%

These occurrences are distributed chronologically as follows (Table 8):

In relation to these forms, the results are also not surprising, in that it is to be expected that the use of 2pl forms will decrease, especially from the 18th century, taking into account the historical development of the verbal and pronominal paradigm of 2pl in standard modern Portuguese.

5.3. Morphological Innovation: Second Person Singular Forms of the Simple Past (Indicative)—The Morpheme *-stes*

Regarding the verbal forms that are the main object of this study, that is, 2sg forms of the simple past (indicative) with the innovative morpheme *-stes*, for person–number marking, 131 cases were registered, as exemplified in the innovative form *mandastes* in (7):

- (7) “eu qua resebi huma carta **tua**_{2sg} cesta feira da comana pasada eu tomei bem atenção no que **mandastes**_{2sg} dizer.”
 ‘I have received **your**_{2sg} letter Friday, last week, a I’ve paid much attention to what you **send**_{2sg} me’ [letter of Francisco Henriques, taylor, to his wife, Inácia de Jesus, 18th century, PS]

Table 9 and Figure 1 illustrate the distribution of these forms over the different centuries:

Table 9. Number of morphological innovation occurrences (morpheme *-stes*) per century.

Morphological Innovation <i>-stes</i>	Occurrences	%
16th	0	0%
17th	0	0%
18th	10	8%
19th	38	29%
20th	83	63%
TOTAL	131	100%

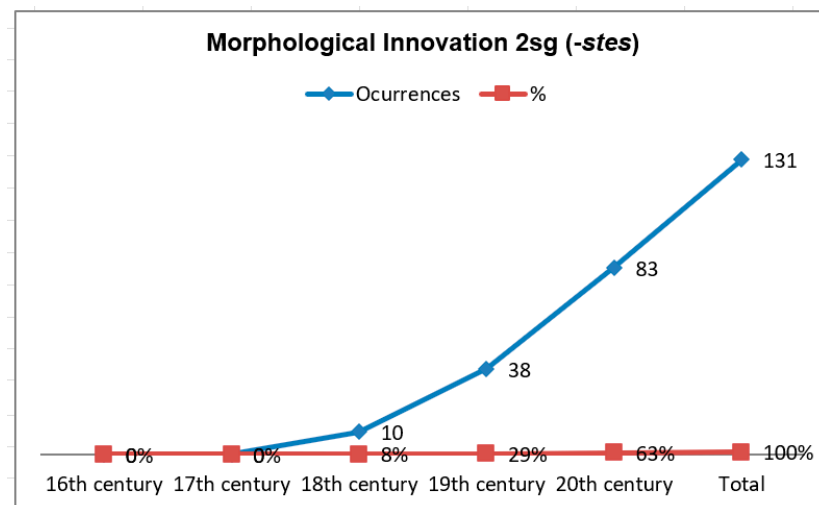


Figure 1. Distribution of occurrences of morphological innovation.

It is clearly observable that the use of these forms has increased considerably in the 20th century, with 63% of cases concentrated in this period. However, these data show that it is not a recent innovation, since these constructions are observable starting from the 18th century, and in the transition from the 18th century to the 19th century, the rise from 8% to 29% of cases is also significant.

5.4. Second Person Usage throughout the Centuries

When comparing all the tokens from 2sg of the simple past (indicative), singular and plural, the growth of 2sg is notorious, with the innovative or the canonical morpheme, as shown in Table 10 and Figure 2.

Table 10. Number of tokens of 2sg simple past (indicative) throughout the centuries.

Century	2sg <i>-ste</i>	2sg <i>-stes</i>	2pl <i>-stes</i>
16th	5	0	108
17th	8	0	50
18th	11	10	14
19th	22	38	11
20th	229	83	1
Subtotal	275	131	184
TOTAL	406		184

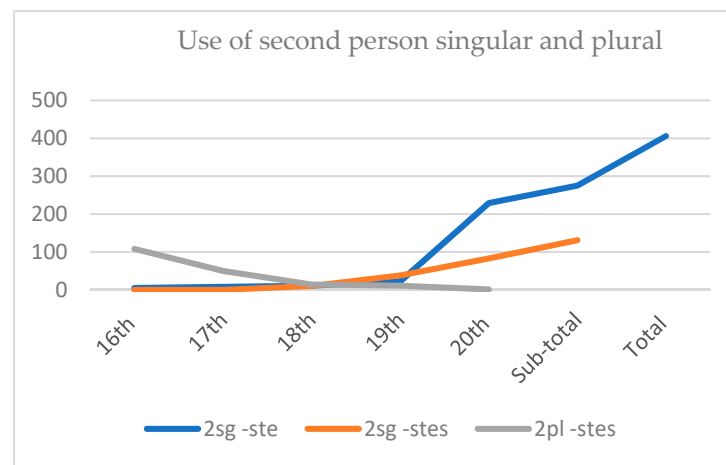


Figure 2. Use of 2sg and 2pl throughout the centuries.

As can be seen, the usage of 2sg, with the innovative or with the canonical morpheme, has been greater than the use 2pl, except in the 16th and 17th centuries. From the 18th century, it is observable that there was a growth of 2sg, which was followed by the decrease in 2pl. These data meet, in general terms, what is described in the literature concerning the use of second person in Portuguese for the treatment system. However, we observe a usage of *vós* to a singular recipient beyond the 18th century.

5.5. Variation between the Canonical Morpheme (*-ste*) and Innovative Morpheme (*-stes*) in the Person–Number Marking of the Second Person Singular

Regarding cases of variation between the use of the regular morpheme *-ste* and the innovative morpheme *-stes*, 25 cases were reported, identical to the example in (8), where the same author uses *fizeste* (canonical form) and *deixastes* (innovative form):

- (8) Já **fizeste**_{2sg} contas com **teus**_{2sg} pais? Durmo mal, passo a vida à procura das coisas que nunca sei onde estão, pois não há forma de fixar os lugares em que as **deixastes**_{sg}.
 ‘Have you **settle**_{2sg} accounts with **your**_{2sg} parents? I sleep badly, I always looking for things that I never know where they are, because I can’t remember the places where **you**_{2sg} **left**_{2sg-stes} them’
 [letter from a husband to his wife, 20th century, Fly]

The distribution of these cases is shown in Table 11 and Figure 3 below:

Table 11. Number of occurrences of variation between the morpheme *-ste~-stes*.

Variation between <i>-ste/stes</i> (2sg)	Ocurrences	%
16th	0	0
17th	0	0
18th	1	4%
19th	2	8%
20th	22	88%
TOTAL	25	100%

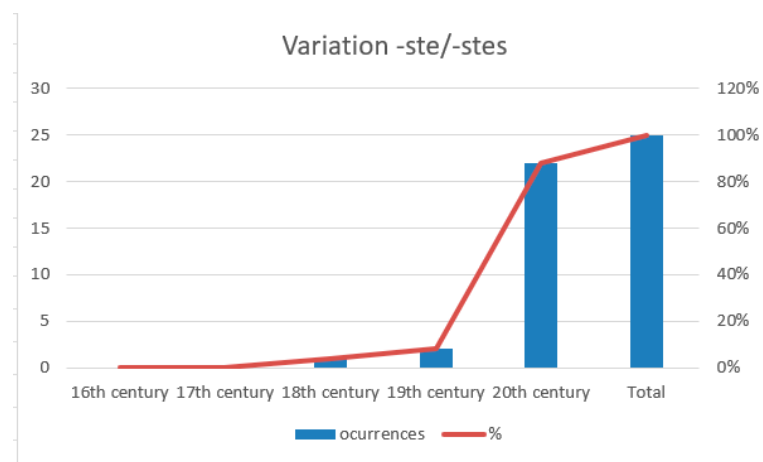


Figure 3. Variation between morpheme *-ste~-stes* over the centuries.

Most of the examples clearly relate to the 20th century, with **22** occurrences (88%), while there were only two cases in the nineteenth century (8%) and one case in the eighteenth century (4%).

5.6. Variation between Second Person Singular and Second Person Plural

The data also attest 11 clear cases of variation between the use of 2sg and 2pl, as illustrated in (9):

- (9) O portador desta é hūomem de setuvela Manoel Rodrigues cazado em Matosynhos **olhay**_{2pl} se **Podes**_{2sg} falar com ele.
 ‘The carrier of this letters is a man from Setúbal called Manuel Rodrigues married in Matosinhos **see**_{2pl} if you **can**_{2sg} talk to him’ [letter from Bartolomeu da Costa to his wife, Madalena Francisca, fishmonger, 17th century, PS].

In (9), the verbal form *olhay* is the imperative of 2pl, whereas *podes* corresponds to the simple present of the 2sg, here used to convey the request initiated with the imperative *olhay* (‘please look’). Nevertheless, in this document and others, this usage of the second person is a clear case of alternation between 2sg and 2pl.

5.7. Other Phenomena Associated with the Use of Second Person

If the data in Section 5.6. attest for clear cases of variation between 2sg and 2pl, some ambiguous forms were found, in cases of 2pl. In Portuguese, the 2pl forms were affected with a specific morphophonological change. This change is the loss of the suffix *-d* (*-de/-des*) present in all 2pl forms, except in the simple past indicative (*-stes*). This loss was registered mainly in intervocalic contexts (*amades* > *amais*, ‘love’, simple present, 2pl) and it was preserved in non-intervocalic cases, namely in inflected infinitive and subjunctive future. However, the deletion of *d-* caused an increment of variation of forms with or without *d-*, even in non-intervocalic contexts, as, for instance, in cases of inflected infinitive (*dizerdes*

> *dizeres*, ‘to say’, 2pl). Middle and Classical Portuguese exhibit great variation between forms. The period of this study points precisely to variation (9) between forms with *-d* and without *-d* (*fazeres/fazerdes*) in contexts where *-d* is preceded by a vibrant (Said Ali [1921] 2001; Brocardo 2006; Brocardo and Lopes 2016). Consequently, forms of 2pl without *d-*, in subjunctive future in inflected infinitive, are syncretic with 2sg—*tu fazeres* (‘you to do’). Therefore, it is not possible to ascertain if verbal forms such as *saveres*, in (10), are really 2pl or 2sg:

- (10) “**Vos**_{2pl} quero avizar de tudo para **saveres**_{2sg/2pl?} o que se passa.”
 ‘I want to let **you**_{2pl} know about everything for you **to know**_{inflected infinitive 2sg/2pl?} what’s going on’ [letter from Helena Costa, to her husband, Francisco Rodrigues, 1654, PS].

The tendency to delete *d-* in these contexts did not generalize, probably to maintain the pragmatic difference between 2pl and 2sg, particularly in a time where 2pl forms were used to address a single interlocutor.

Data also attested cases of ambiguity, generated by a possible clash of the 2pl morpheme in contexts that will later generalize in diphthong, for example, *foses* (11) (*fosses* instead of *fosses*). Again, the possible clash of 2pl morpheme generates syncretism between 2sg and 2pl (‘*tu fosses~vós fosses*’, ‘if you were’, second person imperfect subjunctive); therefore, it is also not possible to determine if in (10) the verbal form refers to a 2sg or 2pl:

- (11) “[A]Conselhandovos_{2pl} q foses_{2pl/2sg?} judeu como eu era (...)”.
 ‘I would advise **you**_{2pl} to **be**_{2pl/2sg?} jew as I was.’ [letter from Diogo da Horta, merchant, to his brother, Fernão da Horta, merchant, 16th century, PS]

5.8. Summary of the Main Findings in the Private Letters

- (i) The frequency of 2pl verb forms of the simple past (indicative) decreases as we progress chronologically and, conversely, the frequency of 2sg verb forms of the simple past (indicative) increases.
- (ii) Naturally, and considering point (i), the frequency of pronominal forms of 2pl and the 2pl morphology decreases over time, while the frequency of pronominal forms and morphology of 2sg increases.
- (iii) The innovative morpheme *-stes* for the marking of person–number of the 2sg of the simple past (indicative) is evident, albeit incipiently, from the 18th century, increases during the 19th century, and it is then in the 20th century that most cases are concentrated.
- (iv) Several complex phenomena related to the use of second person verb forms were located and described. These are cases of ambiguity between 2pl and 2sg generated by a possible deletion in non-intervocalic *-d* and by possible realizations of a clash. Other cases are related to alternations of use between 2sg and 2pl or some evidence of competition between forms of 2sg and 2pl, but the number of examples is small (11). All these cases represent different types of results in which factors of different levels, grammatical and pragmatic, are interrelated.
- (v) In the set of cases of variation, the one in which there was a greater number of examples (25 cases) is related to the person–number marking of the 2sg of the simple past (indicative), in which alternation between the canonical morpheme *-ste* and the innovative morpheme *-stes* is displayed.
- (vi) The innovative morpheme *-stes* for 2sg person–number marking was in documents written by authors from or residing in all parts of the territory, including in 2pl maintenance areas (see Discussion).
- (vii) Observable maintenance of *vós* and the 2pl for an interlocutor beyond the 18th century, even if this use has decreased over the centuries.
- (viii) Diachronically stable and persistent use of *tu* for the axis of informality, whose use increases and stabilizes from the 19th century.

- (ix) Preference for verbal inflection, both 2sg and 2pl, and less use of explicit subject pronouns, *tu* and *vós*. This preference can be explained in pragmatic terms since the singular pronominal forms are more ‘marked’ because they individualize the interlocutor, while communicative interaction mediated through verbal inflection is, in languages of null subject, considered as a more neutral and courteous strategy.

6. Discussion

Several research questions were presented, namely: (i) how old is the phenomenon? (ii) what kind of analogical change—extension or leveling (resulting in syncretism) is at the root of these innovations? (iii) hence, and following (ii) what influence did the disappearance of *vós* and the morphology of the 2pl in standard Portuguese have on this change? (iv) will these innovations spread through the speakers in such a way that might be considered a change in the language? The following discussion of the findings will try to answer these research questions.

The morphological innovations of the ‘*fostes tu*’ type are analogical-type innovations, based on Fertig (2013). This author proposes revised definitions of some concepts, namely the notions of **analogical innovation** and **analogical change**, which were adopted in this work, following Guilherme (2021). As indicated in Section 2, **analogical innovation** emerges when an analogical formation—i.e., a form resulting from analogy that is produced by speakers from an existing standard in the language—deviates from the ordinary pattern of use. **Analogical change** refers to an analogical innovation that is adopted by a significant portion of the speakers in a community (Fertig 2013, p. 12). The innovations of the ‘*fostes tu*’ type correspond to analogical innovation, in the sense of Fertig (2013), since they diverge from the morphological pattern for the marking of person and number of the second person singular of the simple past (indicative), whose historical morpheme marker for the grammatical category person–number is *-ste*. However, and also in the light of Fertig (2013), it is questionable that these innovative forms will correspond to an analogical change. It is an observable variation between the canonical and the innovative forms, but an effective change in language means the adoption of a form by a considerable portion of speakers (Fertig 2013). According to Villava and Mateus (2006, p. 72), the conditions are met for this change to take place, that is, when they are implemented by a large number of speakers. Although these innovative forms are heard more often¹⁰, even in people with some level of education (as Piel [1944] 1989 has already noted), it seems difficult to indicate whether or not a more permanent change will emerge. ‘*Tu fostes, tu fizestes. . .*’ are considered ‘wrong’ forms in standard Portuguese and are therefore assessed negatively. The personal letters from which the data were obtained are mostly written by speakers with a low level of education, who therefore did not receive a formal education, which, as it is well known, is marked by normative pressure. Thus, the negative judgment associated with these forms and normative teaching may contain this change.

As stated, it should also be discussed whether or not this is a case of analogical extension or leveling (in the latter case, leading to syncretism). As leveling is generally understood as a process that eliminates or reduces morphophonological alternations in the radical within a paradigm, alternations resulting from phonetic change, are, therefore, motivated by phonological or morphophonological aspects (Hock 1991, p. 171). Fertig (2013) supports this definition, arguing that leveling only eliminates (totally or partially) the alternations in the radical of an inflectional or derivational paradigm, and innovations of the ‘*fostes tu*’ type affected the inflectional suffix, not the radical. In other words, analogical extension acts in very well-defined contexts, that is, at the level of a morphological pattern and leveling at the level of the paradigm. Hock (1991) admits the possibility of changes by leveling that reach inflectional affixes, like in the case of person-marking affixes, because they had stopped fulfilling their function, as happened in English. In most analogical changes, it is possible to determine, with some degree of clarity, whether they result from leveling or from analogical extension (Hock 1991, p. 179).

6.1. 'Fostes tu'—Morphological and Pragmatic Motivations for Analogical Extension

Regarding the date of these innovations, it is possible to affirm that this is an old phenomenon, because, as indicated in the data, they can be attested to from the 18th century onwards. Although at this time they are still relatively less significant innovations, whose frequency of occurrence later increases, especially from the 19th century, and then grows exponentially in the 20th century.

Another factor that points to the antiquity of the phenomenon is related to the geographic distribution of these innovations, since, as explained, they are present throughout the Portuguese linguistic area. All the verb forms are examples of the innovative morpheme of 2sg, *-stes*.

- (12) "Meu Amado e querido espozo mto e mto Do meu Coração (...)huma **tua**_{2sg} resebi (...).
Dise que não tinha avizo no navio do tal home mas contudo que o dava do dinheir que me **mandastes**_{2sg}."
'My dear and beloved husband from all my heart (...) I've received a letter from **you**_{2sg} (...) he said that there was no sign from that man on the ship yet he would give me the money **you**_{2sg} **have**_{2sg} sent me'
[letter from Margarida Rosa, Lisbon (mid centre dialects), to her husband, 1791, PS]
- (13) Resevi as **tuas**_{2sg} notisias que mto estimei (...)tenhoos mto goardados so pa os dias que me **mandastes**_{2sg} dizer.
'I have received **your**_{2sg} news for which I have really appreciated (...) I have it (medicines) kept only for those days **you**_{2sg} **have**_{2sg} told me (...).
[letter from Maria, from Viana (northern dialects) to her lover, 1760–1769, PS]
- (14) Snr eu tenho huma denuncia (...) a **Seo**_{3sg} Respeito (...) **Vmce he**_{3sg} mão homem e q **he**_{3sg} jácobino [...] e q **és**_{2sg} de mão Sangue e ate **metestes**_{2sg} hum dezortor num Capitullo [...].
'Sir, I have a complaint concerning **you**_{3sg} (...) '**Your mercy is**_{3sg} a bad man and **you are**_{3sg} a jacobine (...) and **you**_{2sg} are bad blood and you_{2sg} even **received**_{2sg} a deserter (...).'
[letter from António Maria Vidal, robber, to Manuel Vaz Lampreia, farmer, from Beja (south interior dialects), 1821, PS]

Examples (12)–(14) attest the existence of the innovative morpheme *-stes* in all parts of the Portuguese territory, including in areas where the use of *vós*, as an address form for a plural interlocutor, remains (12). Example (14) also shows variation between the third sg pronominal and verbal forms (*Vossa Mercê*, 'Your Mercy', *he*, 'is') and 2sg verb forms (*és*, 'you are'; *metestes*, 'you received').

As mentioned in the introduction, the morphological innovations of the 'fostes tu' type are often treated as a case of analogy with the morpheme of person-marking *-s* present in the remaining second person of the singular. However, the literature that addresses this phenomenon does not deepen our understanding of this topic, failing to discuss in more detail which subtype of analogical change it represents (cf. Piel [1944] 1989; Williams 1994).

The main data, taken from the private letters, reinforces the hypothesis that it is a morphological change by analogical extension. Analogical extension, also often called proportional analogy, expands an existing pattern in the language to forms that did not previously follow it. That is, one of the motivations for the analogy is the existence, from the outset, of productive and frequent patterns (Fertig 2013; Hock 1991). In other words, the analogical extension will be more successful if it enlarges a more frequent and productive pattern (Hock 1991, p. 173). When observing the diachrony of *tu* and the verbal morphology of 2sg, we verified that these forms were always persistent, productive, and stable, both in morphological and pragmatic terms. In the development from Latin to the different Romance languages, including Portuguese, verbal infection did not undergo major changes (Maiden 2011, p. 156; Brocardo 2014, p. 76) apart from French, for example. That is, the morpheme *-s* for the marking of person–number in the second person singular persisted through the transition from Latin to Portuguese, with the exception of the 2sg of the simple past (indicative), which also inherited the morpheme *-ste* from Latin. Although historically we find this persistence for the 2sg person–number marking—the morpheme *-s* for the

remaining verb tenses, and the morpheme *-ste* for the simple-past (indicative)—the 2sg person–number marking of the simple past appears to be opaquer. This opacity is reflected in the different morphological analyses that it raises: some authors analyze the morpheme *-ste* as an amalgam of time–mood–tense and person–number (Villalva 2003, p. 936), others see it as a person–number morpheme (Câmara Júnior 1979; Mateus 1991), and still others understand the 2sg as having a person–number morpheme present, but with zero marking for the time–mood–tense (Mota 2020). It should also be recalled that (Siewierska 2004, p. 38) indicated that this type of morpheme, which condenses several grammatical functions, is the most complex to analyze. It was the irregular nature, of less transparent character for the speakers, of the morpheme *-ste*, which was decisive in triggering the emergence of innovations of the *'fostes tu'* type.

Pragmatically, *tu* has always been a productive and stable form. As was described in the history of the system of treatment in Portuguese (Section 3), in the axis of informality and familiarity, we always find *tu* and the verbal morphology of 2sg. Nevertheless, it was necessary to consider the hypothesis of the influence of *vós* and the 2pl on the analogical innovation studied here. We know that *tu* and *vós* competed, but only for pragmatic reasons, because as the data show, *vós* was also used in the context of informality since in personal letters we find this treatment for an interlocutor. However, although *vós* and *tu*, and the respective verbal inflection, were competitors for the same discursive contexts, *vós* and the 2pl verb forms did not establish themselves as informal treatment strategies. The stability of *tu* has won. Moreover, diachronically, the greatest competitions took place between forms that mark formality, that is, between *vossa mercê/você* ('your mercy' and similar) and *vós*, and in the plural paradigm, between *vós* and *vossas mercês/vocês*.

Along with these qualities of persistence, stability, and productivity that we find related to *tu* and the verbal morphology of 2sg, and relevant characteristics to serve as an analogical model to be expanded to other forms, it is observable that the frequency of the innovative forms increased in the 19th century and grew significantly in the 20th century. This development coincided with the obsolescence of *vós* in the standard language. We know that the treatment by *vós* and the 2pl for the singular disappears from the 18th century onwards (Cintra 1986), and the main data analyzed are in line with what is described in the literature: in personal letters, the use of *vós* and the 2pl is predominant in the 16th and 17th centuries, but begins to decline from the 18th century, and in the 19th century, we find only two authors who use this treatment for a singular recipient. If we return to Table 8 and Figure 2, the data exhibit precisely this drop of usage of 2pl to a singular recipient in the context of familiarity, for which these documents are a rich testimony.

So, the fall of *vós* from the standard Portuguese, on the one hand, and the increase in the frequency of the use of *tu* and the inflection of the 2sg to the axis of informality, especially from the 20th century, on the other hand, means that *vós* and the 2pl are no longer able to serve as an analogical model, thus strengthening the analogical pattern of the *-s* for the remaining second person singular cases and causing it to be extended.

Another aspect that seems to favor the hypothesis that this morphological change is the result of an analogical expansion, and not a consequence of leveling resulting in syncretism between the 2sg and the 2pl of the simple past (indicative), is related to the cases of variation found. It would be expected that, at a time when *tu* and *vós* were competing for the same contexts of use, we would find forms in competition or eventually some non-canonical agreements, such as *'vós fizeste'*. However, in the 592 examples analyzed, we found only 83 occurrences of alternations in the private letters, corresponding to 14% of all the material studied. Among these 83 cases, only 27 correspond to evidence of possible alternation between 2sg and 2pl inflection, namely: 14 cases of ambiguity in 2pl forms (11), another eight occurrences of forms whose reading is ambiguous, and only five cases of clear variation between 2sg and 2pl uses (12). It should be noted that the examples we have identified as ambiguous reading, as mentioned above, may not represent examples of forms in competition. The other 25 cases of variation raised correspond to alternation between the uses of the 2sg of the simple past (indicative) with the canonical morpheme

-*ste* and with the innovative morpheme *-stes*, evidencing variation in the marking of the 2sg of the simple past (indicative). This type of variation, as found in (15), registered the most examples, 25 cases. In the following example, the author varies between the canonical morpheme, *fizeste*, with the innovative one, *fizestes*, *fostes*.

- (15) Eu ja **te**_{2sg} estou muito obrigado pelo favor que me **fizeste**_{2sg} (...) que me has de pagar quanto me **fizestes**_{2sg} porque **fostes**_{2sg} a cauza da minha perdecão [...].
 'I am really thankful for the favour **you**_{2sg} **have made**_{2sg} me (...) you shall pay everything **you have done**_{2sg} to me because you are_{2sg} the cause of my lost (...).'
 [menace letter from João Machado to Francisco Ribeiro, Guimarães, 1818, PS].

In view of the above and of the diachronic data, the innovative morpheme *-stes* for 2sg person–number marking seems to result from an analogical extension, and not from leveling, considering the characteristics involved in leveling. We have seen that a productive and frequent pattern will be more likely to extend to other items. The 2sg paradigm, in Portuguese, has always been characterized by its stability, productivity, and frequency, especially since the 19th century. Finally, analogical extension tends to result in morphological changes and leveling in morphophonological changes (Hock 1991, p. 171). The phenomenon analyzed here clearly fits into a morphological change by analogical extension. There are also less transparent changes in which it is not possible to determine whether we are faced with an extension or a leveling, and, as Hock (1991, p. 179) observes, these processes do often cooperate. The disappearance of *vós* may have played a role in this morphological change in the sense that its tendency towards obsolescence will have paved the way for analogical extension by reducing the likelihood of convergence between grammatically and pragmatically distinct forms.

The Galician and Spanish Cases

Another supporting argument favoring analogical extension could be taken from the Galician and Spanish languages. In Galician, the innovative morpheme *-ches/-stes* to mark 2sg simple past (indicative) (*cantaches*) can be traced since the nineteenth century, and its diffusion is advancing rapidly. The innovative form is, in fact, the variant found in oral standard Galician (Paz Ramón 2013, pp. 356, 364). These innovative forms live at the same time with the fully preserved 2pl morpheme for simple past (indicative) (*cantastes*), pointing to the fact that the extension of *-s* is independent from preserving or not preserving 2pl. The same scenario is found in Spanish and is similar to what happens in Portuguese, and following Lapesa (2000, p. 470) "(...) no faltan otros ejemplos [*hicisistes*, *dijistes*] en la literatura moderna, pero el uso los condena". This seems not to be the case in Galicia, as this change has been socially 'accepted' by the speakers.

6.2. 'Fostes tu'—A Proposal for Analogical Extension

I have argued that the morphological change under analysis, the morpheme *-stes* as a person–number marker of the 2sg of the simple past (indicative), results from an analogical extension process. As indicated in Section 2 of this paper, analogical extension is often described in the form of a proportional scheme, which is recalled (a) as:

- (a) a: a'
 b: x = b'

However, the case of analogical extension discussed here can hardly be translated into a proportional equation as in (a), or, at least, it does not appear to be possible to establish a linear relationship by means of a proportional equation (in the terms referred to in Hock (1991, p. 172)) that somehow formalizes the change that occurred. For the equation to operate, it is necessary to have a base, or pivot, from which other forms of the paradigm are formed. Bases are usually described as the 'basic' forms of the paradigm. The forms considered 'basic' are often indicated as the least marked, i.e., those which contain less information and have less affixes to the paradigm (Hock 1991). However, there is truly no consensus on what makes one less marked (Fertig 2013, p. 117; Albright 2005). There is, however, some agreement among scholars about the singular as being the least

marked for the category of number, the present for the category of time, and the indicative for the category of mood (Albright 2005; Fertig 2013). In other words, the third person singular of the present indicative is generally accepted as being the most basic (Bybee 1985; Albright 2005). In natural languages, following Siewierska (2004) and Cysouw (2001), the third person singular is the one that most often presents a zero morpheme, that is, it does not convey much grammatical information. Of course, much will depend on the theoretical framework with which one works. However, although there is also no consensus on whether the most marked is the first person or the third of the singular, the least marked and most polite form, in pragmatic terms, is also the third person singular. Other researchers note that a basic form has nothing to do with being or not being more marked, but is related to the frequency of tokens. And, often, these phenomena can converge, resulting in the basic form being the least marked, containing less affixes, and being the most frequent (Albright 2005).

For the morphological change under study and considering, then, that third person is the basic form of the paradigm from which the analogical extension operates, when we formalize the change through the proportional equation, we obtain an agrammatical form, where in (b) is an example for the first conjugation paradigm and for the second conjugation paradigm:

(b)		
Third person singular—First conjugation		Second person singular—Second conjugation
ama ('he/she loves')		amas ('you love')
amava ('he/she loved/would love')		amavas ('you loved/would love')
amará ('he/she will love')		amarás ('you will love')
amou ('he/she loved')		*amouste (s) ('you loved')
Third person singular—Second conjugation		Second person singular—Second conjugation
come ('he/she loves')		comes ('you love')
comia ('he/she loved/would love')		Comias ('you loved/would love')
comerá ('he/she will love')		comerás ('you will love')
comeu ('he/she loved')		*comeuste (s) ('you loved')

In morphological terms, particularly with regard to person–number marking, third person singular is characterized by the absence of a person–number morpheme, that is, it carries a \emptyset morpheme, while *-s* is the person–number mark of all second person forms of all tenses, with the exception of simple past (indicative)¹¹, which is, precisely, *-ste*. Thus, taking into account the person–number marking morphology of 3sg and 2sg, we propose that the abstract proportional rule below can better interpret this analogical extension:

(c) $3\text{sg}: -\emptyset_{\text{PN}}$
 $2\text{sg}: x' = -s_{\text{PN}}$

The equation in (c) gives an account of the pattern of relationship between forms for the marking of person–number in Portuguese, that is, if \emptyset marks the 3sg then *-s* marks the 2sg. This will be the pattern of relationship that speakers identify in their grammar and therefore, by association, extend the *-s*, an unequivocal mark of the 2sg, to the only morpheme that does not present this regularity—the *-ste* morpheme. Apparently, the speakers, in the case of second person, favor the marking of person–number for the singular (cf. Rodrigues 2004, p. 18).

Although it is proposed in the equation in (c), it is necessary to take into account what Fertig (2013) observed regarding the concept of “proportional”, namely, that a change is still “proportional” even if it does not fit the quadripartite model and, once again, the discussion around this concept is closely related to the theory with which it works and with a more or less comprehensive view of analogy¹². However, these changes remain analogical because it is observable that a pattern of morphological relationship was expanded to forms that did not previously have it.

In (c), it is suggested a formalization that accounts for the morphological change under study—the person–number marking of the second person singular of the simple past (indicative). However, it is also worthwhile to ponder the preservation areas of *vós*

and the 2pl and where is possible to find *vós fizestes* and *tu fizestes*¹³, and to consider if these syncretic forms result from the sub-process of leveling. However, there are indications that, also in these areas, the form *tu fizestes* results from the process of analogical extension already described. Although *vós + -stes* and *tu + -stes* have been registered, the cases are scarce and there is no competition between them in the treatment system. Nonetheless, the occurrences of *vós* and the 2pl are mainly for the plural and it is precisely within the plural that more examples of variation have been found, showing that *vós* and the 2pl compete with *vocês* and the 3pl (as mentioned above). There is also evidence that the subsystem of treatment and verbal inflection that we find in the areas of preservation of *vós* is changing, as already described by Aguiar and Paiva (2017)¹⁴. That is, for these subsystems, it is not clear that the analogical model comes from *tu*, or that it is the point of origin for the cases of syncretism between the 2sg and the 2pl. More typically, syncretic forms derive from phonetic change rather than morphological change, and leveling often involves the elimination of allomorphic forms in the radical (Hock 1991; Fertig 2013).

7. Concluding Remarks

This paper presented a case of morphological change in Portuguese that has received little attention in the literature. The documents studied, private letters written by very illiterate authors, present a multiplicity of phenomena associated with the innovations under discussion as the type of '*tu fostes*', emphasizing, on the one hand, the richness of the diachronic data that can be found in the particular epistolography, and on the other, demonstrating the scientific relevance of these innovations as an object of study, which are configured as "transgressors" of the norm (cf. Peres and Mória 1995). This paper attempted to contribute to the development of research in historical morphology. This was done by discussing the concept of analogy and its adequacy as an explanatory mechanism for change, looking at a phenomenon that has been scarcely studied. In addition, the research on this innovative morpheme (*-stes*) had also to focus on historical pragmatics, showing that a cooperation between historical morphology and historical pragmatics is possible and enriching. This study also attempted to show the relevance of studying private epistolary texts as rich linguistic testimonies of past languages diachronies, as they were written by people with poor education, making the writing of the letters very close to orality.

The data also support the hypothesis that this morphological change results from the analogical extension of the characterizing morpheme of the second person singular. This extension was favored by pragmatic and morphological conditions: the opacity of the morpheme *-ste*, the frequency and productivity of the person–number marking pattern of the second singular people, the pragmatic stability of *tu* for the informal context and the fall of *vós* and the verbal morphology of the second person plural created the necessary conditions for the occurrence of the diachronic analogical process. Innovations are most prominent from the 19th century onwards. In centuries prior to the 18th century, the coexistence of *tu* and *vós* for an interlocutor, within familiarity, possibly did not allow, for pragmatic reasons, the performance of analogy. At the time when *vós* was used in that discursive context, the distinction between second verbal persons through the morphemes *-ste* (2ps) and *-stes* (2pl) was pragmatically relevant.

The proportional equation, in the terms in which it is typically described, fails to adequately represent this analogical extension. The equation suggested here captures, in a more abstract way, the analogical model that served for analogical innovations such as '*fostes tu*'. Above all, it will illustrate more clearly the relationships between inflection patterns and the association that speakers perform mentally. An analogical change occurs when a significant number of speakers adopt the innovation (Fertig 2013); however, as previously indicated, there are several factors, namely, the negative perception of these forms and formal education, which may block the diffusion of these forms.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

2sg	second person singular
2pl	second person plural
3sg	third person singular
3pl	third person plural

Notes

- ¹ These definitions are the researcher's own translations. In addition to these definitions, the author also presents another one that he calls 'associative interference', defined as: 'The influence of one form on phonetics over another with which it is (or appears to be) grammatically or semantically related (Fertig 2013, p. 12)'. It does not appear in the list of definitions presented in the text because we believe it is not preponderant for the construction of the theoretical framework of this dissertation.
- ² Plural formation with -s of English names or regular past tense formation -ed (in modern English) are examples of more productive patterns. However, the unpredictable nature of the occurrence of the analogy is observed, for example, in several English names that were not regularized to the regular pattern (-s) in the plural, and remained as irregular with regard to the formation of the plural as is the case with *foot-feet* and *ox-oxen*. These words seem to resist regularization, with an exception being the production, especially by children, of the most frequent form of plural, when they produce, for example, * *foots*.
- ³ It should be noted that the investigation related to the pronoun *você* in Portuguese has been more productive, probably due to the multifaceted and complex character in socio-discursive terms.
- ⁴ These projects of collecting and digitalizing private letters were carried out by the Linguistics Centre of the University of Lisbon (CLUL). The archive are available in <http://teitok.clul.ul.pt/postscriptum/index.php> (accessed on 1 November 2023) and <http://fly.clul.ul.pt/index.php> (accessed on 2 November 2023).
- ⁵ More letters were located, but many were discarded in order to maintain a more homogeneous set (cf. Vaamonde 2018, pp. 147, 156).
- ⁶ *Casa da Suplicação* is, in general terms, equivalent to what is currently the Court of Appeal.
- ⁷ See Marquilhas (2015) for more details on the nature and size of these processes.
- ⁸ Available at <http://fly.clul.ul.pt/> (accessed on 2 November 2023).
- ⁹ The *Post-Scriptum corpus* has 993,274 words in Portuguese and *Fly has* about 700,000. The occurrence of 592 forms of 2SG and 2PL of the PPS is also justified by the textual genre in question and, probably, by the condition of the authors: they are short private letters, written in everyday life that do not generate very narrative or complex texts, and are written mostly by authors with low or no schooling.
- ¹⁰ Based on empirical/impressionistic evidence.
- ¹¹ In the imperative mode, 2sg affirmative is also marked with Ø.
- ¹² Not all changes will fit linearly into a proportional and harmonious equation, since linguistic change is governed different principles than those in mathematics: "Not all cases considered proportional analogy can be easily represented in this proportional formula, and some cases not normally thought to be proportional analogical change can be fitted into such a formula." (Campbell 1998, p. 91).
- ¹³ Based on research in the *Cordial-Sin corpus*, a corpus for dialectal syntax in Portuguese.
- ¹⁴ In a recent study on the behaviour of *vocês*/3pl and *vós*/2pl in Braga, a conservation area for *vós* and the 2pl for a plural interlocutor, Aguiar and Paiva (2017) observe that the explicit use of *vocês* prevails over that of overt *vós*, mainly in the subject function, however, and taking into account exclusively verbal uses, there is a prevalence of 2pl forms, for the same syntactic function. The authors also point out that the sum of the occurrences of explicit *vocês* with 3pl null subject forms indicates a growing use of the latter, pointing to an ongoing change in this area (Aguiar and Paiva 2017).

References

Primary Sources

- CLUL, ed. 2014. P.S. Post Scriptum. Arquivo Digital de Escrita Quotidiana em Portugal e Espanha na Época Moderna. Available online: <http://ps.clul.ul.pt> (accessed on 2 November 2023).
- Ed. Fly-Forgotten Letters, Years 1900–1974. Available online: <http://fly.clul.ul.pt/> (accessed on 2 November 2023).

Martins, A. M., coord. 2000. *CORDIAL-SIN: Corpus Dialectal para o Estudo da Sintaxe / Syntax-oriented Corpus of Portuguese Dialects*. Lisboa: Centro de Linguística da Universidade de Lisboa. Available online: <http://www.clul.ulisboa.pt/en/10-research/314-cordial-s> (accessed on 2 November 2023).

Secondary Sources

- Aguiar, Joana, and Maria da Conceição Paiva. 2017. Vocês tenham cuidado, sois educadas para isso- Second person pronouns in Braga speech. In *Studies on Variation and Change in Varieties of Portuguese*. Edited by Pilar Barbosa, Maria da Conceição de Paiva and Celeste Rodrigues. Philadelphia: John Benjamins, pp. 136–52.
- Albright, Adam. 2005. The morphological basis for paradigm leveling. In *Paradigms in Phonological Theory*. Edited by Laura Downing, Tracy Alan Hall and Renate Raffelsiefen. Cambridge: Oxford University Press.
- Bacelar, Fernanda do Nascimento. 2020. Formas de Tratamento. In *Gramática do Português*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota and Luísa Segura e Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, vol. III, pp. 2701–78.
- Bechara, Evanildo. 1991. As fases da língua portuguesa escrita. In *Actes du XVIIIe congrès international de linguistique et de philologie romanes*. Tübingen: Max Niemeyer Verlag, vol. III, pp. 68–76.
- Brocardo, M. Teresa. 2006. Sobre periodização da história do português europeu. Contributo para uma discussão. *Iberoromania* 62: 97–117. [CrossRef]
- Brocardo, M. Teresa. 2014. *Tópicos de História da Língua Portuguesa*. Lisboa: Colibri.
- Brocardo, M. Teresa, and Célia Lopes. 2016. History and Current Setting. In *The Handbook of Portuguese Linguistics*. Edited by W. Leo Wetzels, João Costa and Sergio Menuzzi. Philadelphia: John Wiley Sons, Inc., pp. 1–13.
- Bybee, Joan. 1985. *Morphology*. Amsterdam and Philadelphia: John Benjamins.
- Campbell, Lyle. 1998. *Historical Linguistics*. Edinburgh: Edinburgh University Press.
- Câmara Júnior, J. M. 1979. *Estrutura da Língua Portuguesa*. São Paulo: Vozes Editora.
- Cintra, Luís F. L. 1986. *Sobre “Formas de Tratamento” na Língua Portuguesa*, 2nd ed. Lisboa: Livros Horizonte.
- Cysouw, Michael. 2001. The Paradigmatic Structure of Person Marking. Tese de Doutorado, Universidade de Nijmegen, Nijmegen, The Netherlands.
- Faraco, Carlos A. 1996. O tratamento você em português: uma abordagem histórica. *Fragmenta* 13: 51–82. [CrossRef]
- Fertig, David. 2013. *Analogy and Morphological Change*. Edinburgh: Edinburgh University Press.
- Fisher, Olga. 2008. On analogy as a motivation for grammaticalization. In *Studies in Language*. Amsterdão: John Benjamins Publishings, vol. 32, pp. 336–82.
- Guilherme, Ana. 2021. ‘Fostes tu?’ A Marcação de Pessoa e Número em Português Europeu. O Caso da Segunda Pessoa do Pretérito Perfeito Simples. Doctoral Thesis, FCSH, Lisboa, Portugal.
- Hock, Hans. 1991. *Principles of Historical Linguistics*, 2nd ed. Berlin: Mouton de Gruyter.
- Hock, Hans. 2003. Analogical Change. In *The Handbook of Historical Linguistics*. Edited by Brian D. Joseph and Richard D. Janda. Oxford: Blackwell, pp. 441–60.
- Hock, Hans, and Brian D. Joseph. 1996. *Language History, Language Change and Language Relationships*. Berlin: Mouton de Gruyter.
- Jucker, Andreas H., and Irma Taavitsainen. 2002. *Diachronic Perspectives on Address Term Systems*. Philadelphia: John Benjamins.
- Kuryłowicz, Jerzy. 1995. The nature of the so-called analogical processes. *Acta Linguistica* 5: 17–34. First published 1945–1949.
- Lahiri, Aditi. 2000. *Analogy, Levelling, Markdness. Principles of Change in Phonology and Morphology*. New York and Berlin: Mouton de Gruyter.
- Lapesa, Rafael. 2000. *Estudios de Morfosintaxis Histórica del Español*. Madrid: Gredos.
- Lara Bermejo, Víctor. 2018. *La cortesía en la Península Ibérica: Dialectología del Sprachbund suroccidental*. Berna: Peter Lang.
- Lara Bermejo, Víctor, and Ana Guilherme. 2018. The Diachrony of Pronouns of Address in 20th Century European Portuguese. In *Studies in Hispanic and Lusophone Linguistics*. Berlin: De Gruyter, pp. 39–79.
- Leite de Vasconcellos, J. 1911. *Lições de Philologia Portuguesa*. Rio de Janeiro: Livros de Portugal.
- Maiden, Martin. 2011. Morphological Persistence. In *The Romance Languages*. Edited by M. Maiden, J. C. Smith and A. Ledgeway. Cambridge: Cambridge University Press, pp. 155–216.
- Mańczak, Witold. 1958. Tendences générales des changements analogiques. *Lingua* 7: 298–325. [CrossRef]
- Marcotulio, Leonardo L. 2014. Formas de tratamento no português arcaico: Contribuições do teatro português quinhentista. In *Confluência, Revista do Instituto de Língua Portuguesa*. Rio de Janeiro: e-Dialnet, pp. 157–200.
- Marquilhas, Rita. 2015. Non-anachronism in the historical sociolinguistic study of Portuguese. *Journal of Historical Sociolinguistics* 1: 213–42. [CrossRef]
- Mateus, Maria H. M. 1991. *Fonética, Fonologia e Morfologia do Português*. Lisboa: Universidade Aberta.
- McMahon, April. 1994. *Understanding Language Change*. Cambridge: Cambridge University Press.
- Mota, M. A. 2020. Morfologia do Verbo. In *Gramática do Português*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota and Luísa Segura e Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, vol. III, pp. 2933–3009.

- Paz Ramón, Mariño. 2013. O sufixo numero-persoal da P2 do pretérito de indicativo no galego medieval. In *Ao sabor do texto. Estudos dedicados a Ivo Castro*. Edited by Ana Maria Martins, Ana Ramos and Henriques Monteagudo. Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela, pp. 335–68.
- Peres, João Andrade, and Telmo Mória. 1995. *Áreas Críticas na Língua Portuguesa*. Lisboa: Caminho.
- Piel, Joseph-Maria. 1989. A flexão verbal do português (Estudos de morfologia histórica). In *Estudos de Linguística Histórica Galego-Portuguesa*. Lisboa: INCM. First published 1944.
- Rodrigues, Alexandra Soares. 2004. Da norma da língua sob a perspectiva da linguística. In *Conferências das comemorações dos 150 anos do Liceu de Bragança*. Bragança: Escola Secundária Emídio Garcia, pp. 65–91.
- Said Ali, Manuel. 2001. *Gramática Histórica da Língua Portuguesa*. São Paulo: Editora Melhoramento. First published 1921.
- Siewierska, Anna. 2004. *Person*. Cambridge: Cambridge University Press.
- Trask, Robert Lawrence. 1996. *Key Concepts in Language and Linguistics*. London: Psychology Press.
- Traugott, Elizabeth Closs. 2011. Grammaticalization and mechanisms of change. In *The Oxford Handbook of Grammaticalization*. Edited by Bernard Heine and Heiko Narog. Oxford: Oxford University Press.
- Vaamonde, Gael. 2018. Escritura epistolar, edición digital y anotación de corpus. *Cuadernos del Instituto Historia de la Lengua* 11: 139–64. [CrossRef]
- Villalva, Alina. 2003. Aspectos Morfológicos das Gramática do Português. In *Gramática da Língua Portuguesa*. Lisboa: Caminho, pp. 917–978.
- Villava, A., and Maria H. Mira Mateus. 2006. *O Essencial sobre Linguística*. Lisboa: Caminho.
- Wheeler, Max. 2012. Paths of analogy and explanation in the evolution of the weak preterite of the Romance e-conjugation. In *Etudis Romànics*. Edited by Martin Maiden. Barcelona: Institut d'Estudis Catalans, vol. 34, pp. 7–36.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Text Mining Approaches to Language Use in Social Media: The Case of Portuguese *Bué*

Camila Lívio ^{1,*} and Chad Howe ^{2,†}

¹ Department of Research and Computational Data Management, University of Georgia Libraries, Athens, GA 30602-1641, USA

² Department of Romance Languages, University of Georgia, Athens, GA 30602-1815, USA; chowe@uga.edu

* Correspondence: camila.emidio25@uga.edu

† These authors contributed equally to this work.

Abstract: This study describes processes of language change in Angolan Portuguese focusing on the use of intensifiers. Previous studies have shown that intensifiers are a relevant category for the study of language change due to their rapid change and variable meaning. It has been noted that intensifiers are particularly prone to renewal, suggesting speakers' desire to innovate. Informed by a Digital Humanities approach, we collect and analyze data from Twitter (now X), focusing on the multi-functional intensifier *bué*, 'very', in Angolan Portuguese (AP). In this paper, we (1) provide an overview of the word's distribution in AP, (2) consider the processes of change involved in *bué*'s variation, and (3) discuss the role of linguistic borrowing in language change and grammaticalization, shedding light on some of the cultural aspects that play a role in this word's development, such as the influence of the media and the contact situation between Angolan and European Portuguese.

Keywords: intensifiers; Angolan Portuguese; European Portuguese; social media; text mining; language variation and change

Citation: Lívio, Camila, and Chad Howe. 2024. Text Mining Approaches to Language Use in Social Media: The Case of Portuguese *Bué*. *Languages* 9: 82. <https://doi.org/10.3390/languages9030082>

Academic Editor: Patrícia Amaral

Received: 9 November 2023

Revised: 26 January 2024

Accepted: 30 January 2024

Published: 29 February 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Linguistic innovation and intensifiers are concepts that often appear together. The relationship between them can be described through the Maxim of Extravagance proposed by Haspelmath (1999, p. 1055): “talk in such a way that you are noticed”. Intensifiers, also known as degree modifiers, amplifiers, maximizers, and boosters (Tagliamonte 2008, p. 361), are argued to be a highly fluctuating class (Peters 1994, p. 271). This fluctuation is the result of the competition among intensifying forms, which change and are recycled in speech with renewed meaning (Ito and Tagliamonte 2003; Macaulay 2006; Tagliamonte 2008; Kanwit et al. 2017). For example, Tagliamonte observes that the low rate of *very* in Toronto confirms the interpretation of the preference for *really* over *very* in the twentieth century (Tagliamonte 2008, p. 369). Such rapid renewal in the intensifying system leads to the argument that speakers seek expressive (or extravagant) ways to emphasize meaning, “since their impact is only as good as their novelty” (Tagliamonte 2008, p. 391). This leads us to reflect upon issues of agency in language. As far as intensifiers are concerned, the adoption of a newer form (e.g., *wicked*) in the face of more canonical ones (e.g., *very*) can be conceptualized in terms of accommodation (Backus and Spotti 2012, p. 190). In other words, speakers tend to—consciously or unconsciously—adapt to various social contexts and the impressions they wish to make on their interlocutors (Nguyen et al. 2016, p. 550).

In this paper, rather than discussing a novel intensifier, we analyze the distribution, usage, and collocational profile of the well-established intensifier *bué*, 'very', in Angolan Portuguese. The origins of the word are not completely clear; several sources indicate that it comes from Kimbundu, a Bantu language widely spoken in Angola (1,700,160 speakers) and with an ethnic population of 6,000,000 people (Campbell 2008). For instance, in the online Portuguese dictionary *Priberam* (Figueira et al. 2011), the word appears as a synonym

of *muíto* in the informal Portuguese spoken in Portugal. The authors of the dictionary entry affirm that the Kimbundu origin is, however, uncertain. Pimenta (2022, p. 64) affirms that the use of this word was, for the generations born from the mid-1970s onwards, associated with migrating Angolans living in Portugal or with Angolan roots. Pimenta (2022) also refers to them as the *Bué Generation*, a generation with an “urgency for existence” (Pimenta 2022, pp. 64–65), and suggests that this generation was responsible for adopting the word from Kimbundu, *mbuwe*, into Portuguese, adding that the word’s consolidation as a linguistic mark of identity of Lisbon’s suburban culture stems from a song by Portuguese rapper of Cape Verdean roots, Boss AC (Pimenta 2022, p. 64). In addition, Almeida (2008, p. 19) mentions the frequent use of the word as an intensifier in the teen Portuguese soap opera *Morangos com Açúcar*, which aired between 2003 and 2012. Almeida (2008) explains that her choice to investigate the soap opera as a corpus for her study stems from the fact that such creative work featured ‘youngspeech,’ given its audience. Hence, the frequent use of *bué* by this age group is considered evidence of the word’s adoption into European Portuguese. A similar kind of media influence on language use was documented by Tagliamonte and Roberts (2005), in which the authors argue that the sitcom *Friends* has been “an influential cultural phenomenon” that “provides a kind of preview of mainstream language” (Tagliamonte and Roberts 2005, p. 296). The authors affirm that these types of media seem to “pave the way”, as language tends to be more innovative in the media than in the general population (Tagliamonte and Roberts 2005, p. 296). While the influence of television and other media has often been overlooked as a relevant factor in language change within the field of sociolinguistics (Stuart-Smith 2007, p. 140), we consider that media, such as television and music, serves as an indirect catalyst for linguistic diffusion. As Stuart-Smith (2007) puts it, “television may act as a source for new lexis and idioms, or as a model for speakers of a dialect to acquire the core phonology and syntax of the standard variety of a language (. . .), but here such changes require conscious motivation by speakers to orientate towards, and imitate, such a model” (Stuart-Smith 2007, p. 140). Trudgill (1989, p. 228) affirms that the degree of contact among speakers of different varieties of a language influences the rate of change. Therefore, considering the context of contact between Angolans and Portuguese is meaningful for a more comprehensive understanding of *bué*’s development.

During the Portuguese expansion, from the sixteenth to the twentieth century, speakers came into contact with a variety of different communities and languages. Specifically in the case of Angola, Carvalho and Lucchesi (2016, p. 44) affirm that the dissemination of Portuguese was limited even when the Portuguese ruled Luanda, in the late fifteenth century. The language was mostly used by the African–Portuguese administrative elite as a second language until the mid-eighteenth century (Carvalho and Lucchesi 2016, pp. 44–45). This linguistic situation changed in the twentieth century when a new policy requiring that Angolans were fluent in Portuguese was implemented during the Salazar dictatorship (1928–1974), and even then, the proportion of Portuguese speakers in Angola amounted to only 1% by 1940, dropping again by the time it reached political independence in 1975 (Carvalho and Lucchesi 2016, p. 44). Carvalho and Lucchesi (2016) add that in the present day, only a small portion of the Angolan elite speaks Portuguese as their mother language, in a variety that is similar to European Portuguese, as more than 90 percent of the population is a native speaker of a Bantu language, such as Kimbundo, Kikongo, and Umbundo (Carvalho and Lucchesi 2016, p. 45). A wide influx of immigrants from Angola to Portugal took place in the 1970s. An estimated 500,000 to 800,000 Portuguese settlers known as *retornados*—a term used to describe the Portuguese who were born in the African colonies—were forced to return to Portugal during the decolonization (Peralta 2022, p. 54). This exchange between Portugal and Angola is not unidirectional. With the end of the Civil War, many Angolans who used to live in Portugal returned to Angola, as well as new generations of Angolans moved to the former colony, which resulted in a steady growth of this population, according to the Portuguese National Institute of Statistics (INE 2022). Oliveira (2005) affirms that the relations between Angola and Portugal have been marked

by the dynamics of hierarchy and silencing since the fifteenth century, which resulted in great economic, social, cultural, and linguistic changes in Angola and, to a lesser degree, Portugal (Oliveira 2005, pp. 56–57).

This intense contact has, not surprisingly, led to a number of linguistic changes. Carvalho and Lucchesi (2016, p. 42) maintain that in most cases, the effects of language contact are observed through lexical borrowings that do not directly affect the languages involved. This claim seems to apply to our case study since the use of the word *bué* in Portugal tells us a small part of the long-standing relationship between these countries.

Our synchronic analysis utilizes a corpus of geo-tagged tweets from Luanda, Angola, and uses a computational approach to text analysis. Considering the nature of our dataset, we focus on language-internal factors that contribute to the variation and change of *bué*. We find this intensifier to be a compelling form for quantitative analysis of language variation and change due to its (i) development and variable use and (ii) frequency of use. Furthermore, it provides evidence for arguments that have been made in the study of intensifiers in English, such as the notion that intensifiers are considered to be (1) constantly changing (Ito and Tagliamonte 2003; Méndez-Naya 2003), (2) age-related (Roels and Enghels 2020), and, as we contend, (3) prone to being borrowed to/from other languages. Examples (1), (2), and (3) were taken from *Corpus do Português: Web/Dialects* (Davies and Ferreira 2006) in the Angolan subcorpus.

- (1) *Meus amigos, não adianta só filosofar bué*
'My friends, there's no point in philosophizing a lot'
- (2) *Foi uma cena bué positiva pois tive a oportunidade de mostrar outros trabalhos meus...*
'It was a very positive experience because I had the opportunity to show some of my other work'
- (3) *bué de peeps no mambo, boa música e BAR ABERTO!!! Daamm*
'Lots of folks dancing, good music and OPEN BAR!!! Dammm'

We start by outlining an overview of the literature on intensifiers in general and on Portuguese specifically, followed by a description of our materials and methods. We then show the general trends of intensification in the Twitter corpus, which are relevant to the introduction of our research questions in Section 6. The results are divided into three parts: (1) the general distribution of *bué* in the corpus, (2) a trigram-based approach for the analysis of the right and left collocates of *bué*, and (3) a statistical analysis of association strength. We then provide a discussion of the results and a conclusion in Section 7.

2. An Overview of Intensifiers

Speakers use a wide range of strategies to ensure that they, in Haspelmath's words, are noticed. These include types of linguistic structures that intensify, or "boost", the meaning of, among other categories, adjectival modifiers (Tagliamonte 2008)¹. For instance, the example in (4) compares the modifiers *alto/tall* and *muito alto/very tall*. The effect of the lexical intensifiers *muito* and *very* is such that the hearer is expected to understand that Cris is taller than Carmo. This interpretation holds regardless of whether the propositions are presented in parallel, as in (4a), or contrasted, as in (4b). The use of items like *muito* and "very" as adjectival modifiers will constitute the focus of the current analysis, although, as will be presented later, these effects are obtained in other domains as well. Among the other structures used with an intensifier meaning is the morphological superlative *-íssimo*, illustrated in (4c). Although this structure can be used with a meaning akin to that of *muito*, we will set aside a more in-depth discussion of how superlatives might be integrated into the system of degree modification in Portuguese.

- (4) a. *Carmo é alto e Cris é muito alto.*
'Carmo is tall and Cris is very tall.'
- b. *Carmo é alto mas Cris é muito alto.*
'Carmo is tall but Cris is very tall.'
- c. *Carmo é alto mas Cris é altíssimo.*
'Carmo is tall but Cris is very tall/the tallest.'

An exhaustive discussion of the precise semantic and pragmatic details surrounding the use of lexical intensifiers will not be provided in this discussion, though it should be noted that, at least for the purposes of explaining semantic change, it is precisely the comparison between, on the one hand, unmodified adjectives and those whose meanings are “boosted” (as in 4), and, on the other, the interplay between different lexical items drafted into service by speakers in this domain. The variety of lexical intensifiers in English has been documented in numerous sources (Bolinger 1972; Ito and Tagliamonte 2003; Macaulay 2006; Tagliamonte 2008, among others), with two primary observations that stand out in the literature. First, the individual entries in this set—such as *really*, *very*, *pretty*, *wicked*, etc.—emerge in sequence diachronically, with older variants existing alongside or perhaps being replaced with newer ones. This means that, for any given point in synchrony, speakers have at their disposal a palette of intensifiers that can be used for pragmatic purposes (i.e., “boosting” or “maximizing”). The second, and related, observation specifically concerns this notion of replacement, or “recycling”, as proposed by Ito and Tagliamonte (2003) and Tagliamonte (2008). The layering of options available to speakers produces competition among these structures, with some becoming more canonical, and consequently less “extravagant” (e.g., English *very*, than other, often newer options, like English *so* or *wicked*).

In Portuguese, Lívio and Howe (2020) argue, following Foltran and Nóbrega (2016), that there are two canonical intensifiers, *muito* and *bem*, exemplified in (5a) and (5b), respectively. These structures, discussed also by Gomes (2011), produce meanings analogous to their English counterparts, providing speakers with a means of enhancing the degree of the modified adjective. Foltran and Nóbrega (2016) also argue that *muito* is a “prototypical intensifier” in that there are numerous reflexes of this etymon in other Romance Languages².

- (5) a. *Portanto, é muito importante que o usuário tenha disponível cópias de segurança recentes de seus dados.*
 ‘Therefore, it is **very important** that the user has secure recent copies of their data.’
 b. *E isto ensina-nos a terceira lição bem importante: a missão não é nossa.*
 ‘And this is the third **really important** lesson we are taught: this is our mission.’

In their cross-dialectal distribution of *muito* and *bem*, Lívio and Howe (2020) show (i) that *muito* is preferred, at least in the context of adjectival modification, in each of the target varieties (i.e., Angolan, Brazilian, European, and Mozambican) and (ii) that there is some fluctuation with respect to this preference ranging from, on the high end, 87% *muito* in both the Angolan and Mozambican samples to only 70% in Brazil. These differences in overall frequency are accompanied by collocational distinctions as well, where *muito* co-occurs with a more restricted set of adjectives across the dialect samples (e.g., *simples* ‘easy/simple’, *bom* ‘good’, *difícil* ‘difficult’) and *bem* displaying a more heterogeneous profile.

Beyond these structures, Foltran and Nóbrega (2016) provide two additional groupings, “innovative adjectival intensifiers” (as in 6) and “borrowed intensifiers” (shown in 7 and 8). The use of *extremamente* in (6) reflects the typical trend of using adverbs in degree modification, a pattern attested widely across Portuguese and other Romance Languages. Examples (7) and (8), we argue, represent a distinct pattern concerning the development of these forms, suggesting an avenue for “renewal”, following Tagliamonte’s terminology, that draws on items from other languages. This is precisely the mechanism that is being proposed with *bué* in the current analysis.

- (6) *É um processo natural de floclulação de baixo custo e extremamente importante para as regiões secas*
 ‘It’s a low-cost, natural flocculation process and **extremely important** for dry regions’
 (7) *Em o final, fiquei super feliz porque não contava mesmo com aquela classificação*
 ‘In the end, I was **super happy** because I really didn’t count on that classification’
 (8) *Eu tinha uma limitação muito grande na alimentação, nada de frituras, carne apenas de rã, uma alimentação hiper controlada.*
 ‘I had a very big limitation in my diet, no fried foods, only frog meat, a **very controlled** diet.’

It perhaps goes without saying that a discussion of lexical borrowing of the type illustrated by the use of *super*, *hiper*, and *bué* in Portuguese necessitates an exploration of the cultural factors that influence their introduction. Nevertheless, we set aside these issues for now to focus on the structural embedding of these items, specifically *bué*, offering a detailed view of how it has been integrated into the system of degree modification. We will demonstrate that, in addition to its more canonical behavior as an adjective intensifier similar to both *muito* and *bem*, *bué* is distinct in a number of ways, including its usage in modification beyond the adjectival domain.

3. Materials and Methods

Recent studies have shown that Twitter data offer compelling linguistic representation in terms of regional tracking and diversity of the sample (Huang et al. 2016), as well as consisting of a useful source for the study of language variation and innovation (Grieve et al. 2018). Such an approach to linguistic analysis has been increasingly adopted as access to large corpora and texts become more available. Some emblematic examples include the works by Tagliamonte and colleagues on the investigation of intensifiers using corpora (Ito and Tagliamonte 2003; Tagliamonte 2008; Tagliamonte and Roberts 2005), Grieve and colleagues on the use Twitter as a prolific source of language data (Grieve et al. 2018, 2019), and Eisenstein and colleagues on social media language and its contribution and limitations to processing and understanding online natural language (Eisenstein 2013; Bamman et al. 2014). What these works have in common is that they center on different aspects of English and its variants. In Romance Languages, we observe a growing body of corpus-based studies that seek to not only apply similar methods to those used for the study of English in varying datasets but also deals with language-specific questions that go beyond English-oriented methods³.

For the present study, we compiled a corpus of 10,000 geo-tagged tweets from Luanda, Angola, in September of 2021, using the scripting and software environment R (R Core Team 2013) and the package *Rtweet* (Kearney et al. 2016). Retweets were not included in the corpus as a way to prevent gathering repeated tweets and bots. Furthermore, tweets are mixed, specified by the argument “type”. The search string targeted a list of five high-frequency adjectives (*bom* | *boa* ‘good’, *feliz* ‘happy’, *fixe* ‘cool’, *triste* ‘sad’, *especial* ‘special’) determined by a search on the Angolan Portuguese portion of the Reference Corpus of Contemporary Portuguese (do Nascimento et al. 2014). Thus, every tweet contains at least one of these adjectives specified in the search. The reasoning behind this methodological decision is that intensifiers are more frequently found modifying adjectives in comparison to other parts of speech (Ito and Tagliamonte 2003, p. 263). Further processing of the data included checking for its quality (i.e., number of empty values, duplicates, and consistency), cleaning (i.e., stripping the punctuation, case conversion, and removal of stop words), tokenization, and tagging. The tagging was performed using the R package *Udpipe* (Wijffels et al. 2018), due to its high performance and simplicity (Schweinberger 2023b). The package offers a great number of pre-trained language models, including Portuguese.

To carry out the analysis, we first provide overall trends in adjectival intensification in the Tweet corpus, followed by a manual search focusing on adjectives. We then focus on the structural properties of *bué* with the extraction and analysis of trigrams, as well as measuring the association strength between *bué* and other words. These methods allow for the observation of *bué*’s collocational profile, offering insights into the words’ meaning and variation.

The extraction of linguistic samples from digital media sites has become a popular method allowing for the gathering of larger samples that do not involve any type of elicitation from participants, which is desirable in the sense that it enables “unobtrusive measurements” (Nguyen et al. 2020, p. 4). However, it is essential to point out some of the limitations that born-digital data entail. For example, Twitter data are not as widely available in the case of Angolan Portuguese in comparison with many varieties of English. The website *Statista* lists the top 20 countries whose users are more engaged in the mi-

croblogging platform, showing that the only Portuguese-speaking country that displays a high number of Twitter users currently is Brazil, with 24.03 million users (Statista 2023). For that reason, our sample was limited to 10,000 tweets. In addition, we agree with Morin and Grieve (2024) that a disadvantage of working with social media data is related to the generalization of the results. Twitter data can only provide insights into language use on Twitter and, consequently, about users from specific demographics that are more represented in this social media platform (Morin and Grieve 2024, p. 11).

Data Management Plan

Considering the principles of open data in linguistics, as discussed in (Berez-Kroeker 2022), we adhere to Mattern’s guidelines (Mattern 2022, p. 19) on how to manage and preserve our dataset. While we strongly believe that a move toward open research (Gawne et al. 2021, p. 20) is imperative in our field to promote replicability and reproducibility, current regulations in the Twitter (now X) Developer Agreement prohibit full content redistribution, thereby affecting our ability to make the dataset entirely available. With these restrictions fully in mind, our data management plan, which outlines the data collection, software packages, dates, data structure, file types, metadata, variables, and analysis, can be found in this GitHub repository (commit efaf61c), <https://github.com/camlivio/Bueh-DMP> (accessed on 8 November 2023).

4. General Trends of Intensification in Tweets from Luanda, Angola

Initial mining of the corpus reveals the widespread use of intensifiers, such as *muito* (‘very’) and *bem* ‘well’, and comparative intensifying constructions, such as *mais ADJ que* (‘more ADJ than’) and *tão ADJ quanto* (‘as ADJ as’), particularly *muito*. These findings align with the literature on intensification in modern Spanish and Portuguese, which highlights such trends (Kanwit et al. 2017; Foltran and Nóbrega 2016). Figure 1 illustrates such a trend by means of a rapid automatic keyword extraction (RAKE). Rose et al. (2010, pp. 3–4) explain that within the field of information retrieval, keywords are defined as a sequence of one or more words that represent a text’s content. RAKE selects candidate keywords by focusing on meaning-bearing words and how they co-occur in the text. A score is then assigned to each candidate keyword and “defined as the sum of its member words scores” (Rose et al. 2010, p. 7). In other words, this method identifies keywords based on the frequency and the position of a word or set of words in a specific text⁴. Figure 1 shows the top 10 most frequent strings of adverbs preceding adjectives in the dataset, indicating that, as far as this syntactic combination goes, forms like *muito*, *tão*, and *bem* tend to be frequent before adjectives.

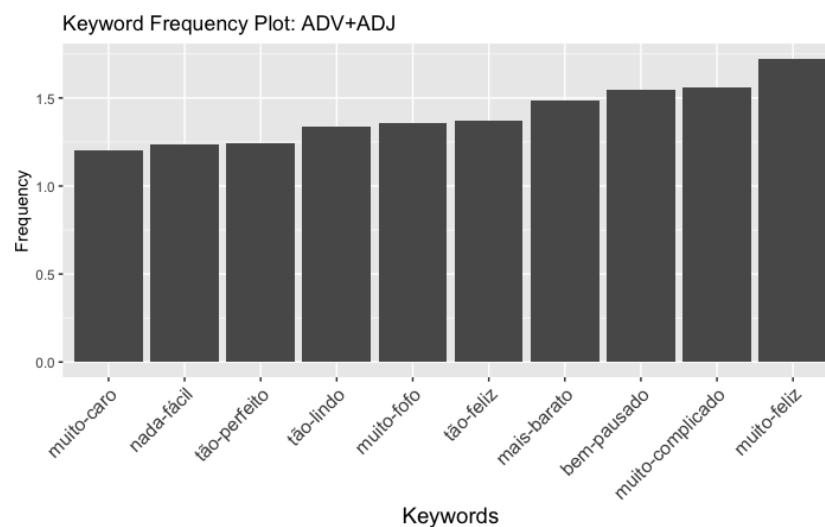


Figure 1. Rapid automatic keyword extraction of the sequence ADV + ADJ.

Lívio and Howe (2020, p. 480) show that the distribution of *muíto* and *bem* preceding adjectives in two African Portuguese varieties, Angolan and Mozambican, is not parallel to that of Brazilian and European Portuguese, whose cases of *bem* + ADJ are significantly higher, especially in the Brazilian variety. Considering that *bem* is newer in comparison with *muíto*—and given that the distribution of the latter is more homogeneous across varieties—we suggest that *bué* corresponds to *bem* in the Angolan variety. This can be better understood with a manual search in our Twitter corpus by adjective, as frequency shows us that *muíto* and comparative forms *mais/que* and *tão/que* tend to be the most used intensifying forms before adjectives in the Angolan variety, as shown in Figure 1. We queried the corpus of Tweets to observe how and which morphemes are used in this morphosyntactic environment (i.e., adjectives modified by intensifiers), with the objective of exploring varying intensifiers regardless of their frequency. Table 1 features both modified high-frequency adjectives (*fixe* ‘cool’, *bom* ‘good’, *feliz* ‘happy’, *triste* ‘sad’) and adjectives that are highly affective but lower in frequency (*puto* ‘shit’, *cool*, *foda* ‘screwed’).

Table 1. Examples of intensified adjectives extracted from the Twitter corpus.

Adjective	Example	Gloss
Fixe ‘cool’	<i>tem sempre uma irmã bweeee fixe</i>	‘There’s always a super cool sister’
Bom ‘good’	<i>kkkk mto bom me sentir excluída to amando</i>	‘LOL so good to feel excluded I’m loving it’
Feliz ‘happy’	<i>Aí tô mó feliz, comprei uma blusa</i>	‘Check it out I’m so happy, I got this new shirt’
Triste ‘sad’	<i>estou tao triste mano queria bue ler aquela revista</i>	‘I’m so sad bro I really wanted to read that magazine’
Puto ‘bitch’	<i>Cê tá onde, mó puto?</i>	‘Where are you/How are you feeling, so fed up?’
Cool	<i>O som está muíto cool</i>	‘The music is really cool’
Foda ‘shit’	<i>KKKKKKKK EU SOU MUITO FODA</i>	‘LOL I’m the shit’

The concordances⁵ in Table 1 show that intensifiers *bué*, also spelled *bue* and *bwé*, and *mó*—a written representation of the phonologically reduced *maior* (‘bigger’) (Xatara and Seco 2014, p. 511)—are the forms that appear to have a more informal use, similar to *bem*. We choose to focus on *bué* because it is the lexical intensifier with more language-specific characteristics and shows an interesting history and usage, illustrating why intensifiers have been considered to be (1) constantly changing (Ito and Tagliamonte 2003; Méndez-Naya 2003), (2) age-related (Roels and Enghels 2020), and (3) prone to being borrowed to and from other languages and communities.

As far as the language-specific argument is concerned, a quick search on *Corpus do Português* (Davies and Ferreira 2006) shows that *bué* is indeed widely employed in both Angolan and European Portuguese, as shown in Table 2. In this figure, both spellings are considered, *bué* and *bue*.

Table 2. Distribution of the intensifier *bué* | *bue* across varieties on *Corpus do Português Research Questions*.

Word	All	Brazil	Portugal	Angola	Mozambique
Bué	1367	39	942	330	56
Bue	577	123	228	212	14

Considering the background provided in the previous sections, we are well-positioned to answer the following research questions:

RQ1: How is *bué* used as an intensifier in the dataset?

RQ2: Given the general distribution of the word in Angola, Brazil, Mozambique, and Portugal, how has *bué* become widespread among Portuguese speakers?

RQ3: What does *bué*’s collocational profile tell us about the path of change of intensifiers?

5. Results

5.1. Overall Distribution of Bué

To visualize how *bué* is used in the corpus, we compare it to the distribution of other lexical intensifiers—as opposed to intensifying affixes—such as *muito*, and alternative forms that are frequent in other variants of the language, like the use of *super* and *totalmente* in Brazilian Portuguese. We also show how the standard spelling of the word compares with alternative forms, such as *bue*, *bwe*, and *bued*. Tatman (2015, p. 97) argues that while it may seem counter-intuitive to look for phonetic information in written mediums such as Twitter, previous research has shown that “the parallels between face-to-face and computer mediated communication (CMC) are robust” and “use a high proportion of variant spellings” (Tatman 2015, p. 99). Spelling variation is relevant in that it reflects, at least to some extent, the speaker’s own speech habits (Tatman 2015, pp. 99–100), and has been documented as a relevant way to observe social membership in some contexts (Androutsopoulos 2000; Thurlow and Brown 2003). We find that the word’s spelling variation generates further data to observe *bué*’s various grammatical functions, particularly between intensifying and quantification constructions (e.g., *bué de* ‘a lot of’). We provide a more detailed discussion about the many functions of the word in Section 6. Figure 2 illustrates these comparisons. In the first graph, only the form *bué* with the orthographic accent is considered.

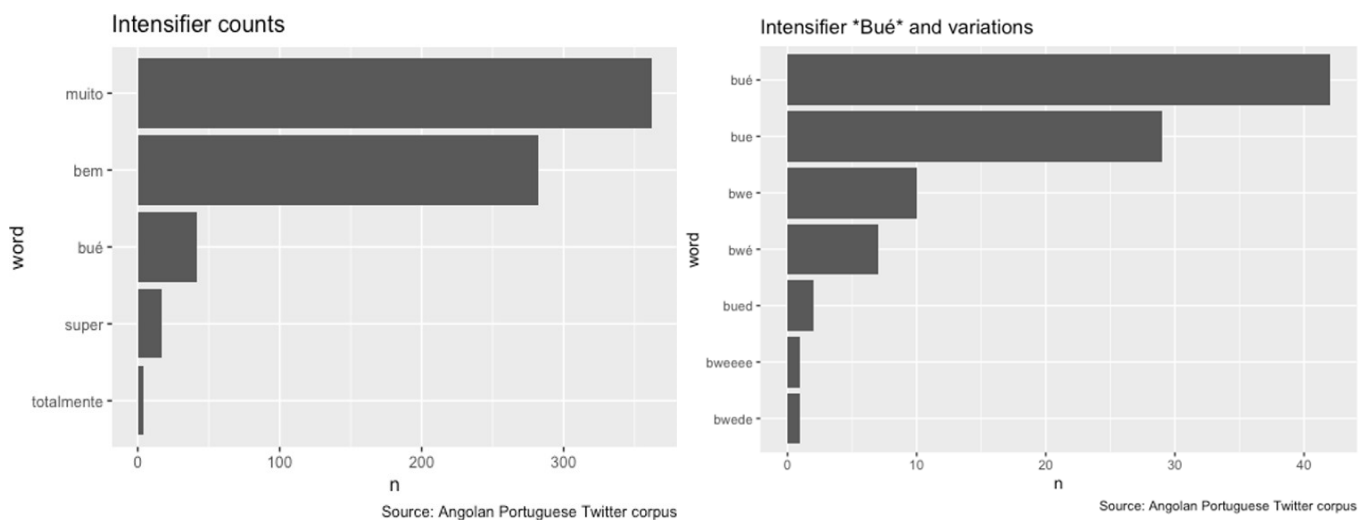


Figure 2. Comparing *Bué* to other lexical forms and variations in the same form.

It is interesting to notice that the overall counts of *bué*—even when considering only the normative spelling of the word—are higher than *super* since recent studies have shown that *super* has become widespread in many Portuguese varieties (Romerito Silva 2019; Lívio Emídio 2023), while significantly lower than both *muito* and *bem*. As far as the second graph in Figure 2 is concerned, we observe that both forms, with and without the accent, are the predominant choice among speakers in the sample. *Bued* and *bwede* point to a quantifying function of the word. Finally, Figure 3 shows the distribution of *bué* | *bue* by individual users. In this plot, users’ handles have been anonymized and substituted for an identification number on the x-axis. The figure shows that a few users tend to use the word at higher rates than most users. The graph suggests that some are “super users” of the word, while most users tend to use it once⁶.

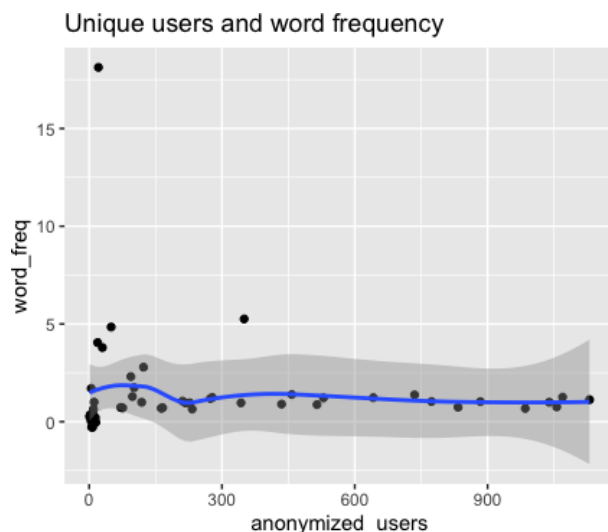


Figure 3. Accounting for individual users.

5.2. Trigrams

In this section, we propose a corpus-based approach to understand the patterning and variability in expressions containing the string *bué*. Through the extraction of trigrams, defined as three-word sequences (Jurafsky and Martin 2023, p. 2), we observe the terms that most frequently appear before and after *bué*. For the present analysis, we first tokenized the Twitter corpus into trigrams, as illustrated below in Table 3, in which we observe the first ten lines of the output where word2 corresponds to *bué*⁷.

Table 3. The first ten lines of trigrams containing *bué* in the word2 slot.

id	word1	word2	word3	Gloss
1	carne	bue	excited	‘meat very excited’
2	cena	bue	caricata	‘very ludicrous scene’
3	cena	bue	normal	‘very normal scene’
4	concordo	bue	mas	‘I totally agree but’
5	estao	bue	boas	‘they are so good’
6	estao	bue	nitidas	‘they are super clear’
7	fica	bue	bom	‘it’s really good’
8	foi	bue	rough	‘it was really rough’
9	foi	bue	um	‘it was really one’
10	guardam	bue	de	‘they keep lots of’

Essentially, this step results in a data frame in which each line contains a trigram. We then separate these trigrams so that each word is in its own, separate, column. With the inspection and extraction of all trigrams containing the word of interest, we created a new dataset in which *bué* occupies the second slot in the trigram, identified as *word2*. This arrangement facilitates the analysis of the words that precede and follow our key term. An illustration of a trigram featuring *bué* in the second slot is provided in Table 4.

Table 4. Representation of trigram extraction.

Word1	Word2 (=bué)	Word3
<i>é</i>	<i>bué</i>	<i>fixe</i>
'It's	very	cool'

Jurafsky and Martin (2023, pp. 2–3) elaborate on the concept of n-grams, which rely on the conditional probability of a word *w* given its history *h*, formally denoted as $P(w|h)$. Essentially, the n-gram approach aids in addressing the question, “How frequently does the word *w* follow the history *h*?” (Jurafsky and Martin 2023, p. 2), thereby uncovering frequency trends in the co-occurrence of specific words. Our findings, illustrated in Figure 4, are crucial for the examination of the morphosyntactic environment of the term.

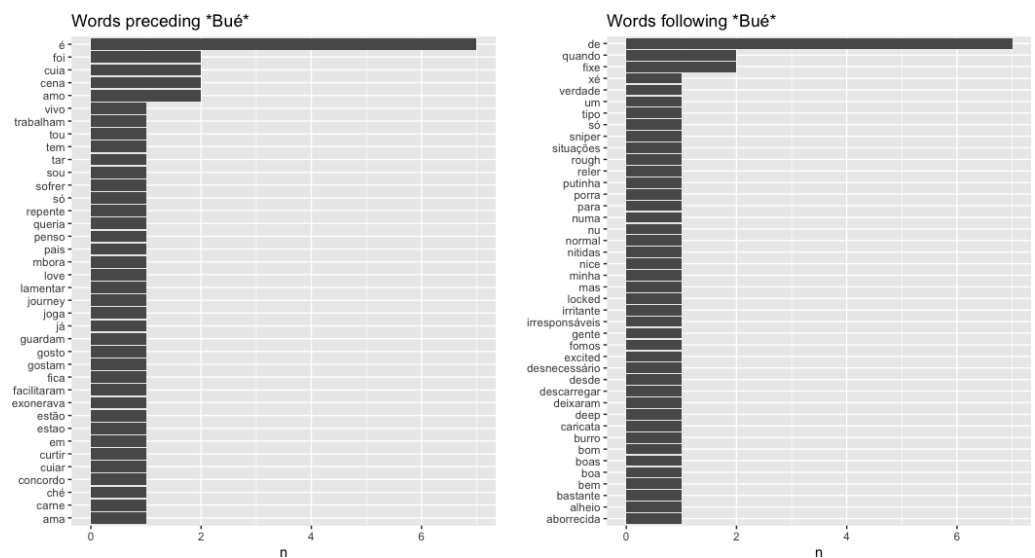


Figure 4. Words that precede and follow *bué*.

Starting with the first plot in Figure 4, the most frequent terms preceding *bué* is the singular form of the copular verb *é* ‘is’, followed by the less frequent preterit form of the same verb, *foi* ‘was’. These forms indicate that speakers typically use *bué* as a modifier of predicative adjectives, such as the example given in Table 4. Tagliamonte (2008, p. 373) points out that a higher frequency of intensifiers modifying predicative adjectives is a reflection of “a later point in an intensifier’s development”, showcasing a well-documented pattern of intensification variation, namely, deslexicalization (Ito and Tagliamonte 2003; Tagliamonte 2008; Schweinberger 2020). This argument is motivated by the observation of the development of *very*: “According to some researchers, a later stage in the development of intensifier *very* was when it came to modify predicative adjectives” (Tagliamonte 2008, p. 373).

5.3. Association Strength

Corpus-based approaches have played a pivotal role in describing linguistic patterns of variation and use (Biber 2015, p. 197). The well-known phrase “You shall know a word by the company it keeps”, attributed to Firth (1968, p. 179), remains pertinent in the context of our analysis. Specifically, we use an association measurement to shed light on *bué*’s collocates. We find this method to be useful in that the “use of computational tools ensures high reliability, since a computer program should make the same analytical decision every time it encounters the same linguistic phenomenon” (Biber 2015, p. 197). The trigram analysis allowed us to observe raw frequency counts of what terms occur as the right and

left collocates of *bué*. However, as Evert points out, plain frequencies are not enough “as a measure for the amount of ‘glue’ between two words” (Evert 2005, p. 20).

Our statistical interpretation of the frequency data is based on repeated Fisher’s exact tests and subsequent Benjamini–Hochberg correction for repeated testing, following Schweinberger’s (2023a) analysis of co-occurrences and collocations in R. The results indicate that the statistically significant collocations of *bué* in the Twitter data are *merdas* ‘shit’, *carne* ‘meat’, *tou* ‘I am’, and *morto* ‘dead’, as shown in the table below. The results in the column *CorrSignificance* represent the corrected p-values, which are necessary after performing several tests to avoid errors (Schweinberger 2023a). The corrected values display the only truly significant collocates of *bué* | *bue*. These findings hold practical value as they not only highlight the statistically significant collocates (*type*) but also the terms that are repelled (*antitype*) by the key term. The presence of the word *não*, as an antitype, shows that it co-occurs less frequently with the word *bué*.

Term	CoocTerm	TermCoocFreq	CoocFreq	p	X2	phi	expected	Significance	CorrSignificance	Type
bue	merdas	4	15	0.000193	30.4	0.24	0.345	p < 0.001	p < 0.01	type
bue	carne	4	22	0.000945	18.9	0.19	0.506	p < 0.001	p < 0.01	type
bue	tou	2	14	0.0378	4.5	0.09	0.322	p < 0.05	NA	type
bue	morto	2	14	0.0378	4.5	0.09	0.322	p < 0.05	NA	type
bue	nao	0	97	0.135	1.7	0.06	2.23	NA	NA	antitype
bue	ver	0	83	0.229	1.3	0.05	1.91	NA	NA	antitype
bue	pq	0	90	0.236	1.5	0.05	2.07	NA	NA	antitype
bue	n	0	62	0.377	0.7	0.04	1.43	NA	NA	antitype
bue	bue	0	41	0.612	0.2	0.02	0.943	NA	NA	antitype
bue	acho	0	35	1	0.1	0.02	0.805	NA	NA	antitype

6. Discussion

Breban and Davidse (2016, p. 245) and Méndez-Naya (2003, p. 389) argue that the various functions and layers of meaning of the intensifier *very* are observable due to its complete process of grammaticalization, which is the reason why *very* cannot be renewed (i.e., reintroduced in the speech with renewed expressive power as discussed in Tagliamonte (2008)). These authors, as well as several other studies on intensifiers (Ito and Tagliamonte 2003; Kanwit et al. 2017; Méndez-Naya 2003; Serradilla Castaño 2006), maintain that the high usage of an intensifier gives rise to alternative or competing forms in the language due to the speaker’s desire for innovation and expressivity. Similar to *very*, *bué* has undergone many changes that are observable in our Twitter corpus, noticeably by its generalization of meaning (Méndez-Naya 2003, p. 372). For example, Almeida (2008) argues that the change is from positive adjective *bué* in Angola, as in examples (9) and (10), to intensifier *bué* in Portugal, as in (11).

- (9) *A senhora é bué*
‘The lady is fashionable’
- (10) *Kota, o serviço foi bué, portanto vais dar só cem dólares*
‘Kota, the service was good, therefore you should tip only \$100’
- (11) *Alunos estrangeiros, fixe. Vocês vão curtir bué a Barra*
‘Foreign students, cool. You guys will enjoy a Barra a lot’

Adamson (2000, p. 47) describes a similar change in the adverb *lovely*. The author argues that *lovely* starts out as an adjective of positive connotation and comes to be used as an intensifier via a shift from denotational to a “speaker-related meaning”, or from descriptive to affective, as well as a change that entails moving from being an independent lexical word to a grammatical operator (Breban and Davidse 2016, p. 222). The author contrasts the change in meaning and structure in *She is lovely* to *A lovely quiet engine* (Adamson 2000, p. 54), which is parallel to the process described by Almeida (2008) (*Ela é bué* to *é bué interessante*). To put it another way, our n-gram analysis of *bué* showed that the keyword tends to be preceded by copular verbs (e.g., *é* and *foi*), as well as other verbal forms (*amo* ‘I love’, *vivo* ‘I live’, *trabalham* ‘they work’, *tou* ‘I am’, *tem* ‘have/has’, *tar* ‘are’, *sou* ‘I am’, and *sofrer* ‘to suffer’), indicating functional layering. This observation is further

supported when examining the right-hand collocates of *bué*, as the frequency in which the preposition *de* occurs points to the use of a quantifying construction. Examples (12) and (13) illustrate some instances of the construction in our Twitter corpus. Note that these examples display a “counting” interpretation (Doetjes 2008, p. 149) as the modifier targets the plural predicates.

- (12) *Bué de gajos*
'Lots of guys'
(13) *Bué de gatos*
'Lots of cats'

As a quantifier, *bué de* displays similar behavior to the quantifying construction *monte de*, which is literally translated as ‘a mountain of’. According to Alonso and Fumaux (2019, p. 141), the more recent quantifying interpretation of *um monte de* stems from a locative understanding of the construction. For instance, in example (14), from the 16th century, obtained by the authors from *Corpus do Português* (Davies and Ferreira 2006), they argue that the original construction is historically linked to a qualitative expression where *monte* (‘mountain’) refers to *morro* (‘summit’), and the prepositional phrase establishes a connection to a specific location where the mountain is situated, as shown in example (14), taken from the historical section of *Corpus do Português* (Davies and Ferreira 2006) from the 16th century, as cited by Alonso and Fumaux (2019, pp. 141–42).

- (14) *Guar-te de praguejar de homiês poderosos porque t oras hehû monte de africa. Onde foy enforcado daphitasgrãmatico porque dezia mal dos reys em verso*
'Beware of cursing powerful men because you are but a mountain in Africa. Where Daphitas the grammarian was hanged because he spoke ill of kings in verse.'

Moreover, they add that such use functions as a specifier in that it identifies where the mountain is located. A number of syntagmatic variations are found in Portuguese, particularly as far as constructions of the type NP_1 of NP_2 : *copo de vidro*, *caderno de atividades*, *monte de Lisboa*, as well as similar quantifying constructions, such as *milhares de pessoas* ‘millions of people’ and *centenas de problemas* ‘hundreds of problems’ (Alonso and Fumaux 2019, p. 122).

A further development of the quantifying construction is described by Almeida (2008, p. 122). The author states that *bué da* has evolved from *bué de* and was commonly used by young European Portuguese speakers to convey heightened expressiveness. This assertion is supported by examples (15) and (16) provided in her work (Almeida 2008, p. 118). By contrasting these examples with (12) and (13), we can observe that (15) and (16) produce an intensifying reading—rather than a counting one—despite the presence of the preposition *da*. This effect is attributed to the predicates they modify, as both *fixe* and *preocupado* are adjectives.

- (15) *Este jogo é bué da fixe eu ainda não joguei mas eu tenho o final fantasy x e é bué da fixe*
'This game is super cool, I still haven't played it but I have Final Fantasy X and it is super cool'
(16) *Acho que ficaste bué da preocupado*
'I think you were super worried'

In addition to *bué de*, the results for the right-hand collocates display a variety of adjectives that are modified by the intensifier: *fixe* ‘cool’, *nítidas* ‘clear’, *nice*, *irritante* ‘annoying’, *irresponsáveis* ‘irresponsible’, *excited*, *burro* ‘dumb’, and *bom | boa* ‘good’. This result 467 supports the claim that the more grammaticalized a form is, the wider the semantic evaluation of the words it modifies will be (Ito and Tagliamonte 2003; Kanwit et al. 2017). In other words, intensifiers that have been used by a community of speakers for a long time tend to modify adjectives of both positive and negative semantic evaluation, such as *bom* ‘good’ and *burro* ‘dumb’, as is the case with *muito* and, to a lesser extent, *bem* (*bem feio* ‘well ugly’). Given that *bué* is no longer considered an innovative intensifier in Portuguese, evidenced by its inclusion in the *Dicionário da Academia das Ciências* (Casteleiro 2001), it is not surprising that speakers use it with a variety of adjectives.

The statistical analysis confirms that *bué* has undergone grammaticalization, aligning with the pattern found in similar constructions, such as *very* and *lovely*, and evidenced by the functional layers that it displays in the corpus (intensifier, quantifier, and adjective). Observing both the right-hand collocates of *bué* and its statistically significant collocates, we add that the type of intensification introduced by this word seems to be more pragmatic than grammatical in Beltrama and Bochnak's (2015) terms. The authors posit two main ways in which degree modification can manifest. The first, called "true degree intensification", occurs when an intensifier tracks scales lexically encoded in the modified word, as seen in examples like *very tall*, where "very" modifies a scalable property of the adjective "tall". In other words, they modify gradable predicates. Conversely, another class of degree modifiers targets aspects of the meaning rather than gradable properties, as in the Italian example *Michael Jordan è un campionissimo* 'Jordan is a super champion', in which the intensifying suffix *-issimo* reinforces the meaning of the modified word, enhancing the expressiveness of the utterance (Beltrama and Bochnak 2015, p. 845). The collocates in Figure 4 show that *bué* modifies both gradable (*fixe* 'cool'), and non-gradable predicates (*nu* 'naked' and *morto* 'dead'), though the true collocates of the word are non-gradable, suggesting that the interpretation of such constructions are context-dependent, or at least more dependent on the context than grammatical intensification.

In addition to showing the words with which *bué* is more likely to co-occur, an advantage of the code elaborated by Schweinberger (2023a) to measure word association is that we can concomitantly observe the terms that are repelled by our key term, *bué*. Negation words such as *não*, and its abbreviated form, *n*, are less likely to occur in contexts where *bué* is present. We suggest that at least part of the explanation for this phenomenon is connected to the issue of polarity. Similar to *bem*, *bué* partially retains its original meaning (Luo et al. 2019, pp. 1–2), which is a positive polarity meaning in this instance, conflicting with what Israel (2006, p. 10) describes as the "unpleasant sort of construction" that negation introduces.

In summary, we pose that *bué*'s "foreign quality" makes for an ideal intensifier as these modifiers are known for being "colorful" and "inventive" (Bolinger 1972, p. 18), and are often sourced from lexical items (Ghesquière and Davidse 2011, p. 252) as they are meaning-bearing words and speakers target some aspect of their meaning. For example, the English adjectives *awful* and *dead* function as intensifiers in utterances such as *an awfully long line* and *dead chuffed* due to their extreme meanings.

7. Conclusions

In this paper, we focused on the use of the Angolan Portuguese intensifier *bué* in social media data. We showed that as a synonym for 'abundance', and having its origins being generally attributed to Kimbundo, this frequent word in Angolan Portuguese displays many and variable uses. We argued that it is due to the intense contact among speakers of Angolan and European Portuguese, combined with its use in popular culture, that the word is borrowed into European Portuguese. Furthermore, we posit that, as a borrowing, *bué* possesses a foreign quality that can also be described in terms of Haspelmath's (1999) Maxim of Extravagance, or the Level of Surprisal that an intensifier introduces in speech. According to Scheffler et al. (2023, p. 2), "high surprisal of a sign means that it is unexpected given its preceding context of occurrence, and that it carries high information". In that sense, we believe that certain loanwords are apt to become thriving intensifying forms in a given language due to the high levels of information that they carry, explaining the popularity of other similar intensifiers categorized as foreign, such as *super* and *hiper* in many varieties of Portuguese (Foltran and Nóbrega 2016). Documented in the Portuguese Dictionary of the Lisbon Academy of Sciences since 2001 (Almeida 2008) in three different entries (adjective, adverb, and interjection), the element of surprise or extravagance associated with *bué* has decreased over time as it became increasingly popular among speakers. Currently, such popularity is observed in online forums in which native and non-native speakers of different varieties of Portuguese ask and comment about the words' usage, meaning, and

origins, such as in an entry in the question-and-answer website Quora (Como surgiu o termo “bué” em português de Portugal n.d.) that displays over thirty replies from different perspectives about the word.

One interesting aspect of loanwords that is worth being noted is that it is not clear why speakers borrow a word from a different language when a “fully equivalent word existed beforehand” (Haspelmath 2009, p. 35). And even beyond that, Haspelmath inquires why borrowings occur at all as every language has the means to create new words and expressions. Moreover, while research on loanwords shows that borrowing may take place in both directions—to and from cultures in contact—it is argued that there tends to exist an asymmetry in this process, given that source languages often hold an advantage of power over recipient languages in some way (Hoffer 2002, p. 3). What we observe in the case of *bué* is the opposite, revealing that social aspects and attitudes play a role in borrowing (Poplack 2017, p. 186).

By focusing on the contexts in which the word appears in a collection of contemporary tweets, our analyses showed that this intensifier modifies a variety of word classes, such as adjectives, verbs, and nouns, of both positive and negative semantic evaluation, as well as possesses a quantifier function, suggesting that *bué* displays similar development to other intensifiers crosslinguistically, such as English *lovely* (Adamson 2000) and Spanish and Portuguese *bien/bem* (Kanwit et al. 2017; Livio and Howe 2020). It starts out as an adjective of positive evaluation, transitioning from a descriptive function to an affective one as speakers use these forms to “externalize their subjective point of view”, as argued by Traugott (1999, p. 189). This finding is relevant in the wider literature on intensification as it shows that similar processes of change, such as bleaching and grammaticalization (Ito and Tagliamonte 2003; Luo et al. 2019), are at play in the development of this grammatical class across languages.

In terms of our methods, we demonstrated the usefulness of manipulating social media data, as it provides access to a snapshot of synchrony. This approach allows us to observe how speakers utilize certain terms and constructions, fostering productive dialogues between fields such as language use, pragmatics, historical linguistics, text mining, digital humanities, and computational social science (Huang et al. 2016; Scheffler 2017).

Author Contributions: C.L. and C.H. contributed equally to this manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Twitter’s current Developer Agreement and Policy prevent us from making the full dataset publicly available.

Acknowledgments: The authors would like to express their gratitude to two anonymous reviewers and Patrícia Amaral for providing insightful comments throughout the process. The first author also extends thanks to Carlos Pio for her time and insights regarding the use of the word *bué* in Portugal, along with general comments about language use.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

¹ The terms “intensifier” and “booster” are often accompanied by a reference to “maximizers” as well, all with the intention of underscoring the fact that these structures play a role in enhancing the meaning of the modified element. These should not be confused with “minimizers”, as in *a word* in “He said not a word”, which produces a parallel effect, though in the opposite direction on the degree scale.

² As will be seen in subsequent examples, there are orthographic variants of the canonical intensifiers—e.g., *muito* > *mto* and *mó*. The form *mui* was attested in Old Portuguese (Huber 1986, paras. 423 and 424) and may be observed in contemporary varieties.

- ³ The publication of *The Routledge Handbook of Spanish Corpus Linguistics* (Parodi et al. 2021) responds to this trend and features recent works on different types of corpora, methodologies, and tools in the field of Spanish linguistics.
- ⁴ The code used for the extraction of RAKE returned some strings that were not pairs of ADV+ADJ, such as in *nem precisas, já devia*, and *nunca parar*. We manually filtered out those cases and plotted only the ones that matched the search.
- ⁵ Among the many features observed in the data provided in Table 1 are, primarily orthographic, strategies used by authors in digital contexts. For instance, in addition to the general use of all caps and repeated letters, laughter in Portuguese is represented specifically as *kkkk/KKKKKK*. Truncated forms such as *tô* (<estou 'I am') and *cê* (<voçe 'you') are also commonly attested in this register.
- ⁶ One of the reviewers points out that this graph may represent two distinct speech communities. We believe that observing the demographic information about these speakers could have been one way to identify such communities.
- ⁷ One of the reviewers points out the recurrence of non-Portuguese words, especially in the "word3" slot. We generally concur with Androutsopoulos (2004, p. 83) regarding the use of English words by non-native English speakers in social media. This phenomenon, named 'Englishization' by the author, is described as being motivated by identification with English-speaking pop culture. Furthermore, Androutsopoulos emphasizes the importance of studying Englishization in specific linguistic settings to understand the sociolinguistic factors that may influence such use.

References

- Adamson, Sylvia. 2000. A lovely little example. Word order options and category shift in the premodifying string. In *Pathways of Change*. Edited by Olga Fischer, Anette Rosenbach and Dieter Stein. Amsterdam and Philadelphia: John Benjamins, pp. 39–66.
- Almeida, Maria Clotilde. 2008. Youngspeak, Subjectification and Language Change: The case of *bué*. In *Questions on Language Change*. Edited by Maria Clotilde Almeida, Bernd Sieberg and Ana Maria Bernardo. Lisboa: Colibri, pp. 117–32.
- Alonso, Karen Sampaio Braga, and Nuciene Caroline Amphiphio Fumaux. 2019. Um monte de mudanças nessa construção: Sintaxe e semântica do quantificador 'um monte de' no português. *LaborHistórico* 5: 119–41. [CrossRef]
- Androutsopoulos, Jannis. 2004. Non-native English and sub-cultural identities in media discourse. In *Den fleirspråklege utfordringa/The Multilingual Challenge*. Edited by Helge Sandøy. Oslo: Novus, pp. 83–98.
- Androutsopoulos, Jannis K. 2000. Non-standard spellings in media texts: The case of German fanzines. *Journal of Sociolinguistics* 4: 514–33. [CrossRef]
- Backus, Ad, and Massimiliano Spotti. 2012. Normativity and change: Introduction to the special issue on agency and power in multilingual discourse. *Sociolinguistic Studies* 6: 185. [CrossRef]
- Bamman, David, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18: 135–60. [CrossRef]
- Beltrama, Andrea, and M. Ryan Bochnak. 2015. Intensification without degrees cross-linguistically. *Natural Language & Linguistic Theory* 33: 843–79.
- Berez-Kroeker, Andrea L. 2022. *The Open Handbook of Linguistic Data Management*. Cambridge: MIT Press.
- Biber, Douglas. 2015. Corpus-based and corpus-driven analyses of language variation and use. In *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press.
- Bolinger, Dwight. 1972. *Degree Words*. Berlin: De Gruyter.
- Breban, Tine, and Kristin Davidse. 2016. The history of *very*: The directionality of functional shift and (inter)subjectification. *English Language & Linguistics* 20: 221–49. [CrossRef]
- Campbell, Lyle. 2008. *Ethnologue: Languages of the World*. Dallas: SIL International.
- Carvalho, Ana Maria, and Dante Lucchesi. 2016. Portuguese in contact. In *The Handbook of Portuguese Linguistics*. Hoboken: Wiley-Blackwell, pp. 41–55.
- Casteleiro, João Malaca. 2001. *Dicionário da língua Portuguesa Contemporânea*. Lisboa: Academia das Ciências de Lisboa e Editorial Verbo, vol. 2.
- Como surgiu o termo "bué" em português de Portugal? n.d. Available online: <https://pt.quora.com/Como-surgiu-o-termo-bu%C3%A9-em-portugu%C3%AAs-de-Portugal> (accessed on 19 January 2022).
- Davies, Mark, and Michael Ferreira. 2006. *Corpus do Português*. Available online: <http://www.corpusdoportuguese.org> (accessed on 20 November 2023).
- do Nascimento, Maria Fernanda Bacelar, Amália Mendes, Sandra Antunes, and Luísa Pereira. 2014. The Reference Corpus of Contemporary Portuguese and related resources. In *Working with Portuguese Corpora*. London, New York, New Delhi, Oxford and Sydney: Bloomsbury Academic, p. 237.
- Doetjes, Jenny. 2008. Counting and degree modification. In *Recherches Linguistiques de Vincennes*. Saint-Denis: Presses Universitaires de Vincennes, pp. 139–61.
- Eisenstein, Jacob. 2013. What to do about bad language on the internet. Paper presented at the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, Georgia, June 9–14; Toronto: Association for Computational Linguistics, pp. 359–69.
- Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. Ph.D. dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart, Stuttgart, Germany.

- Figueira, Helena, Alfonso Mendes, Pedro Mendes, and Cláudia Pinto. 2011. Dicionário Priberam da Língua Portuguesa. *Boletim da Academia Galega da Língua Portuguesa*.
- Firth, John. 1968. A synopsis of linguistic theory, 1930–55. In *Studies in Linguistic Analysis*. Edited by Frank Robert Palmer. Oxford: Blackwell.
- Foltran, Maria José, and Vítor Augusto Nóbrega. 2016. Adjetivos intensificadores no Português Brasileiro: Propriedades, distribuição e reflexos morfológicos. *Alfa: Revista de Linguística (São José do Rio Preto)* 60: 319–40. [CrossRef]
- Gawne, Lauren, Suzy Styles, Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller, and Lauren B. Collister. 2021. Situating linguistics in the social science data movement. In *The Open Handbook of Linguistic Data Management*. Cambridge: MIT Press, pp. 9–25.
- Ghesquière, Lobke, and Kristin Davidse. 2011. The development of intensification scales in noun-intensifying uses of adjectives: Sources, paths and mechanisms of change. *English Language & Linguistics* 15: 251–77.
- Gomes, Ana Quadros. 2011. Uma proposta de distinção semântica para os intensificadores muito e bem. *Estudos Linguísticos* 41: 379–94.
- Grieve, Jack, Andrea Nini, and Diansheng Guo. 2018. Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics* 46: 293–319. [CrossRef]
- Grieve, Jack, Chris Montgomery, Andrea Nini, Akira Murakami, and Diansheng Guo. 2019. Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence* 2: 11. [CrossRef] [PubMed]
- Haspelmath, Martin. 1999. Why is grammaticalization irreversible? *Linguistics* 37: 1043–68. [CrossRef]
- Haspelmath, Martin. 2009. Lexical borrowing: Concepts and issues. In *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: De Gruyter, vol. 35, p. 54.
- Hoffer, Bates L. 2002. Language borrowing and language diffusion: An overview. *Intercultural Communication Studies* 11: 1–37.
- Huang, Yuan, Diansheng Guo, Alice Kasakoff, and Jack Grieve. 2016. Understanding U.S. regional linguistic variation with Twitter data analysis. *Computers, Environment and Urban Systems* 59: 244–55. [CrossRef]
- Huber, Joseph. 1986. *Gramática do português antigo*. Translated by Maria Manuela Gouveia Delille. Lisboa: Fundação Calouste Gulbenkian.
- INE. 2022. Censos 2021. In *O que nos dizem os Censos sobre a População de Nacionalidade Estrangeira Residente em Portugal*. Relatório Técnico. Lisboa: Instituto Nacional de Estatística. ISBN 978-989-25-0633-3.
- Israel, Michael. 2006. The pragmatics of polarity. In *The Handbook of Pragmatics*. Hoboken: Wiley-Blackwell, pp. 701–23.
- Ito, Rika, and Sali Tagliamonte. 2003. Well weird, right dodgy, very strange, really cool: Layering and recycling in English intensifiers. *Language in Society* 32: 257–79. [CrossRef]
- Jurafsky, Dan, and James H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Saddle River: Prentice Hall.
- Kanwit, M., V. Terán, and S. P. Sarrió. 2017. Un fenómeno bien curioso: New methods for analyzing variable intensification across four dialects of Spain and Argentina. *Studies in Hispanic and Lusophone Linguistics* 10: 259–95. [CrossRef]
- Kearney, Michael, Andrew Heiss, and François Briatte. 2016. Package 'Rtweet' Briatte. Available online: <https://cran.r-project.org/web/packages/rtweet/rtweet.pdf> (accessed on 19 March 2019).
- Lívio, Camila, and Chad Howe. 2020. Intensification in Portuguese: A Cross-dialectal Study of Muito and Bem. *Hispania* 103: 475–87. [CrossRef]
- Lívio Emídio, Camila. 2023. Digital Approaches to Intensification in Portuguese and Spanish. Ph.D. dissertation, University of Georgia, Athens, GA, USA.
- Luo, Yiwei, Dan Jurafsky, and Beth Levin. 2019. From Insanely Jealous to Insanely Delicious: Computational Models for the Semantic Bleaching of English Intensifiers. Paper presented at the 1st International Workshop on Computational Approaches to Historical Language Change, Florence, Italy, August 2; Florence: Association for Computational Linguistics, pp. 1–13. [CrossRef]
- Macaulay, Ronald. 2006. Pure grammaticalization: The development of a teenage intensifier. *Language Variation and Change* 18: 267–83. [CrossRef]
- Mattern, Eleanor. 2022. The linguistic data life cycle, sustainability of data, and principles of solid data management. In *The Open Handbook of Linguistic Data Management*. Cambridge: MIT Press, pp. 61–72.
- Méndez-Naya, Belén. 2003. On Intensifiers and Grammaticalization: The Case of SWIE. *English Studies* 84: 372–91. [CrossRef]
- Morin, Cameron, and Jack Grieve. 2024. The semantics, sociolinguistics, and origins of double modals in American English: New insights from social media. *PLoS ONE* 19: e0295799. [CrossRef]
- Nguyen, Dong, A. Seza Dođruöz, Carolyn P. Rosé, and Franciska De Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* 42: 537–93. [CrossRef]
- Nguyen, Dong, Maria Liakata, Simon DeDeo, Jacob Eisenstein, David Mimno, Rebekah Tromble, and Jane Winters. 2020. How we do things with words: Analyzing text as social and cultural data. *Frontiers in Artificial Intelligence* 3: 62. [CrossRef] [PubMed]
- Oliveira, Ricardo Soares. 2005. Sobre as relações entre Portugal e Angola ao fim de trinta anos: Um ensaio crítico'. *Relações Internacionais* 8: 55–72.
- Parodi, Giovanni, Pascual Cantos, and Chad Howe. 2021. *The Routledge Handbook of Spanish Corpus Linguistics*. London: Routledge.
- Peralta, Elsa. 2022. The return from Africa: Illegitimacy, concealment, and the non-memory of Portugal's imperial collapse. *Memory Studies* 15: 52–69. [CrossRef]
- Peters, Hans. 1994. Degree adverbs in early modern English. *Studies in Early Modern English* 13: 269–88.

- Pimenta, Susana. 2022. The Mestiço in the “Urgency of Existence”. *Essa Dama Bate Bué!* (2018), by Yara Monteiro. *Comunicação e Sociedade* 41: 61–73. [CrossRef]
- Poplack, Shana. 2017. *Borrowing: Loanwords in the Speech Community and in the Grammar*. Oxford: Oxford University Press.
- R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
- Roels, Linde, and Renata Enghels. 2020. Age-based variation and patterns of recent language change: A case-study of morphological and lexical intensifiers in Spanish. *Journal of Pragmatics* 170: 125–38. [CrossRef]
- Romerito Silva, José. 2019. Intensificação do verbo e mudança construcional. *Revista Soletas* 37: 224–45. [CrossRef]
- Rose, Stuart, Dave Engel, Nick Cramer, and Wendy Cowley. 2010. Automatic keyword extraction from individual documents. In *Text Mining: Applications and Theory*. Hoboken: John Wiley & Sons, Ltd., pp. 1–20.
- Scheffler, Tatjana. 2017. Conversations on twitter. In *Researching Computer-Mediated Communication: Corpus-Based Approaches to Language in the Digital World*. Edited by Darja Fišer and Michael Beišwenger. Ljubljana: Ljubljana University Press, Faculty of Arts, pp. 124–44. [CrossRef]
- Scheffler, Tatjana, Michael Richter, and Roeland van Hout. 2023. Tracing and classifying German intensifiers via information theory. *Language Sciences* 96: 101535. [CrossRef]
- Schweinberger, Martin. 2020. A corpus-based analysis of differences in the use of very for adjective amplification among native speakers and learners of English. *International Journal of Learner Corpus Research* 6: 163–92. [CrossRef]
- Schweinberger, Martin. 2023a. *Analyzing Co-Occurrences and Collocations in R*, 2023.05.31 ed. Brisbane: The University of Queensland, Australia. School of Languages and Cultures. Available online: <https://ladal.edu.au/coll.html> (accessed on 1 June 2023).
- Schweinberger, Martin. 2023b. *Part-of-Speech Tagging and Dependency Parsing with R*. Brisbane: The University of Queensland, School of Languages and Cultures. Available online: <https://ladal.edu.au/postag.html> (accessed on 1 November 2023).
- Serradilla Castaño, Ana Maria. 2006. “BIEN”+“ adjetivo” como perífrasis de superlativo en español: Particularidades semánticas y sintácticas. *Verba: Anuario galego de filoloxia* 33: 215–33.
- Statista. 2023. Leading Countries Based on Number of X (formerly Twitter) Users as of January 2023. Available online: <https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/> (accessed on 10 June 2022).
- Stuart-Smith, Jane. 2007. The influence of the media. In *The Routledge Companion to Sociolinguistics*. London: Routledge, pp. 140–48.
- Tagliamonte, Sali A. 2008. So different and pretty cool! Recycling intensifiers in Toronto, Canada. *English Language & Linguistics* 12: 361–94. [CrossRef]
- Tagliamonte, Sali, and Chris Roberts. 2005. So weird; so cool; so innovative: The use of intensifiers in the television series friends. *American Speech* 80: 280–300. [CrossRef]
- Tatman, Racheal. 2015. # go awn: Sociophonetic Variation in Variant Spellings on Twitter. *Working Papers of the Linguistics Circle* 25: 97–108.
- Thurlow, Crispin, and Alex Brown. 2003. Generation Txt? The sociolinguistics of young people’s text-messaging. *Discourse Analysis Online* 1: 30.
- Traugott, Elizabeth Closs. 1999. The rhetoric of counter-expectation in semantic change: A study in subjectification. In *Historical Semantics and Cognition*. Berlin and New York: Mouton de Gruyter, pp. 177–96.
- Trudgill, Peter. 1989. Contact and isolation in linguistic change. In *Language Change: Contributions to the Study of Its Causes*. Berlin: Walter de Gruyter, pp. 227–37.
- Wijffels, Jan, Milan Straka, and Jana Straková. 2018. Package ‘Udpipe’. Available online: <https://cran.microsoft.com/snapshot/2018-06-65112/web/packages/udpipe/udpipe.pdf> (accessed on 15 June 2022).
- Xatara, Claudia, and Mariele Seco. 2014. Culturemas em contraste: Idiomatismos do português brasileiro e europeu. *Domínios de Linguagem* 8: 502–19. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Marching towards Contrast: The Case of *ao passo que* in Portuguese

Manuel Delicado Cantero ^{1,*} and Patrícia Amaral ²

¹ School of Literature, Languages and Linguistics, The Australian National University, Canberra, ACT 2601, Australia

² Department of Spanish and Portuguese, Indiana University, Bloomington, IN 47405, USA; pamaral@indiana.edu

* Correspondence: manuel.delicado@anu.edu.au

Abstract: This paper explores the diachrony of the Portuguese contrastive connective *ao passo que* ('whereas'). First, we describe its syntactic and semantic properties in present-day European Portuguese. With this contemporary analysis in mind, we explore the semantic and syntactic changes from the PP *ao passo que* (lit. 'at the step/pace that') into first a temporal connective of simultaneity ('at the same time as') and, ultimately, a contrastive expression. The evolution of expressions with temporal meanings into contrastive ones has been documented in many languages. In our paper, we show that another related meaning, that of the gradual development of events that are temporally simultaneous, may also evolve into a contrastive meaning. We also examine the role of the syntax and semantics of the noun *passo* in this process. Furthermore, we discuss the internal analyzability of the connective and provide evidence for the retention of some internal syntax, which has implications for current theories on the nature of complex categories.

Keywords: contrast; connective; adversative; noun-based construction; adverbial subordination

1. Introduction

In contemporary Portuguese, *ao passo que* is a connective that establishes a contrast between two states of affairs, as in (1):

- (1) O Rui gosta do café cheio, ao passo que o Zé gosta dele curto. (Lobo 2013, p. 2029)
'Rui likes his coffee full, whereas Zé likes it short.'

The *Dicionário Priberam da Língua Portuguesa* (Dicionário Priberam 2008–2021, online)¹ defines *ao passo que* as both a temporal/proportional expression ('at the same time as') and as a contrastive one ('whereas') (see also Duarte 2003, p. 107).

In this paper, we explore the semantic and syntactic changes undergone by the prepositional phrase (PP) *ao passo que* (originally meaning 'at the step, at the pace that') to become a complex contrastive connective (complex complementizer), as in (1). The evolution of temporal connectives into contrastive ones has been documented. For example, the Spanish *mientras (que)* 'while, whereas' and the English *while* originally conveyed temporal simultaneity, then began implying the contrast between two simultaneous events, and finally came to encode the contrast between two eventualities² that are not necessarily simultaneous (Pavón Lucero 2003, pp. 281–282, a.o.). This semantic shift is similar to that of the Portuguese *enquanto (que)* 'while/whereas' (Lopes 2001; Costa 2005). Simultaneity is a temporal relation between events that is cross-linguistically linked to the creation of contrastive connectives (König 1985, pp. 10–11; Traugott and König 1991; Bybee et al. 1994). In this paper, we show that a related meaning, the *gradual development of temporally simultaneous events*, may also evolve to take on a contrastive meaning. In this case, instead of a temporal relation between events, we have a comparison between the development

Citation: Delicado Cantero, Manuel, and Patrícia Amaral. 2024. Marching towards Contrast: The Case of *ao passo que* in Portuguese. *Languages* 9: 6. <https://doi.org/10.3390/languages9010006>

Academic Editor: John Lipski

Received: 28 October 2023

Revised: 30 November 2023

Accepted: 7 December 2023

Published: 20 December 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

of two events: their gradual realization is asserted to happen at the same pace. Adopting a Neo-Gricean approach to semantic change (Traugott and Dasher 2002), we propose that an implication of contrast between both events, which could initially arise in some contexts containing *ao passo que*, became conventionalized over time, with the original comparative-proportional meaning being lost (for the most part).

This semantic shift is accompanied by a syntactic change resulting in the creation of a connective that no longer relies on a relative clause structure headed by the noun *passo*. We examine the evolution of the internal structure of the PP, especially the loss of the typical nominal properties of the noun *passo*, and the resulting internal syntax and analyzability. While the original structure grammaticalizes into a connective, we show that we may not overcast the complex relation with the meanings of the noun itself as ‘pace, rhythm, manner.’ Hence, we highlight the crucial role of the noun in the creation of the connective, a methodological approach that was proven to be crucial in earlier work (Amaral and Delicado Cantero 2019, 2022). In doing so, we propose a more nuanced approach to compositionality (or lack thereof) in the resulting grammatical item within accounts of syntactic and semantic change.

The structure of this article is as follows. We first examine the properties of *ao passo que* in contemporary Portuguese, relying mostly on existing work on European Portuguese³. Next, we present the diachronic data and describe the syntactic and semantic changes that we can observe, using some of the tests previously presented (Section 3). In Section 3.2, we zoom in on the meaning of the noun *passo* and examine the structures in which it occurs. Section 4 discusses theoretical questions related to the development of *ao passo que*, specifically with respect to its reanalysis and the analyzability of the expression at its different stages. Section 5 concludes with a summary of our contribution and suggestions for further work.

2. *Ao passo que* in Present-Day Portuguese: Syntactic and Semantic Properties

2.1. Syntactic Properties

In the literature, *ao passo que* is mentioned in sections on Portuguese adverbial subordination, where it is presented as similar to the contrastive uses of *enquanto (que)* ‘while’ (Costa 2005; Lobo 2013) and as expressing antithetic contrast (Duarte 2003). The criterion for adverbial subordination in European Portuguese (EP) is the triggering of proclisis, as can be seen with the direct object clitic *o* in (2), just like *enquanto* (3). In subordinate clauses, proclisis is preferred in EP, a language in which enclisis is found in main clauses and is generally preferred (see Luís and Kaiser 2016 for an overview and the current tendency to increase enclitic placement in EP).

- (2) O João leu parte do livro ao passo que a Maria o leu todo/*leu-o todo (Costa 2005, p. 503).
The João read.PST.3SG part.of.the book ao passo que the Maria CL.ACC.MASC.SG
read.PST.3SG all/read.PST.3SG-CL.ACC.MASC.SG all
‘João read part of the book, while Maria read the whole book.’
- (3) a. O João leu parte do livro enquanto a Maria o leu todo.
b. Enquanto o João leu parte do livro, a Maria leu-o todo.
‘João read part of the book, while Maria read the whole book.’

This is in contrast to coordinating conjunctions like *mas*, ‘but,’ which does not trigger proclisis, as shown by the enclitic position of the direct object clitic in (4); in this case, proclisis would be ungrammatical:

- (4) O João leu parte do livro mas a Maria leu-o todo/*o leu todo.

Usually, *ao passo que* is described together with *enquanto*, but there are other connectives with similar meanings, which, like *mas*, do not trigger proclisis:

- (5) O João leu parte do livro só que a Maria leu-o todo.
‘João read part of the book, while Maria read the whole book.’
- (6) O João leu parte do livro. Já a Maria leu-o todo.
‘João read part of the book, while Maria read the whole book.’

Lobo (2013) points out that *enquanto* may head either the first or the second clause, as in (3) and (7), while *ao passo que* can only head the first clause:⁴

- (7) Enquanto o João leu parte do livro, a Maria leu-o todo.
 (7') *Ao passo que o João leu parte do livro, a Maria leu-o todo.

In her description of contrastive connectives in EP, Costa (2005) mentions three types of contrast, based on Duarte (2003): antithetic, substitutive, and concessive (Duarte 2003). Duarte presents *enquanto* and *ao passo que* as heading contrastive subordinated clauses ('frases subordinadas contrastivas': 'exprimem valores semelhantes ao das justapostas ou coordenadas contrastivas antitéticas,' 'they express values similar to those of juxtaposed or coordinated contrastive antithetic [clauses],' p. 107, our translation). However, she does not define the criteria or tests for 'antithetic contrast'.

The literature also notes that *ao passo que* displays a dual nature, at least for some speakers: on the one hand, it is a temporal–proportional connective of simultaneity synonymous to *à medida que*, 'at the same time that, as, [temporal] while' (8) (cf. Cunha and Cintra 2017, pp. 604, 623); on the other hand, it is a contrastive connective interchangeable with *mas* 'but' and *enquanto*, as we have shown.

- (8) Ao passo que vamos a caminho do carro totalmente elétrico ou até com os modelos híbridos, veremos mais oportunidades para controlar cada motor (CdP, Portugal)
 'At the same time that/as we are on our way to a completely electrical car or even with the hybrid models, we will see more opportunities to control each engine.'

Interestingly, this value of *ao passo que* is not mentioned in the section on adverbial subordination that mentions proportional clauses (cf. Brito and Matos (2003), p. 765). As the authors point out, such clauses may be introduced by correlative expressions, like *quanto mais... tanto mais* 'the more... the X-er', or by isolated expressions, like *à medida que*, *à proporção que*, *enquanto*. Note that it is possible to replace the proportional expressions that occur in Brito and Matos's examples (2003, p. 765), our (9) and (10), with *ao passo que*, as in (11), and the sentences have a similar meaning:

- (9) Quanto mais se avança na montanha, mais difícil se torna a respiração.
 (10) À medida que as pessoas avançam na montanha, a respiração torna-se mais difícil.
 (11) Ao passo que se avança na escalada da montanha, a respiração torna-se mais difícil.
 'The more one climbs up the mountain, the harder it becomes to breathe.'

As we show below, originally, *ao passo que*, had a proportional comparative meaning, expressing both temporal and manner information (about pace, rate or rhythm).

While the literature adequately classifies *ao passo que* as a contrastive connective in contemporary Portuguese, little is said about the type of contrast expressed, or why it is not always interchangeable with the adversative connective *mas* (in terms of meaning, i.e., independently of the syntactic difference between subordination and coordination). We address such questions by turning now to a more detailed examination of the meaning of *ao passo que*.

2.2. Semantic Properties

In this section, we ask, how do we classify the contrast expressed by *ao passo que*? The literature on types of contrast and their expression across languages is vast (Lakoff 1971; Abraham 1979; Mann and Thompson 1988; Rudolph 1996; Malchukov 2004; Mauri 2008, a.o.). We first review some proposals about the meanings expressed by adversative connectives.

In Lakoff's seminal study on English *but* (1971), a distinction is established between contrastive and concessive meaning, or the difference between 'semantic opposition *but*' and 'denial of expectation *but*.' Corrective meaning, which is also associated with adversative connectives, is not discussed in this work because Lakoff is concerned with clausal contrast, not contrast within smaller units. The contrastive value identified by Lakoff is also known as antithetic (Quirk et al. 1985) or 'semantic opposition' (Malchukov 2004).

While Lakoff focuses on two types of contrastive meaning, Foolen (1991) and Izutsu (2008) propose a three-way classification for contrast. For Foolen, the difference is pragmatic, but Izutsu convincingly argues that it should be understood in semantic terms and that languages lexicalize these different meanings. Here, we adopt Izutsu's proposal and we focus on her notion of 'contrast,' which are expressed in English by *in contrast*, *conversely*, and *by comparison*.

Ao passo que can only express contrast, not concession (denial of expectation) or correction, as shown by the tests in Izutsu's study (2008). Compare the differences of acceptability in the examples below:

- (12) O Lino gosta de massa, ao passo que/mas a Patrícia gosta de arroz (contrast/semantic opposition)
'Lino likes pasta, while/but Patrícia likes rice.'
- (13) O Pietro é italiano, mas/*ao passo que não gosta de massa (concession/denial of expectation).
'Pietro is Italian, but/*while he doesn't like pasta.'
- (14) O Pietro não é italiano, mas (sim)/*ao passo que suíço (correction).
'Pietro is not Italian, but rather/*while Swiss.'

Izutsu presents several tests for contrast: contrast allows the following: (i) the reversal of the order of the clauses without a significant change in the meaning of the sentence, as in (12'); (ii) the replaceability of the contrastive connective by *and* without a significant change in the meaning of the sentence (12''); and (iii) the possibility of omitting the connective (12'''). As the author puts it: 'contrast is a simple opposition between the propositional contents of two symmetrical clauses' (Izutsu 2008, p. 658).

- (12') A Patrícia gosta de arroz, ao passo que o Lino gosta de massa.
- (12'') O Lino gosta de massa e a Patrícia gosta de arroz.
- (12''') O Lino gosta de massa. A Patrícia gosta de arroz.

Izutsu proposes that gapping is possible with contrast. Note, however, that syntactic factors may play a role here: the sentence with *ao passo que* in (15), with the predicate 'be valued' omitted from the second clause, is degraded for some speakers. This may be related to a coordinated structure instead of subordination.

- (15) Os homens são valorizados pelo dinheiro que fazem, ?ao passo que as mulheres pela aparência (acceptability is degraded without repeating the predicate).
'Men are valued by how much money they earn, ?while women by their looks.'

According to Izutsu, the preferred interpretation of pronouns in potentially ambiguous sentences is also a hallmark of contrast. Since the propositional contrast relies on symmetrical sentences, the personal pronoun in the second clause is interpreted as referring to the referent of the second referring expression in the first clause:

- (16) O João gritou com o Pedro, ao passo que o Paulo se emocionou com ele (ele = Pedro) (pronoun disambiguation relies on the parallelism between clauses).
'João screamed at Pedro, while Paulo became emotional with him (him = Pedro).'

Izutsu proposes that contrast can be analyzed along a set of parameters. We show that *ao passo que* meets these criteria by exemplifying each of these parameters. The first one is the mutual exclusiveness of the compared items in a shared domain. To exemplify this parameter, consider (17) (for CETEMPúblico n.d., see Rocha and Santos 2000):

- (17) Quanto ao peso dos encargos financeiros no valor das vendas, 57 por cento das empresas declara que são inferiores a cinco por cento, ao passo que três por cento revela encargos superiores a 20 por cento (Cetempúblico, ext616-eco-95a-1).
'As for the burden of expenses on the amount of sales, 57% of the companies declare that they are less than 5%, while 3% reports expenses above 20%.'

The compared items here are '57% das empresas' and '3% das empresas', which refer to different companies. The same topic is being discussed: how many expenses with sales do companies have? The values provided are mutually exclusive. As noted above for contrast, in (17) we could invert the order of the clauses without a significant change in meaning:

- (17') 3% das empresas revela encargos superiores a 20%, ao passo que 57% declara que são inferiores a 5%.

Importantly, 'compared items' are not equivalent to entities: one may compare the same entity at different points in time, and each would be a compared item. We can see each compared item as a 'contrastive topic' (cf. Umbach 2005):

- (18) O Newton não era conhecido na altura dele, ao passo que hoje em dia é muito conhecido.
'Newton was not known in his time, while nowadays he is very well-known.'

The parameter under consideration involves a shared domain and the mutual exclusiveness of the compared items in that domain (i.e., occupying mutual exclusive regions of one shared property). For example, in (19), the domain could be formulated as 'availability of players,' and two mutually exclusive values are possible: players are either available or not available:

- (19) O Olympiakos tem vantagem teórica, por jogar em casa e por ter todos os jogadores disponíveis, ao passo que o Ajax tem lesionada quase uma equipa inteira: Ronald De Boer, Babagida, Oliseh, Litmanen Hoekstra (*Cetempúblico ext3095-des-98b-1*).
'The Olympiakos has an advantage (in theory), because it plays at home and all its players are available, while almost the entire team of Ajax has lesions. . .'

Crucially, both the existence of compared items and a shared domain are necessary for a similarity relation. However, what distinguishes similarity from contrast is mutual exclusiveness⁵. The close relation between similarity and dissimilarity is also noted by Mann and Thompson (1988, p. 253), as well as by authors working on *sidi* terms like *other* (Charnavel 2015), a point that will be relevant when we discuss the semantic change undergone by *ao passo que*. The identification of contrast often relies on a "'similarity and dissimilarity" interpretation. . . [that] presupposes semantic homogeneity of both objects compared" (Malchukov 2004, p. 183).

Another example that illustrates mutual exclusivity is provided by (20):

- (20) . . . as acções de Torres Pereira seriam sempre mais notórias, porque pertencia ao partido do Governo, ao passo que ele pertence à oposição (*Cetempúblico ext1872-soc-92a-2*).
'[T]he actions of Torres Pereira would always be more striking because he belonged to the party in government, while he belonged to the opposition.'

Mutual exclusiveness is clear in this example: one cannot belong to both the government and to the opposition party.

Another relevant parameter in the analysis of adversative connectives is the number and type of compared items. *Ao passo que* is typically associated with two compared items, as shown by (21):

- (21) O Lino gosta de massa, o António gosta de batata, mas/# ao passo que a Patrícia gosta de arroz (comparison is binary).
'Lino likes pasta, António likes potatoes, but/#while Patrícia likes rice.'

In summary, Izutsu defines contrast as requiring the following: '(i) Two or more different compared items [that] occupy mutually exclusive regions in a shared domain. (ii) The compared items (CIs) must be explicitly differentiated' (Izutsu 2008, p. 661).

Izutsu notes that unlike what happens in concession or correction, in a contrastive relation, the propositional contents of all the segments are asserted as valid, i.e., in uttering a sentence that expresses contrast (such as S1 CONNECTIVE S2), 'the speaker simply asserts the validity of both S1 and S2; neither claim is rejected.' (661). This means that contrastive connectives are factive, since the truth of both clauses is presupposed to be true. This is common to temporal connectives that express a proportional comparison: in asserting that an event occurred at the same pace as another event, both events are presupposed to hold.

2.3. Compatibility with Aspectual Classes of Verbs

According to Lobo (2013), the contrastive value of *ao passo que* is found only with stative predicates. It is true that stative predicates provide the best proof of the semantic change

undergone by *ao passo que*, since they are not compatible with a proportional interpretation (an eventuality taking place at the same rate as another). However, as we show below, it is possible to find examples of the contrastive meaning of *ao passo que* with non-stative predicates. In (22), we find the contrastive reading with achievement predicates (a change of state verb: *passar de . . . para* ‘become’) and in (23), below, with accomplishment predicates (*classificar-se em quarto lugar*, ‘to reach the fourth place’) and achievements (*alcançar um lugar cimeiro* ‘to reach a higher place’):

- (22) Na parte ocidental do país, a taxa de desemprego estabeleceu-se nos 8,2 por cento em Dezembro, contra 7,9 por cento no mês anterior, ao passo que no Leste ela passou de 13 por cento em Novembro, para 13,5 por cento em Dezembro. (*Cetempúblico ext2769-eco-95a-2*).
 ‘In the western part of the country, the unemployment rate reached 8.2% in December against 7.9% in the previous month, while in the east it became 13.5 in December, a change from 13% in November.’

The components of a temporal reading (simultaneity) are present in (22). The context allows us to retrieve a time interval of which the change in the unemployment rate is predicated:⁶ it is the time interval from November to December. Note that in this example, the change is sudden, as it is expressed by discrete values. However, in the earlier examples of *ao passo que*, we find gradual changes, which are often expressed by incremental change verbs like *arrefecer*, ‘to cool,’ or *crescer*, ‘to grow,’ which are known as degree achievements (Dowty 1979) or gradual completion verbs (Bertinetto and Squartini 1995). In other words, the comparative proportional meaning, which is the source meaning of *ao passo que*, presupposes the existence of a time interval (either provided by a temporal adverbial or contextually retrievable) during which two eventualities hold. The pace of development of these eventualities is compared. Hence, the presence of a time interval in the context is a helpful clue to identify the original temporal meaning of the expression. In Section 3, as we examine the semantic change of *ao passo que*, we will use this test for the (earlier) temporal interpretation. In (22), the time interval in which the changes take place is the same in both clauses (from November to December), but both the starting and the final amounts are different, indicating a contrast between western and eastern regions. The example shows that while a temporal component is present in (22), different amounts and rates of change are predicated of the compared regions. In sum, we have both a temporal and a contrastive interpretation.

While in (22), the contrast relies on the different values of the amounts presented, in many examples, the contrast is expressed by the meanings of the predicates in the two clauses. In (23), there is no explicit time interval, although one may reconstruct a time interval corresponding to the extent of the car competition. Here, although we have eventive predicates, the interpretation of the sentence is not temporal. The comparison being established is between ‘reaching a good classification’ and ‘reaching a poor classification’ as a result of the performance in the races. We have compared items (the two pilots whose performance is compared), a shared domain (their classification in the races), and mutually exclusive values in the property being considered (a good vs. a poor result).

- (23) Brundle, mesmo assim, classificou-se em quarto lugar, à frente de Blundell e Panis, autores de provas de muito bom nível, particularmente o inglês, ao passo que o campeão europeu de F3000, sexto na primeira volta, perdeu demasiado tempo nas «box» para poder lutar por um posto mais cimeiro (*Cetempúblico ext18954-des-94b-2*).
 ‘Brundle, despite everything, reached a 4th place classification, ahead of Blundell and Panis, who performed at very high levels, especially the English [competitor], while the European champion of F3000, who was 6th in the first round, wasted too much time in the box to be able to fight for a higher position.’

The relation between the temporal and the contrastive meanings and the role of the predicates in each clause will be discussed in the diachronic analysis, to which we turn now.

3. The Diachrony of *ao passo que*

3.1. *Ao passo que*

The presence of *ao passo que* in the Portuguese language is documented since the 17th century. The number of tokens in the *Corpus do Português* (CdP) (Davies 2006) is limited before the 19th century (with nine examples in the period from the 13th to the 18th century, as opposed to 394 tokens from the 19th c. alone). The earliest examples show a PP with a relative clause modifying the noun *passo*: the event denoted by the verb in the relative clause is realized at the same pace as the event denoted by the verb in the main clause. Semantically, they convey the temporal simultaneity of dynamic eventualities, including gradual change (cf. the degree achievements like *aquecer*, *melhorar* in (24)), and indicate that the eventualities denoted by such predicates occur ‘at the same rate’:

- (24) a. ao passo que o tempo vai aquecendo vou eu também melhorando do achaque do estômago (*Cartas*, Vieira, 17th c.).
 ‘as the weather gets warmer I also gradually improve of my stomach ailment.’
 b. “Ao passo que iam procedendo os tempos-diz São Gregório-ia juntamente crescendo a sabedoria dos antigos Padres, conhecendo sempre mais de Deus os segundos que os primeiros. Moisés soube mais das cousas divinas que Abraão; os Profetas mais que Moisés; os Apóstolos mais que os Profetas” (*História do futuro*, Vieira, 17th c.).
 ‘As time went by, says S. Gregory, so increased the wisdom of the old fathers, such that the second ones knew more about God than the first ones. Moses knew more about the divine matters than Abraham, the prophets more so that Moses, the apostles more so than the prophets.’

Note that in addition to the temporal and proportional meaning, at least in some of these examples, a causal implication can be inferred: for example, in (24a), the warmer weather can be seen as a cause of the health improvement.

In both examples, there is a wh-gap in the embedded relative clause, namely, in (24a), the weather is heating up at a particular rate/pace/rhythm, while in (24b), time is moving forward at a particular rate. In both examples, a time interval is contextually assumed: in (24a), it is a time interval that contains utterance time, and in (24b), it is a time interval in the remote past that is being discussed by the Saint. As we showed above, the existence of a salient time interval is an essential component of the temporal interpretation. Because the eventualities involved are simultaneous, a comparison of pace (‘at the same rate/pace as’) is obtained as well. Further evidence comes from the use of *ao passo que* and *à medida que*⁷ in the same text (25); notice that both combine with the same verb *crescer* (‘to grow’):

- (25) crescendo sempre nela ao passo que ia crescendo nos anos . . . ; assim a Igreja . . . vai sempre crescendo mais e mais na luz e na sabedoria, à medida que cresce nos anos e na idade (*História do Futuro*, Vieira, 17th c.).
 ‘growing in it at the same rate as it was getting older [lit. growing in years] . . . thus the church grows more and more in light and wisdom at the same pace as it gets older [lit. it grows in years and age].’

The limited evidence in the 18th c. shows that *ao passo que* was combined with stative verbs, as in (26):

- (26) Já se ve que—ao passo que os educandos souberem a sua língua, a latina e a francesa, a geografia, a cronologia e os elementos da história—devem passar a outras classes, onde se ensinarão as ciências que dependem destes conhecimentos (*Cartas sobre a educação da mocidade*, Sanches, 18th c.).
 ‘We can see that, as soon as the students know their language, Latin, French, Geography, the chronology and elements of History, they must go on to other classes, where the sciences that rely on that knowledge will be taught.’

In (26), *ao passo que* combines with *saber* (‘to know’), and we may interpret the subordinate clause as ‘as soon as they come to know those subjects,’ i.e., *à medida que*, rather than as a relative clause.

In (27), we find that a relative analysis is even more questionable; rather, *ao passo que* functions as a temporal–comparative connective, ‘at the same time as’, even with a potential implication of contrast.

- (27) Lamentavão a sua falta com copiosas lagrimas as suas ovelhas quando em de Fevereiro de 1739 dos seos olhos foy tirado para Arcebispo da Bahia; ao passo que nesta cidade era recebido com as demonstraçoens mais fervorosas de contentamento, mas cedo experimentarao as da Bahia o mesmo golpe sendo transferido para Bispo da Guarda no mesmo anno de 1739 (*Desagravos do Brasil e glórias de Pernambuco*, Coutto, 18th c.).
 ‘His sheep (followers) regretted his absence with many tears when in February 1739 he was relocated as archbishop of Bahia, at the same time as he was received with demonstrations of joy in that city; but soon the people of Bahia were to experience the same sorrow because he was transferred to Guarda that same year of 1739’.

Despite the fact that the number of tokens is limited for this period in the corpus we consulted, we must note examples with habitual aspectual values that receive contrastive readings. In these cases, one may still assume that an underlying time interval is being considered, during which two (disparate or otherwise conflicting) eventualities take place:

- (28) a. Castigao com notavel rigor latrocinios, ao passo que outros delictos, ou são canonisados como virtudes, ou não castigados como culpas (*Desagravos do Brasil e glórias de Pernambuco*, Domingo do Loreto Coutto, 18th c.).
 ‘They punish thefts with severe rigor, whereas other crimes are either canonized as virtues or are not punished as crimes.’
 b. a dor se diminuirá ao passo que a resignação crescer (*Eva e Ave ou Maria Triunfante*, de Macedo, 18th c.)
 ‘the pain will decrease as the resignation grows.’

In some of these examples, we find stative verbs, as in (28c):

- c. A Lypsio, que se gloriava, que ao passo que a sua Obra nao tinha cousa sua, era cousa sua: Omnia nostra, et nihil (*Escola moral, politica, christãa, jurídica*, Aboym, 18th c.)
 ‘To Lypsio, who boasted that his work was his at the same time as it did not have anything of his.’

While the temporal–comparative meaning remains, the fact that the predicates involved denote continuous activities (habits) or are stative renders exact simultaneity irrelevant. Furthermore, the lexical meaning of the predicates themselves contributes to the contrastive implicature. For example, in (28a), *castigar* (‘to punish’) is opposed to *ser canonisado* (‘to be canonized’). The negation of one of the predicates contributes to a contrastive reading in (28c); such predicates may not be ordered in time because states do not happen, they hold. Regarding the syntax, no relative-clause analysis seems plausible.

We have shown that, originally, *ao passo que* could only connect two clauses with eventive predicates and no contrast was expressed. Using the terms introduced in Section 2, in the original examples, there were compared items and a shared domain (the same time interval), but no implication of mutual exclusivity. This semantic implication, which is a hallmark of the meaning of *ao passo que* in the contemporary language, is the conventionalization of an inference that could arise in certain contexts, when there was an unexpected difference between the compared items.

By the 19th century, we find examples of temporal simultaneity (29a,b), examples displaying ambiguity (notice both the contrastive predications ‘writing an order to leave’ and ‘writing a resolution to obey,’ and the presence of *ao mesmo tempo*, ‘at the same time’ in (29c)), and examples expressing contrast (30):

- (29) a. O escravo recebeu a carta e na sala de jantar, entregou-a a Dusá, ao passo que dizia (*Maria Dusá*, Rocha, 19th c.).
 ‘The slave received a letter and gave it to Dusá in the dining room while saying: ...’
 b. Entretanto, ao passo que assim pensava, uma agitação extrema o perturbava, como se tivesse diante de si um tesouro inapreciável a que bastasse estender a mão para o possuir (*O Missionário*, Sousa, 19th c.).

‘Meanwhile, while he was thinking thus, an extreme agitation shook him, as if he had in front of him an invaluable treasure which could be owned simply by extending his hand.’

c. Angelina escreve ocultamente, uma ordem de partir, ao passo que Teófilo escreve em outro papel, ao mesmo tempo, a sua resolução de obedecer; os dois bilhetes são lidos na mesma ocasião (*Textos críticos*, Machado de Assis, 19th c.).

‘Angelina writes an order to leave while Teófilo writes, on a different piece of paper, at the same time, his resolution to obey; both notes are read on the same occasion.’

Example (29c) instantiates a stage at which the two implications associated with *ao passo que*, contrast and simultaneity, are conveyed: (i) there is contrast between the two compared items (what is predicated of Angelina and what is predicated of Teófilo) and (ii) these two events occur at the same time. The contrastive meaning was originally an invited inference, i.e., an implication triggered in some contexts only, and the simultaneity meaning was the semantic content of the construction. However, in this example, it is impossible to determine the status of each implication, given the contrast in meaning between the predicates and the explicit temporal information about the events occurring simultaneously.

(30) a. Mas ao passo que o marchante não poupava finezas nem esforços para prender definitivamente o negociante, lançava-lhe estas outras contas muito diferentes (*O Matuto*, Távora, 19th c.).

‘But, whereas the dealer profusely courted the businessman, the businessman told him very different things.’

b. Ambas avançam para o desconhecido. Mas, ao passo que a ciência caminha, a poesia voa (*Flores da Noite*, Paiva, 19th c.).

‘Both advance toward the unknown, but, whereas science walks, poetry flies.’

Syntactically, these examples do not accept a relative analysis. In (29a) and (29b), the subject is not saying something or thinking at a particular rate/pace/rhythm. No comparison of rate or pace is logical in (30a,b) either. Notice that the movement predicates in (30b) do not force a temporal–proportional interpretation. What is contrasted is not the rates at which science moves and poetry flies, but rather the fact that science walks whereas poetry flies (see ex (23) above).

In this period, the contrastive *ao passo que* is attested in two positions, even in the same text, both preposed and after the main clause. In (31a,b), stative predicates are present and the comparative–proportional interpretation is no longer possible:

(31) a. Vivia muito bem com a mulher e ajudava o pai na lavoura, ao passo que o outro era um vadio, cheio de idéias esquisitas, um poeta, afinal! (*O Missionário*, Sousa, 19th c.).

‘He had a very good life with his wife and helped his father at work, whereas the other one was lazy, full of weird ideas, a poet after all!’

b. ao passo que esta oferecia um perfeito espécime da mais virgem e rude mata do Amazonas . . . , o local do sítio do velho tuxaú fora completamente modificado por mãos inteligentes de homem de bom gosto (*O Missionário*, Sousa, 19th c.).

‘whereas this one provided a great sample of pure Amazonian forest, the old location of Tuxaú had been completely modified by the intelligent hands of a man with good taste.’

These examples clearly demonstrate that a semantic change has occurred: at this stage, the expression *ao passo que* primarily encodes contrast. We build here on the theory of invited inferencing in semantic change (Traugott and Dasher 2002), which assumes that contextual implications that are sufficiently systematic to be generally associated with the use of a certain expression may become semanticized over time. As Eckardt puts it in her definition of semantic reanalysis: ‘What may have previously been in part assertion, in part implication, turns entirely into a literal assertion after reanalysis. Semantic reanalysis may have repercussions on the meaning of parts of the sentence (constructions, phrases, words, affixes), leading to a changed meaning under the new semantic composition of the sentence’ (Eckardt 2006, p. 236). While, in earlier examples, we identified a contrastive implication that can be inferred from the simultaneous development of two (somehow incompatible) events, in these two examples, there are no longer simultaneously developing events, only contrasting states of affairs.

3.2. The Noun *Passo*

The noun *passo* is historically attested with different senses, namely, ‘step, distance advanced by the act of walking’ and ‘pace, progress.’ Other meanings, such as ‘crossing’, are also documented but are irrelevant for our discussion. *Passo* is also attested as an adverb with the meaning ‘slowly’⁸.

In terms of its morpho-syntactic properties, *passo* as ‘step’ shows the properties of regular countable nouns, namely adjectival and PP modification and pluralization:

- (32) a. Assý começarõ a yr en bõo passo et ajuntados en tropel (*Cronica Troyana*, 14th c.).
‘Thus they started to go at a good pace and all together as a mob.’
b. Et avia em cada quadra em longo sete mill et quinētos passos, segundo amedida da geometria (*General estoria*, 15th c.).
‘Each block was 7500 steps long, according to the measurements of geometry.’
c. quãdo semtio os passos dos escuitas & o rramalhar que faziã pello milho, cuydou que herã os porcos que vinhã comer (*Crónica do Conde D. Pedro de Meneses*, 15th c.)
‘when he heard the steps of the explorers and the noise they were making in the cornfield, he thought that it was the pigs coming to eat.’

Note that in (32a), the interpretation of pace/rate/rhythm is arguably already present.

Within the PP *a X passo*, the N may be modified by quantifiers (like the distributive *cada* in (33a)), demonstratives (the deictic form *este* in (33b)), and adjectives, such as *mesmo* or *cheio*, as in (33c–e):

- (33) a. Pelo que em nossas cazas estavamos de noite e de dia esperando quando havião tambem de dar sobre nós, e a cada passo nos vinhão diversas novas (*Historia do Japam* 3, Frois, 16th c.).
‘For which reason we stayed home day and night waiting for the moment when they would attack us, and at each step we got different news.’
b. Bem está tudo isso, mas nos mal, se a este passo havemos de examinar toda esta livraria. Sendo assim que, só para a estante dos poetas portugueses, que agora nos fica a mão, necessitamos de muitos dias de conferencia (*O hospital das letras*, Melo, 17th c.).
‘All that is good but it is bad for us if we must examine this whole library at this pace, seeing that, just for this shelf with Portuguese poets, which we now have at hand, we need many days of meeting.’
c. Depois que entrou Abril se esfriaram notavelmente os dias, e ao mesmo passo se atrasou a saúde (*Cartas*, Vieira, 17th c.).
‘After April started, the days became colder and, at the same rate, our health deteriorated.’
d. Cresce com ellas ao mesmo passo mayor o risco (*Escola moral, política, christãa, jurídica*, Aboym, 18th c.).
‘With them, the risk grows bigger at the same time.’
e. marchando a passo cheyo ao som dos seus estromentos (*Peregrinação*, Pinto, 17th c.).
‘marching at full pace, to the sounds of their instruments.’

In (33) a and b, *passo* has a temporal meaning: it denotes the rate/pace/rhythm in which the action takes place. The same meaning is available in (33e), which includes a reference to the musical tempo, the physical movement of the march (i.e., literally, the steps) combining with its rhythm. The modification with *mesmo* (33c,d) expresses a comparison of equality between the pace of the development of the events in each clause (for example, the days becoming colder, health diminishing in 33c).

We have attested a few examples with *passo* modified by a PP introduced by the P *de* ‘of’:

- (34) a. Ao passo da vontade anda o merecimento (*A vida de Frei Bertolameu dos Mártires*, Luis de Sousa, 17th c.)
‘merit walks at the (same) pace of will.’
b. os membros fraquejaõ, o juízo vacilla, as remissoens crescem ao passo das obrigaçoens a que se deverá acodir? (*Eva, e Ave, ou Maria Triunfante*, António de Sousa Macedo, 18th c.)
‘the limbs weaken, the judgement falters, the remissions grow as grow the obligations to which one must attend.’

Furthermore, *passo* is documented in other PPs with similar interpretation as those documented in combination with the P *a*:

- (35) a. para nao andarem em igual passo os pensamentos com as suspeitas (*Casamento perfeito*, Andrada, 17th c.)
 ‘so that thoughts and suspicions would not go at the same pace.’
 b. Mas quanto ao gasalhado, do mesmo passo se foram ao Hospital de Todos os Santos, pretendendo igualmente viver com os pobres em pobreza e exercitar com os enfermos a misericórdia (*História da vida do Padre S. Francisco Xavier*, Lucena, 17th c.)
 ‘But regarding their clothes, at the same time they went to the Hospital of All Saints, aiming to live with the poor in poverty and to apply mercy to the sick people.’

The role of the P *a* is also accountable on independent grounds. As Fábregas (2007) and Romeu Fernández (2014, pp. 109–16) indicate for Spanish, the locative P *a* combines with scalar nouns (e.g., ‘level’), a description we extend to Portuguese. These PPs indicate a point in a stative locative scale; while not literally locative, the semantics of *passo* are similar to such scalar nouns, which justifies its combination with *a*.

The importance of examining the properties of the N *passo* when studying the construction *ao passo que* is further supported by the fact that the same author may use both in the same text. In (36a), Frois uses both *a cada passo*, as in ‘at each step,’ with an interpretation of action progressing at a particular rate, and the temporal–proportional *ao X passo que* in the same book. Similarly, Melo uses the temporal–proportional *ao mesmo passo* and *ao mesmo passo que*, as shown in (37):

- (36) a. Metidos os Padres no fogo de tamanha aflicção e renovando-se a cada passo com mais terror as novas que sem faltar corrião pela terra, determinarão os Padres de consultar com os poucas christãos que alli havia o remedio que se poderia achar para se salvarem daquella furia (*Historia do Japam 3*, Frois, 16th c.)
 ‘Once the priests were very sad and, since more terrific news were propagating around the world at each step, the priests decided to consult with the few Christians there on the solution they could find in order to be saved from that horror.’
 b. vinhão todos cheios de gente fugindo do destroço, os quaes a cada passo que encontravão o Padre lhe dizião mil injurias (*Historia do Japam 3*, Frois, 16th c.)
 ‘(all the paths) were full of people feeling destruction, people who, every time they met the priest, insulted him profusely.’
 (37) a. que ao mesmo passo crece a luz e crece a sombra (*Cartas familiares*, Melo, 17th c.).
 ‘that light and shade grow at the same rate.’
 b. tendo por certo, que ao mesmo passo que o Reyno merecesse a Castella hum grãde castigo, ficaria elle absoluto senhor dos Portuguezes (*Epanaphora politica primeira*, Melo, 17th c.).
 ‘being sure that, at the same time that the kingdom would punish Castile, he would become the absolute lord of the Portuguese.’

To finish this section, we want to add that, while rare, we have attested one example of *ao mesmo passo que* that appears to be more readily interpreted as contrastive:

- (38) A paz, a opulência, o luxo . . . eram os tiranos algozes que apertavam o garrote ao comércio de Portugal, ao mesmo passo que sem eles fora impossível beneficiá-lo (*Obras completas*, Correia Garção, 18th c.).
 ‘Peace, opulence, luxury . . . were the cruel tyrants that suffocated Portugal’s commerce, at the same that, without them, it would have been impossible to benefit it.’

We conclude this section by providing evidence of the regular nominal syntax of *passo* as ‘step’ outside of a connective, which is especially common with the dynamic verb *dar*:

- (39) A cada passo que damos no discurso da vida, imos nascendo de novo, porque a cada passo imos deixando o que fomos, e começamos a ser outros (*Reflexão sobre a Vaidade*, Matias Aires, 18th c.)
 ‘Each step we take in life, we are born again because with each step we leave behind who we were and start to be different people.’

Interestingly, the same example contains the gradual temporal meaning of *a cada passo*, ‘at each step’, expressing a comparison of equality between the progression of the events associated with change (leaving behind who we were, starting to become new people).

4. Discussion

The earliest instances of the expression *ao passo que* show a PP with a N modified by a relative clause (‘X takes place at a (particular) pace that. . .’). A century later, a second form of *ao passo que* as a temporal–proportional–comparative connective (‘X at the same pace/time as Y’) with no relative syntax (there is no wh-gap in the embedded clause) emerged. The final outcome is the contrastive connective. All three are in use today—layering (cf. Hopper 1991)—even though the contrastive connective is more common than the temporal–proportional connective in present-day Portuguese.

An important question has to do with the internal syntactic structure or the analyzability of the entire expression. The examples with a clear relative analysis display free syntax; the temporal–comparative and contrastive connectives, however, are less free. Research on complex constructions—referred to as *locutions* in some works—has long noted their loss of internal syntax, arguing that the final outcome is an unanalyzable unit. At the same time, much of that research highlights that this loss is gradual (e.g., Brems and Davidse 2010, p. 112; Traugott and Trousdale 2013).

Montoro del Arco (2006, pp. 218–20) lists different criteria to test the degree of internal fixation, namely:

- a. Neither *que* nor the N may be substituted. However, the N may sometimes vary with no semantic change (e.g., the Spanish *de manera/modo/forma/suerte que*, ‘in order that’).
- b. The N is invariable (e.g., no pluralization) and may not be modified.
- c. Some degree of semantic specialization, with a more abstract meaning of the N, which is not referential.
- d. It functions as a simple connective.

These criteria are neither absolute, as there may be some degree of variation, nor necessary conditions, as not all tests need to pass for the proposed definition of *locución conjuntiva* (lit. *conjunctive locution*; connective as complex complementizer, in our terms) to stand (see also Tornel Sala 2000, pp. 160–70; Cifuentes Honrubia 2003, pp. 104–10 for further discussion and additional references on fixation; cf. Fagard et al. 2020, p. 9, 13; a.o.).

The connective *ao passo que* passes the previous tests. We have shown in the data that *que* may not be changed, that the N is morphologically invariable, etc. Let us explore some of these points:

1. The N was always invariable in the connective. The use of the N *passo* in plural can be found, but we found no attestations of **aos passos que*, even though the N *passos* in plural is used in other constructions (e.g., *dar passos*). As noted in the literature (cf. Ruiz Gurillo 1997, n/p, who cites Zuluaga (1980, p. 99)), the loss of pluralization is not testable in *ao passo que*, but rather only in comparison to the evidence of *passo* in other contexts.
2. We documented some evidence of the modification of the N in the temporal–comparative interpretation (e.g., *mesmo*), although the majority of the examples we retrieved do not include modification. The contrastive *ao passo que* does not seem to accept any adjectival internal modification of the N (see the exception in (36)).
3. In the connective, we do not have examples of *passo* anaphorically referred to later on in the same textual context, i.e., in a hypothetical (unattested) example like **ao passo que. . . e a tal passo. . .*

The nominal properties of *passo* are limited in the data that we identify as displaying the connective. However, the occasional modification with *mesmo*, for instance, poses a theoretical challenge to the idea of unithood.

In fact, in the literature, it is accepted that some variation is possible in fixed constructions. Montoro del Arco (2006, pp. 180–81) states that only a few constructions are

completely fixed from a formal point of view. On the list of accepted variations, the author lists, for Spanish, some modification by *mismo*, *justo* or *exactamente* ('same', 'just', 'exactly') and the possibility of interpolation (Corpas Pastor 1996, pp. 27–30; Cifuentes Honrubia 2003). We note interpolation in contemporary European Portuguese, with *ao passo que* having a clear contrastive meaning:

- (41) a. *Ao passo porém que* Plínio coloca o seu povo na margem esquerda do Minho . . . , Ptolemeu coloca-o na margem direita (1930, https://www.patrimoniocultural.gov.pt/static/data/publicacoes/o_arqueologo_portugues/serie_1/volume_29/157_seurros.pdf), accessed on 16 November 2023.
'However, whereas Pliny situates his town on the left bank of river Minho, Ptolemy places it on the right bank.'
- b. *Ao passo, porém, que* os deuses gregos são objectivações formais dos instintos humanos, os deuses nórdicos são objectivações amorfas (Pessoa, <http://arquivopessoa.net/textos/3909>), accessed on 16 November 2023.
'However, whereas the Greek gods are representations of human instincts, the Nordic gods are amorphous representations.'

Interpolation is relevant because it is another key test used in the literature to diagnose the gradual loss of internal analyzability. For example, it is one of the criteria in Torres Cacoullos's study (2006) of the Spanish *a pesar de* ('in spite of'); in this work, a fixed construction disallows interpolation.

Whereas previous studies acknowledged degrees of loss of internal and even degrees of syntactic unithood (Corpas Pastor 1996, p. 106), we consider that such flexibility ultimately means that a reduction in internal analyzability does not equal the presence of an unanalyzable unit. From a theoretical point of view, the question has to do with whether there is any internal syntax and with the theoretical concept of (syntactic) *unit*.

Jiménez Juliá (2017, pp. 100, 104–5) states that a string of words may be a unit or a construction but, crucially, that the former is defined by its being completely unanalyzable (as in univerbation, e.g., Spanish *pero* 'but'), whereas the latter may be part of a productive model and have internally recognizable parts. Jiménez Juliá notes that even free syntax is subject to idiomatic interpretations and to co-collocation restrictions. We note here that *ao passo que* exists alongside other examples, such as *ao tempo que* or *à medida que*.

The evolution of the connective *ao passo que*, with its less free syntax compared to the examples of relative clauses adduced above, is in line with the conclusions for other complex constructions in some recent formal literature. Formal studies have accounted for the lexically reduced nature of the nouns involved, e.g., AxParts in complex locative Ps in Svenonius (2006, 2010) or weak nouns in Matushansky and Zwarts (2019, 2021). This approach argues that some of the constituents in complex constructions have been recategorized as functional categories, the decategorialization typically argued for in usage-based grammaticalization studies, which explains why their syntax is not lexical. The key point is that these formal studies propose some internal structure; there is no univerbation or complete loss of internal analyzability. In our case, we argue that the lexical N *passo* was—and continues to be—a regular lexical N when modified by a relative clause. As part of the connective *ao passo que*, *passo* became recategorized as a functional, weak, lexically reduced N, similar to the concept of AxParts but with the scalar meaning discussed by Fábregas (2007) and Romeu Fernández (2014). This lexically reduced N is part of a complex complementizer that is formally a PP and that also includes a semantically void determiner *o* and the complementizer *que*⁹. The presence of internal syntax is more evident in the temporal–proportional–comparative connective, since it allows some modification by *mesmo*, and much less so in the contrastive connective. In the contrastive connective, the possibility of internal modification is null and the loss of semantic compositionality contributes to its interpretation as a block¹⁰. Nevertheless, the interpolation in (41) suggests that some internal syntax also remains in the contrastive *ao passo que*.

5. Conclusions

In this paper, we examined the syntactic and semantic development of the expression *ao passo que*, which, to the best of our knowledge had not been studied previously. We show that this is a relatively recent change in Portuguese from a relative clause structure, conveying temporal and comparative meaning (e.g., ‘at the rate/pace that X grows’—‘X grows at a particular rate’), to a temporal–proportional–comparative connective (‘at the same pace/rate/time as’) and, finally, to an adversative connective (‘whereas’).

Semantically, *ao passo que* could initially only connect two clauses with eventive predicates, whose pace of development was compared. In the source expression, there was no contrast, as we saw in examples like (23a). In such comparative expressions, there were two compared items and a shared domain (a temporal interval), but no mutual exclusivity. The semantic change into adversative meaning resides in the original temporal meaning, whereby the development of two events in the same time interval and at the same pace invites a comparison. As noted above, a “similarity and dissimilarity” interpretation. . . [that] presupposes semantic homogeneity of both objects compared’ (Malchukov 2004, p. 183): while both situations compared gradually develop over the same time interval, their outcomes receive unequal evaluations. We proposed that currently, this semantic implication is the conventionalization of an inference that could arise in certain contexts, when there was an unexpected opposition between the compared eventualities, i.e., by Gricean reasoning, expectations about the pace and the outcome of two events lead to a new meaning of the comparative expression (contrast), which eventually generalizes, now allowing for stative predicates. The proportional implication is no longer present, i.e., the meaning of pace or rhythm has been lost and the asserted meaning in contemporary Portuguese is one of semantic opposition. While temporal simultaneity is a common source of contrastive meanings, semantic shifts from comparative meanings of gradual development over time (‘at the same pace, at the same rhythm’) have been less studied.

Syntactically, we discussed the loss of internal analyzability typically found in the creation of complex constructions. We provided some evidence that questions the total loss of internal syntactic structure in the connective, following theoretical assumptions in recent formal syntactic research. Furthermore, we traced the syntactic and semantic evolution of the noun *passo* itself. This evidence proves that taking the evolution of the noun *passo* is crucial in understanding the overall evolution of the connectives that include it. In doing so, we are able to situate the evolution of the connective *ao passo que* within the diachrony of other noun-based constructions in Portuguese (and in Spanish), in line with the conclusions reached in previous work (Amaral and Delicado Cantero 2022).

This paper contributes to the study of the diachronic development of adversative connectives in Portuguese, a topic that has barely been investigated. The semantic paths that are mentioned in works on Portuguese historical syntax (e.g., from causal to contrastive) still lack support from an empirical perspective. More broadly, this paper contributes to current debates on the semantic and syntactic development of complex constructions, including key concepts such as compositionality and internal analyzability.

Author Contributions: Conceptualization, M.D.C. and P.A.; methodology, M.D.C. and P.A.; formal analysis, M.D.C. and P.A.; investigation, M.D.C. and P.A.; resources, M.D.C. and P.A.; data curation, M.D.C. and P.A.; writing—original draft preparation, M.D.C. and P.A.; writing—review and editing, M.D.C. and P.A.; visualization, M.D.C. and P.A.; supervision, M.D.C. and P.A.; project administration, M.D.C. and P.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data were extracted from the online corpora cited in the text and listed in the References section below.

Acknowledgments: We want to thank the two anonymous reviewers for their comments and suggestions.

Conflicts of Interest: The authors declare no conflict of interest.

Notes

- ¹ Available online: <https://www.priberam.pt/dlpo/passo> (accessed on 18 December 2017).
- ² As is normal practice in semantics, we use the term eventuality to encompass both events and states (Bach 1981).
- ³ To the best of our knowledge, the studies that mention *ao passo que* as a contrastive connective present data from European Portuguese. Hence, they focus on the syntactic properties of the connective in this variety. However, the dictionaries we consulted mention the same meanings for both European and Brazilian Portuguese. Therefore, we are not assuming a difference between these varieties in our study. Our diachronic data include examples of both varieties.
- ⁴ This issue and some counterexamples to this claim are discussed below.
- ⁵ Mutual exclusivity explains the diachronic path from exceptive words meaning ‘outside of’ or ‘except’ to contrastive meanings (cf. also the contrastive meaning of EP *longe de, fora*).
- ⁶ We will not pursue this analysis in depth, since our focus is on the contrastive meaning of *ao passo que*. For an analysis of the semantics of temporal correlatives, and the cross-linguistic connections between temporal connectives, relative clauses, and equative constructions, see (Bhatt and Lipták 2009).
- ⁷ Barreto (1999, pp. 406–11) briefly describes the grammaticalization of both expressions as having a similar structure and meaning.
- ⁸ An example of *passo* as an adverb in medieval Portuguese, meaning ‘slowly,’ is given in (i):
 - (i) Entam se chegou a elle mais que ante e pos mão em elle muj passo pollo spertar (*Demanda do Santo Graal*, 113, Toledo Neto 2012–2015). ‘Then [he/she] approached him more than before and touched him very slowly to wake him up.’
 We are grateful to an anonymous reviewer for this example.
- ⁹ We analyze this *que* as a complementizer, part of the complex connective. The key syntactic evidence here is the fact that, in general terms, *que* does not refer back to the N *passo* as there is no wh-gap; for instance, we can see that in (1) *O Rui gosta do café cheio, ao passo que o Zé gosta dele curto*, the syntax is not *Zé gosta dele curto a esse passo*.
- ¹⁰ In any case, we must keep in mind that semantic compositionality and syntactic analyzability are different (Bybee 2010, p. 45).

References

- Abraham, Werner. 1979. But. *Studia Linguistica* 33: 89–119. [CrossRef]
- Amaral, Patrícia, and Manuel Delicado Cantero. 2019. Not a fact: A synchronic analysis of *el hecho de* and *o facto de*. *Probus* 31: 1–27. [CrossRef]
- Amaral, Patrícia, and Manuel Delicado Cantero. 2022. *Noun-based Constructions in the History of Portuguese and Spanish*. Oxford: Oxford University Press.
- Bach, Emmon. 1981. On time, tense, and aspects: An essay in English meta-physics. In *Radical Pragmatics*. Edited by Peter Cole. New York: Academic Press, pp. 63–81.
- Barreto, Therezinha Maria Mello. 1999. Gramaticalização das conjunções na história do português. Ph.D. thesis, Universidade Federal da Bahia, Salvador, Brasil.
- Bertinetto, Pier Marco, and Mario Squartini. 1995. An attempt at defining the class of “gradual completion verbs”. In *Temporal Reference, Aspect and Actionality*. Torino: Rosenberg and Sellier, pp. 11–26.
- Bhatt, Rajesh, and Anikó Lipták. 2009. Matching effects in the temporal and locative domains. In *Correlatives Cross-Linguistically*. Edited by Anikó Lipták. Amsterdam: John Benjamins, pp. 343–72.
- Brems, Lieselotte, and Kristin Davidse. 2010. Complex subordinators derived from noun complement clauses: Grammaticalization and paradigmaticity. *Acta Linguistica Hafniensia* 41: 101–16. [CrossRef]
- Brito, Ana Maria, and Gabriela Matos. 2003. Construções de graduação e comparação. In *Gramática da língua portuguesa*. Edited by Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário and Alina Villalva. Lisboa: Caminho, pp. 729–66.
- Bybee, Joan. 2010. *Language, Use and Cognition*. Cambridge: Cambridge University Press.
- Bybee, Joan L., Revere Perkins, and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect and Modality in the Languages of the World*. Chicago: The University of Chicago Press.
- CETEMPúblico. n.d. Available online: <https://www.linguateca.pt/CETEMPUBLICO> (accessed on 30 November 2023).
- Charnavel, Isabelle. 2015. Same, different and other as comparative adjectives. A uniform analysis based on French. *Lingua* 156: 129–74.
- Cifuentes Honrubia, José Luis. 2003. *Locuciones Prepositivas. Sobre la Gramaticalización Preposicional en Español*. Alicante: Universidad de Alicante.
- Corpas Pastor, Gloria. 1996. *Manual de Fraseología Española*. Madrid: Gredos.
- Costa, Ana Luísa. 2005. Aspectos sintácticos-semânticos de estruturas contrastivas. In *Actas do XX Encontro Nacional da Associação Portuguesa de Linguística*. Edited by Inês Duarte and I. Leiria. Lisboa: Associação Portuguesa de Linguística, pp. 495–507.
- Cunha, Celso Ferreira da, and Luis Filipe Lindley Cintra. 2017. *Nova gramática do português contemporâneo. 7ª Edição. De acordo com a nova ortografia*. Rio de Janeiro: Lexikon.

- Davies, M. 2006. *Corpus do português*. Available online: www.corpusdoportugues.org/hist-gen/ (accessed on 30 November 2023).
- 2008–2021. *Dicionário Priberam da Língua Portuguesa*. Available online: <https://dicionario.priberam.org/> (accessed on 13 October 2023).
- Dowty, David. 1979. *Word Meaning and Montague Grammar*. Dordrecht: Reidel.
- Duarte, Inês. 2003. Aspectos linguísticos da organização textual. In *Gramática da língua portuguesa*. Edited by Maria Helena Mira Mateus, Ana Maria Brito, Inês Duarte, Isabel Hub Faria, Sónia Frota, Gabriela Matos, Fátima Oliveira, Marina Vigário and Alina Villalva. Lisboa: Caminho, pp. 85–123.
- Eckardt, Regine. 2006. *Meaning Change in Grammaticalization*. Oxford: Oxford University Press.
- Fagard, Benjamin, José Pinto de Lima, Dejan Stosic, and Elena Smirnova. 2020. Introduction: Complex adpositions and complex nominal relators. In *Complex Adpositions in European Languages. A Micro-Typological Approach to Complex Nominal Relators*. Edited by Benjamin Fagard, José Pinto de Lima, Dejan Stosic and Elena Smirnova. Berlin: De Gruyter, pp. 1–30.
- Fábregas, Antonio. 2007. (Axial) Parts and Wholes. In *Tromsø Working Papers on Language & Linguistics: Nordlyd 34.2, Special Issue on Space, Motion, and Result*. Edited by Monika Băsić, Marina Pantcheva, Minjeong Son and Peter Svenonius. Tromsø: CASTL, pp. 1–32.
- Foolen, Ad. 1991. Polyfunctionality and the semantics of adversative conjunctions. *Multilingua* 10: 79–92.
- Hopper, Paul J. 1991. On some principles of grammaticization. In *Approaches to Grammaticalization. Volume I. Theoretical and Methodological Issues*. Edited by Elizabeth Closs Traugott and Bernd Heine. Amsterdam-Philadelphia: John Benjamins, pp. 17–35.
- Izutsu, Mitsuko Narita. 2008. Contrast, concessive, and corrective: Toward a comprehensive study of opposition relations. *Journal of Pragmatics* 40: 646–75. [CrossRef]
- Jiménez Juliá, Tomás. 2017. Notas sobre locuciones y construcciones sintácticas. *Cuadernos AISPL: Estudios de Lenguas y Literaturas Hispánicas* 10: 91–112.
- König, Ekkehard. 1985. On the history of concessive connectives in English. Diachronic and synchronic evidence. *Lingua* 66: 1–19. [CrossRef]
- Lakoff, Robin. 1971. If's, and's, and but's about conjunction. In *Studies in Linguistic Semantics*. Edited by Charles J. Fillmore and D. Terence Langendoen. New York: Holt, Rhinehart & Winston, pp. 114–49.
- Lobo, Maria. 2013. Subordinação adverbial. In *Gramática do Português*. Edited by Eduardo Buzaglo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Maria Antónia Coelho da Mota, Luísa Segura and Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, pp. 1981–2060.
- Lopes, Ana Cristina M. 2001. Elementos para uma análise semântica das construções com enquanto. In *Maria Helena Mira Mateus and Clara Nunes Correia. Saberes no tempo: Homenagem a Maria Henriqueta Costa Campos*. Lisboa: Colibri, pp. 371–80.
- Luís, Ana, and Georg Kaiser. 2016. Clitic pronouns: Phonology, Morphology, and Syntax. In *The Handbook of Portuguese Linguistics*. Edited by W. Leo Wetzels, João Costa and Sergio Menezzi. Malden: Wiley, pp. 210–33.
- Malchukov, Andrej. 2004. Towards a semantic typology of adversative and contrast marking. *Journal of Semantics* 21: 177–98. [CrossRef]
- Mann, William C., and Sandra A. Thompson. 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8: 243–81. [CrossRef]
- Matushansky, Ora, and Joost Zwarts. 2019. Tops and bottoms: Axial nominals as weak definites. In *Proceedings of the 36th West Coast Conference on Formal Linguistics*. Edited by Richard Stockwell, Maura O'Leary, Zhongshi Xu and Z. L. Zhou. Somerville: Cascadilla Proceedings Project, pp. 270–80.
- Matushansky, Ora, and Joost Zwarts. 2021. Parameters of Nominal Deficiency in Complex Prepositions. Available online: <https://hal.science/hal-03511560/document> (accessed on 13 October 2023).
- Mauri, Caterina. 2008. *Coordination Relations in the Languages of Europe and Beyond*. Berlin: Mouton De Gruyter.
- Montoro del Arco, Esteban T. 2006. *Teoría fraseológica de las Locuciones Particulares. Las Locuciones Prepositivas, Conjuntivas y Marcadoras en español*. Frankfurt am Main: Peter Lang.
- Pavón Lucero, María Victoria. 2003. *Sintaxis de Las Partículas*. Madrid: Visor.
- Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Pearson Longman.
- Rocha, Paulo, and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *Actas do V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR'2000)*. Edited by Maria das Graças Volpe Nunes. São Paulo: ICMC/USP, pp. 131–40.
- Romeu Fernández, Juan. 2014. *Cartografía mínima de las Construcciones Espaciales*. Ph.D. thesis, Universidad Complutense de Madrid, Madrid, Spain.
- Rudolph, Elizabeth. 1996. *Contrast. Adversative and Concessive Expressions on Sentence and Text Level*. Berlin: Mouton De Gruyter.
- Ruiz Gurillo, Leonor. 1997. *Aspectos de fraseología teórica española. Anejo XXIV de Cuadernos de Filología*; Valencia: Universitat de València. Available online: https://www.academia.edu/4078186/Ruiz_Gurillo_L_1997_Aspectos_de_fraseolog%C3%ADa_te%C3%B3rica_espa%C3%B1ola_Valencia_Universitat (accessed on 11 December 2023).
- Svenonius, Peter. 2006. The emergence of axial parts. *Nordlyd: Tromsø Working Papers in Linguistics* 33: 49–77. [CrossRef] [PubMed]
- Svenonius, Peter. 2010. Spatial P in English. In *Mapping Spatial PPs. The Cartography of Syntactic Structures*. Edited by Guglielmo Cinque and Luigi Rizzi. Oxford: Oxford University Press, pp. 127–60.
- Toledo Neto, Sílvio. 2012–2015. *Edição da Demanda do Santo Graal*. Available online: <http://alfclul.clul.ul.pt/wochwel/index.html> (accessed on 30 November 2023).

- Tornel Sala, José Luis. 2000. *Gramaticalización y vectores lingüísticos. Las perífrasis verbo-nominales*. Alicante: Editorial Club Universitario.
- Torres Cacoullos, Rena. 2006. Relative frequency in the grammaticization of collocations: Nominal to concessive a pesar de. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*. Edited by Timothy L. Face and Carol A. Klee. Somerville: Cascadilla Proceedings Project, pp. 37–49.
- Traugott, Elizabeth C., and Ekkehard König. 1991. The semantics-pragmatics of grammaticalization revisited. In *Approaches to Grammaticalization*. I vols, Theoretical and Methodological Issues. Edited by Elizabeth C. Traugott and Bernd Heine. Amsterdam and Philadelphia: John Benjamins, pp. 189–218.
- Traugott, Elizabeth C., and Graeme Trousdale. 2013. *Constructionalization and Constructional Changes*. Oxford: Oxford University Press.
- Traugott, Elizabeth C., and Richard Dasher. 2002. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.
- Umbach, Carla. 2005. Contrast and information structure: A focus-based analysis of but. *Linguistics* 43: 207–32. [CrossRef]
- Zuluaga, Alberto. 1980. *Introducción al estudio de las expresiones fijas*. Frankfurt: Peter Lang.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

European Portuguese *lá*: Use-Conditional Meaning and Pragmaticalization

Lukas Müller

Department of Romance Studies, University of Cologne, Albertus-Magnus-Platz, 50923 Köln, Germany; lukas.mueller@uni-koeln.de

Abstract: This study focusses on non-adverbial uses of *lá* in European Portuguese, whose exact meaning contribution still remains an open research question. Applying a multidimensional semantics framework, the central claim is that non-adverbial uses of *lá* represent use-conditional items. Passing the standard tests suggested in the literature, they thus do not contribute to the truth conditions of an utterance but specify particular use conditions. It is argued that they are felicitously used if a speaker wants to convey illocutionary modification, which pragmatically leads to mitigation or reinforcement effects. Diachronically, substantiated by historical data from the *Corpus do Português*, use-conditional *lá* is argued to be a product of a pragmaticalization process that led to so-called pragmatic fission at some point, i.e., to the polysemy of two synchronically available configurations of a truth-conditional and a use-conditional *lá*.

Keywords: illocutionary modification; multidimensional semantics; expressivity; modal particle; pragmaticalization

1. Introduction

In European Portuguese, *lá* is primarily known as an adverbial that serves to designate a location relative to the speaker's *origo*, i.e., as a spatial deictic expression, as illustrated in the examples in (1).

- (1) a. *Foi para a cidade e por lá ficou.*¹
'He went downtown and there he stayed.'
b. *Compraram um terreno e lá edificaram a sua casa.*
'They bought a plot of land and built their house there.'

In particular, it signals a location which, according to the standard norm of European Portuguese, is proximate to neither the speaker nor the addressee (Costa and Martins 2010, p. 61). In addition to this basic function of *lá*, there are non-adverbial uses (cf. e.g., Duarte 1989), as exemplified in (2), which have seen growing attention over the past years.

- (2) [À mesa, diz a mãe para o filho:] (Pereira 2009, p. 87)
'[At the table, says the mother to the son:]'
a. *Acaba lá de comer a sopa!*
'(Just) eat the soup!'
b. *Dá-me lá a colher que eu ajudo-te.*
'(Just) give me the spoon, I will help you.'

While it is clear that in these examples, *lá* is not meant to indicate a spatial location but something else, its exact meaning contribution to the entire utterance remains an open research question. Previous descriptions are mainly based on detailed paraphrases of individual examples to cover that *lá* adds certain nuances to the overall meaning of an utterance, which is commonly verified by checking the effect of its omission, i.e., by means

Citation: Müller, Lukas. 2024. European Portuguese *lá*: Use-Conditional Meaning and Pragmaticalization. *Languages* 9: 189. <https://doi.org/10.3390/languages9060189>

Academic Editor: Patrícia Amaral

Received: 7 March 2024

Revised: 10 May 2024

Accepted: 13 May 2024

Published: 21 May 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

of minimal pairs. The general assumption is that non-adverbial *lá* carries a highly context-sensitive, speaker-oriented meaning. Although not stated explicitly, this suggests the status of a conversational implicature, which is to be resolved taking into account the individual context that *lá* is uttered in. In this vein, Pereira (2009) assigns *lá* in (2-a) with the function of affecting the directive speech act ('Eat the soup!') to signal impatience. Elsewhere, *lá* has been described as a marker of emphasis, a marker of cognitive distance and as a so-called emotional deictic expression (cf. Marques and Duarte 2017). Meisnitzer (2012) and (Meisnitzer and Müller 2024) assign non-adverbial *lá* with the status of a modal particle, very much in the tradition of Franco (1988, 1991), which means that they assume a conventionalized meaning entry, which goes against the assumption of a conversational implicature. However, *lá* is not paid detailed attention to in these studies.

What is needed for a better understanding of the meaning of *lá* is an elaborated account of its functional spectrum that can be encountered in synchrony, as well as a careful delimitation of its conventional semantics and its pragmatic potential. Therefore, the goal of this contribution is to describe its meaning building on the framework of multidimensional semantics as discussed in Gutzmann (2013, 2015), a successor of Potts (2005)'s work on the logics of conventional implicatures. As we will see, this approach provides a range of insights that serve to describe the conventional semantics of non-adverbial *lá* in a unified framework based on use conditions. In addition, this framework provides a way to account for the diachronic dimension underlying *lá*, suggesting quite a regular pragmaticalization path, which has not been studied yet to my knowledge.

The article is organized as follows. In Section 2, I briefly introduce the main characteristics of the framework of multidimensional semantics featuring the fundamental distinction between truth-conditional and use-conditional meaning. Applying standard tests from the literature, I then identify a truth-conditional *lá*_{TC}, as well as a use-conditional *lá*_{UC} in Section 3. Building on the assumptions developed by Davis and Gutzmann (2015) concerning the semantics of pragmaticalization of use-conditional items, I extend the analysis to the diachronic dimension in Section 4 to investigate the historical development of *lá*_{UC} out of *lá*_{TC}. The Section 5 concludes the study.

2. Hybrid Semantics: Truth- and Use-Conditional Meaning

The analysis builds on the multidimensional semantics framework in the sense of Gutzmann (2013, 2015) and the crucial idea that the conventional meaning of a linguistic expression disposes of a truth-conditional, as well as a use-conditional dimension. This approach is heavily influenced by work on expressivity by Cruse (1986); Kaplan (1999); (Recanati 2004, p. 447); and Potts (2005, 2007), that also assume multidimensionality (cf. e.g., "two modes of expression" in Kaplan 1999). Accordingly, there is both a descriptive/propositional/truth-conditional as well as an expressive dimension in the semantics of a linguistic expression, which explains the term multidimensional semantics. Basically, the truth-conditional dimension concerns descriptive meaning contributing to truth values in the traditional sense, whereas expressive meaning is about non-truth-conditional, subjective, speaker-oriented, evaluative meaning. Expressive items may be used felicitously and infelicitously, which means that they do not add anything to the truth-conditions of an expression. The prime example for expressives are slurs (3).

- (3) Example discussed in (Gutzmann 2013, p. 8).
- a. *Lessing was a German.*
 - b. *Lessing was a Kraut.*

Both examples display the same descriptive, truth-conditional meaning, which is traditionally understood to define the conditions that need to be fulfilled so that a sentence can be rendered true. In most general terms, this is in line with Wittgenstein (1922)'s *Bildtheorie des Satzes* and the idea that understanding a sentence means to know what the case is if it is true. Accordingly, speakers have to know the conditions that (3-a) and (3-b) impose on the world for judging it true or false. Still, the minimal pair clearly displays a

contrast in meaning, which goes beyond the truth-conditional dimension but is captured by the expressive component contained in the slur *Kraut*, which might be paraphrased more or less as “Generally, I do not like Germans”. If a speaker chooses to linguistically commit to this negative evaluation conventionally denoted by the slur *Kraut*, they utter (3-b) felicitously, regardless of the question of whether it is a true or false statement.

- (4) Multidimensional denotations for (3-b) (modified example based on Davis and Gutzmann 2015, p. 200).
 - a. Mapping to a set of worlds
“Lessing was a Kraut” is **true** if Lessing was a German.
 - b. Mapping to a set of contexts
“Lessing was a Kraut” is **felicitously used** if the speaker has a negative attitude towards Germans.

The multidimensional system gives rise to four possible combinations concerning the truth and use conditions of a sentence, i.e., $\langle 1, \checkmark \rangle$, $\langle 1, \cancel{\checkmark} \rangle$, $\langle 0, \checkmark \rangle$, $\langle 0, \cancel{\checkmark} \rangle$. While expressives in the narrow sense are mostly associated with negative or positive evaluations, Gutzmann (2015)’s notion of use-conditional meaning covers a broader range of kinds of subjective, speaker-oriented meaning, which turn out to share many properties with expressives (cf. also the discussion in Amaral 2018, p. 330f.), such as particles (e.g., German modal particles), certain intonation patterns, or certain morphosyntactic constructions of mirativity or topicalization. The multidimensionality of these examples can be illustrated informally by the fraction notation shown in (5).

- (5) Examples of use-conditional content complementing truth conditions and the fraction notation (Gutzmann 2013).
 - a. Conventional meaning of a linguistic expression = $\frac{\text{use conditions}}{\text{truth conditions}}$.
 - b. *Lessing was a Kraut* = $\frac{\text{Generally, I do not like Germans}}{\text{Lessing was a German}}$
 - c. *Hans-i* (German) = $\frac{\text{Familiar relationship between the speaker and Hans}}{\text{Hans}}$.
 - d. *John, Mary loves.* = $\frac{\text{The speaker’s mental representation of John is active}}{\text{Mary loves John}}$.
 - e. *Carl did finish their book.* = $\frac{\text{It is true that Carl finished their book}}{\text{Carl finished their book}}$.
 - f. *How tall Michael is!* = $\frac{\text{It is unexpected how tall Michael is}}{\emptyset}$.
 - g. *Ouch!* = $\frac{\text{I’m in pain}}{\emptyset}$.

In (5-c), it is the social meaning of a nickname that can be used felicitously or infelicitously without changing its reference; (5-d) suggests linking the information structure to use conditions based on the idea that the marked word order does not describe but expresses different things, e.g., that the speaker’s mental representation of John is active in the respective topicalization. In (5-e), the idea is that the verum focus associated with “did” may be used (in)felicitously without interfering with the sentence’s truth value. Finally, (5-f) shows that there are also expressions that completely lack a truth-conditional component, meaning that they are not mixed but purely use-conditional items, which also applies to (5-g).

In the tradition of Potts (2005)’s seminal book, there has been a lot of subsequent work on the main properties of expressives that led to the development of linguistic tests for expressive items. Partly, these have been controversially discussed (cf. overview in Amaral 2018, p. 327ff. and the second issue of volume 33 of the journal *Theoretical Linguistics*). Still, they provide a solid base to test for expressivity and, in fact, have been shown to apply to use-conditional meanings similarly (cf. Gutzmann 2013, p. 31ff.). Among the main properties of use-conditional items are independence, nondisplaceability, immediacy, descriptive ineffability, and perspective dependence. I will turn back to these in more detail in the analysis of $l\acute{a}_{UC}$ in Section 3.2.

3. Two Semantic Configurations of European Portuguese *lá*

In this section, European Portuguese *lá* is described based on the multidimensional semantics framework introduced above. I argue that there are two configurations, namely a purely truth-conditional $lá_{TC}$ and a purely use-conditional $lá_{UC}$. In Section 3.3, I comment on whether there might actually also be a third type: a mixed type denoting both truth- and use-conditional content at the same time. In fact, there is a candidate, namely a particular use of *lá* which in the literature is often referred to as a marker of meta-linguistic negation (cf. Martins 2014, 2021; Schwenter 2016). Although the detailed study of metalinguistic negation with *lá* within a multidimensional framework remains for future work, I develop the hypothesis in Section 3.3 that it represents just another pragmatic outcome of $lá_{UC}$ as discussed in the following.

3.1. The Truth-Conditional Configuration: $lá_{TC}$ with Deictic Meaning

The truth-conditional $lá_{TC}$ arguably constitutes the basic meaning of *lá*. The term basic intends to account for that (i) it is commonly mentioned in the first place in lists of uses of *lá*, as, e.g., in the *Dicionário da Língua Portuguesa* by the *Academia das Ciências de Lisboa* (cf. footnote 1), and (ii) non-adverbial *lá* is commonly understood as derived; i.e., it has a somewhat secondary meaning in the previous literature, as we will see below.

- (6) a. *Foi para a cidade e por lá ficou.*
'He went downtown and there he stayed.'
- b. *Compraram um terreno e lá edificaram a sua casa.*
'They bought a plot of land and built their house there.'

In these examples, $lá_{TC}$ functions as an adverbial to designate a location relative to the speaker *origo*. In particular, it signals a location which, according to the standard norm, is proximate to neither the speaker nor the addressee (Costa and Martins 2010, p. 61). Close synonyms of *lá* are *aí* and *ali*, featuring subtle differences regarding the parameters of proximity between speaker and addressee.²

- (7) Spatical deictic expressions (Costa and Martins 2010, p. 61).
 - a. *aqui/cá* 'here' [+ close to speaker].
 - b. *aí* 'there' [– close to speaker, + close to addressee].
 - c. *ali/lá* 'there' [– close to speaker and addressee].

In (6), $lá_{TC}$ clearly contributes to the conditions that render the utterances true, which are evaluated based on the agreement with the extralinguistic world's facts. If in (6-a), they went downtown but then returned, (6-a) is false. Likewise, the location of the construction of the house is crucial for deriving the truth value of the whole expression in (6-b). Applying the fraction notation of the framework of hybrid semantics introduced in the previous section, we can informally describe the meaning of $lá_{TC}$ by means of the formula indicated in (8), which shows a non-occupied use-conditional dimension.³

- (8) Multidimensional semantics of $lá_{TC}$ in the fraction notation.
 - a. $lá_{TC} = \frac{\emptyset}{(\text{over}) \text{there}}$.
 - b. *Foi para a cidade e por lá ficou.* = $\frac{\emptyset}{\text{she went downtown and there she stayed}}$.

3.2. The Use-Conditional Configuration: $lá_{UC}$ and Illocutionary Modification

Although non adverbial-uses of *lá* have not been identified as expressing use-conditional content in previous work, their functions have been described in several studies. A first crucial aspect concerns diamesic variation, namely that there is a restriction to oral, informal registers, an observation already made by Ali ([1930] 1971) and maintained in recent studies. As for the meaning contribution, (Duarte 2010, p. 180) speaks of *marqueurs d'aiguillage communicatif* ('communicative referral markers') but also alludes to that they may count as modal particles in the tradition of Franco (1991) (cf. also Meisnitzer 2012;

Meisnitzer and Müller 2024). Further, she qualifies their type of meaning as either of a conventional or a generalized conversational implicature in the sense of Grice (1975). The list of functions associated with *lá* contains distance in time/uncertainty (*la distance dans le temps/l'incertitude*), reinforcement of directives (*le renforcement de l'injonction*), attenuation, and fulfillment of expectations (*l'accomplissement des attentes*). In Marques and Duarte (2015), this list is reduced to the main functions of mitigation and reinforcement (strengthening) featuring specific interactions with particular speech acts and potentially further pragmatic factors.

The present analysis builds on these assumptions, though shifting from pragmatics to semantics, claiming that *lá_{UC}* carries the conventional use-conditional meaning of illocutionary modification, as illustrated in the fraction in (9).

$$(9) \quad \mathit{l\grave{a}}_{TC} = \frac{\text{illocutionary modification}}{\emptyset}$$

I follow (Gutzmann 2015, p. 168) and the distinction of sentence mood on the semantic level (a reflection of syntactic sentence type) and illocutionary force on the pragmatic level, as discussed in (Grewendorf and Zaefferer 1991, p. 270):⁴

	<i>sentence mood</i>	<i>illocutionary force</i>
	declarative	assertion
(10)	interrogative	question
	imperative	directive
	exclamative	exclamation

In the tradition of speech act theory, illocutionary force thus concerns the pragmatic level. This means that *lá_{UC}* operates right at the semantics–pragmatics interface in that it carries the conventional semantics of modifying on the pragmatic level. In line with (Hengeveld 2004, p. 1192), illocutionary modification does not however concern the identification of a certain speech act. Put differently, *lá_{UC}* does not conventionally contribute to the creation of indirect speech acts. Instead, illocutionary modification is part of a general strategy serving to “reinforce or mitigate the force of the speech act”. In particular, this means that assertions, questions, directives, and exclamations can be reinforced or mitigated in order to manipulate their degree of assertivity, directivity, etc. Although the concepts of mitigation and reinforcement might be objected to for their pre-theoretical and intuitive nature (cf. Schwenter 2003, p. 1026), they are broadly used in research, enabling onomasiological comparisons in a cross-linguistic perspective (cf. also the discussion in Favaro 2023, p. 184). In this vein, there has been extensive work on the range of linguistic strategies associated with the mitigation (sometimes also labelled as attenuation, downtoning) and reinforcement (also boosting, strengthening) of speech acts (Holmes 1984; Sbisà 2001; Schneider 2010). According to (Holmes 1984, p. 348), speakers choose to attenuate or boost certain speech acts either to express their “attitude to the content of the proposition” or to express their attitude “to the addressee in the context of utterance”, serving the higher goal of increasing either solidarity or social distance between speaker and hearer, very much in the spirit of politeness theory (cf. Brown and Levinson 1978). Note that this comes quite close to descriptions of expressivity, where “a speaker can establish a relation of intimacy with a hearer merely by choosing one lexical item rather than another in the course of a conversation” (Cruse 1986, p. 285). Although cross-linguistic research on illocutionary modification has often focussed rather on strategies of mitigation than on reinforcement, perhaps due to the “negative-politeness cultures” prevalent in the anglophone academic community (cf. Holmes 1984, p. 348), both are symmetrical and “of the same nature, since both are to be explained as variations in the illocutionary effects of speech acts” (Sbisà 2001, p. 1811). Put differently, they represent two sides of a coin, two pragmatic outcomes that trace back to the basic semantic instruction of illocutionary modification.

In terms of descriptive data for *lá_{UC}*, Pereira (2009) provides an elaborated list of examples together with descriptions of the situations and paraphrases of the meaning effect

as verified by minimal pairs manipulating the presence of *lá*. Starting with imperatives (11), it is illustrated that $lá_{UC}$'s illocutionary modification may swing towards both reinforcement and mitigation.

- (11) $lá_{UC}$ in imperatives (Examples from Pereira 2009, p. 87).
- a. [À mesa, diz a mãe para o filho:]
 - (i) *Acaba lá de comer a sopa!*
'(Just) eat the soup!'
~ Reinforcement of directive speech act, signaling impatience.
 - (ii) *Dá-me lá a colher que eu ajudo-te.*
'(Just) pass me the spoon, I will help you.'
~ Mitigation to turn an order into a request.
 - (iii) *Ó filho, vê lá se comes a sopa para seres um homem.*
'Son, make sure you eat enough soup to be a man'
~ Reinforcement (with a slight threat.)

Since one could think of contexts leading to opposite effect, respectively, it becomes clear that it is a pragmatic issue how $lá_{UC}$'s illocutionary modification is exactly to be resolved. In this regard, (Marques and Duarte 2015, p. 123) point out that intonation also impacts whether there is an effect of attenuation or reinforcement in directives (without identifying typical intonational patterns). Furthermore, they identify politeness strategies in the sense of Leech (1983) as an underlying motivation for using $lá_{UC}$ in directives. This is in line with the general observation that politeness is often mentioned as one (but not the only) reason for linguistic attenuation, as it contributes to efficiency in terms of achieving communicative goals (cf. Briz and Albelda 2013, p. 292). Using the fraction notation again, (11-a-i) can be understood as shown in (12). The perfectly grammatical and acceptable *Acaba de comer a sopa!*, conveying a directive speech act, is combined with the illocutionary modifier $lá_{UC}$. This composition outputs a multidimensional expression that is felicitously used if and only if the speaker wants to manipulate the degree of directivity. In the given context, it is most likely pragmatically resolved as an impatient reinforcement.

$$(12) \quad \textit{Acaba lá de comer a sopa!} = \frac{\emptyset}{\textit{acaba de comer a sopa!}} + \frac{\textit{illocutionary modification}}{\emptyset}$$

$$= \frac{\textit{illocutionary modification}}{\textit{acaba de comer a sopa!}}$$

With regard to declaratives, Marques and Duarte (2015) refer to mitigation (they use the term attenuation) that leads to the effect of epistemic distance or uncertainty and potentially to a negative evaluation of the underlying proposition for reasons of self-protection. In the same direction, Marques and Duarte (2017) refer to the concept of emotional deixis as derived based on the original spatial deictic item (cf. also Lopes 2019 for emotional deixis as a broader category, e.g., also to be found with personal deictics).

- (13) $lá_{UC}$ in declaratives (Pereira 2009, p. 91).
- a. *Tem estado muito doente, mas ontem lá se sentou à mesa.*
'He's been very ill, but yesterday he sat at the table.'
 - b. *O João tem estado muito doente, mas ontem, a muito custo ou com muito esforço, conseguiu tomar a sua refeição à mesa, o que indicia que o seu estado de saúde já não é tão grave como antes.*
'João has been very ill, but yesterday, with great difficulty or effort, he managed to eat their meal at the table, which indicates that their state of health is no longer as serious as before.'
 - c. *Só acabarei este trabalho lá para Setembro.*
'I will not finish this job until (approximately) September.'

Pereira (2009) provides (13-b) as a paraphrase of (13-a), indicating that *lá* contributes to a contrast, since a very sick person is commonly not expected to sit down at a table. The

mitigation effect of $lá_{UC}$ thus serves to weaken the assertion of p *tem estado muito doente ultimamente*. In terms of (Favaro 2023, p. 132), it “reduce[s] the level of assertivity”, i.e., the degree of the speaker’s commitment to what is conveyed by p (cf. also Shapiro 2020 on assertion and commitment). Mitigating the seriousness of the message, there is a flavor of “let us not worry too much, it will be alright”. While there might be a way to express the same meaning in (13-a) via certain intonational contours, it is lexically conveyed by $lá_{UC}$ in (13-b).

- (14) *Tem estado muito doente, mas ontem lá se sentou à mesa.*
 = $\frac{\emptyset}{\text{Tem estado muito doente, mas ontem sentou-se à mesa.}} + \frac{\text{illocutionary modification}}{\emptyset}$
 = $\frac{\text{illocutionary modification}}{\text{Tem estado muito doente, mas ontem sentou-se à mesa.}}$

Another note on this example concerns the presence of *mas*, which clearly acts as a trigger of a conventional implicature in the classical sense of Grice (1975), also indicating some kind of contrast between the two propositions p and q . Although leading to a related meaning effect, *mas* and $lá_{UC}$ act separately.

Concerning (13-c), Pereira (2009) argues that *lá* is not unequivocally disjoined from a local (i.e., temporal) interpretation, which underlines the polysemy of *lá* and the associated high degree of ambiguity obtained in many examples also mentioned by (Marques and Duarte 2015, p. 121). However, Pereira ascribes *lá* in (13-c) a modalizing meaning that conveys the speaker’s uncertainty about the date, which thus turns to be perceived as imprecise, resulting again in a reduced level of assertivity. Crucially, this effect is lost when omitting $lá_{UC}$. In fact, Marques and Duarte (2017) claim that this mitigation effect in declaratives oftentimes coincides with indefinite NPs or other linguistic markers of vagueness, such as hesitation, creating epistemic distance, both illustrated in (15).

- (15) *E depois e depois [hesitação], pronto, uns acontecimentos lá, que [hesitação] não gosto muito de coisas trocadas [...].*
 ‘And then there was [hesitation], a few events there, [hesitation], I do not really like things being changed.’

In questions, the data suggest that $lá_{UC}$ ’s use condition of illocutionary modification uniquely creates reinforcement that leads, e.g., to exclamative-like questions (16-a). Note that this might suggest that $lá_{UC}$ has the potential to create indirect speech acts, e.g., to turn a question into an order (16-b). However, omitting $lá_{UC}$ in the examples in (16) verifies that this is not the case, as indirect speech acts are also easily created without it. This means that $lá_{UC}$ bears the potential to contribute to the creation of an indirect speech act together with a range of other factors, e.g. intonation. However, its function is not to specify but to mitigate or reinforce the illocutionary force. Regarding the missing mitigation, it remains unclear if these actually match with questions at all (similarly to exclamatives, cf. below), which might explain the absence of examples.

- (16) $lá_{UC}$ in questions (Pereira 2009, p. 93).
 a. *Então, como vai lá isso?!*
 ‘So, how’s it going?’
 b. *Podes lá parar com isso?!*
 ‘Will you stop it?!’
 c. *E que tem lá isso?*
 ‘What is up with that?’

Last but not least, Pereira (2009) discusses exclamatives and idiomatic expressions. Since exclamatives do not carry truth-conditional meaning per definitionem, they obviously represent prime examples of expressive/use-conditional meaning. In this sense, $lá_{UC}$ acts again as an item of reinforcement, as the idea of mitigating an exclamation seems to be somewhat contradictory. Considering (17-b), Pereira (2009) ascribes $lá_{UC}$ the emotional expression of surprise and admiration for what is perceived.

- (17) Exclamatives (Pereira 2009, p. 89).
- a. *Vá lá, podia ter sido pior!*
'Come on, it could have been worse!'
 - b. *Eh lá! Que catástrofe!*
'Hey! What a catastrophe!'
 - (i) **Lá! Que catástrofe!* (constructed by LM)
 - (ii) *Eh! Que catástrofe!* (constructed by LM)

Pereira further analyzes *eh* as reinforcing *lá* (instead of the other way round). Given that *lá* alone does not represent a well-formed exclamation though (17-b-i), which *eh* does (17-b-ii), it rather seems that it is *eh* that sets up the exclamation, which then is boosted by $lá_{UC}$, as illustrated in (18).

$$(18) \quad Eh \text{ lá!} = \frac{\text{emotional involvement}}{\emptyset} + \frac{\text{illocutionary modification}}{\emptyset} \\ = \frac{\text{emotional involvement} + \text{illocutionary modification}}{\emptyset}$$

Finally, there are idiomatic constructions that typically host *lá*. They share that they pass *lá* to highly subjective, speaker-oriented constructions, e.g., *seja quem for* (19-b), again with a function of illocutionary modification, as shown in (20). Given that these are easily amenable to expressivity per se, the conventionalization of $lá_{UC}$ at some point in diachrony is plausible and should be reconstructable based on the account discussed here and particularly in Section 4, which I leave for future study. Concerning the type of illocutionary modification, they can be seen as declaratives whose degree of assertivity is boosted.

- (19) Idiomatic constructions (Pereira 2009, p. 94).
- a. *Já lá vai o tempo em a família se reunia ao serão.*
'Gone are the days when the family would get together in the evening.'
 - b. *Seja lá quem for deve ser castigado pelo acto que cometeu*
'Whoever it is must be punished for the act they have committed.'
 - c. *Seja lá como for deve ser castigado pelo acto que cometeu.*
'Whatever the case, he should be punished for the act he committed.'
 - d. *Sabe lá Deus se chegamos amanhã.*
'God knows if we'll arrive tomorrow.'
 - e. *Não penses mais no que te aconteceu. O que lá vai lá vai.*
'Do not think about what happened to you anymore. What goes around comes around.'

$$(20) \quad \textit{Seja lá quem for} = \frac{\emptyset}{\textit{seja quem for}} + \frac{\textit{illocutionary modification}}{\emptyset} = \frac{\textit{illocutionary modification}}{\textit{seja quem for}}$$

Summing up, $lá_{UC}$ is felicitously used if a speaker wants to express illocutionary modification that is pragmatically resolved in terms of mitigation or reinforcement. This holds across several sentence moods, as summarized in Table 1. This generalization is restricted though to the extent that mitigation turns out to be somewhat contradictory to the meaning of the exclamative speech act. It remains furthermore unclear if interrogatives can be mitigated by means of $lá_{UC}$.

Table 1. Illocutionary modification with $lá_{UC}$ and sentence mood.

	Mitigation	Reinforcement
imperative	✓	✓
declarative	✓	✓
interrogative	?	✓
exclamative	■	✓

Question mark means “no example encountered” and black square means “contradictory”.

Testing for Common Properties of Use-Conditional Items

There are some typical properties of expressives suggested by (Potts 2007, p. 166f.) that help identify truth- and use-conditional meaning and that have been discussed a lot subsequently (cf. e.g., Gutzmann 2013, p. 31 and the overview provided in Amaral 2018). In line with the claim that non-adverbial uses of *lá* carry use-conditional meaning, they should pass the typical tests listed in (21).

- (21) Properties of expressives proposed by (Potts 2007) (brackets indicating partly problematic criteria, cf. Gutzmann 2013; Amaral 2018).
- a. Independence.
 - b. Nondisplaceability.
 - c. Perspective dependence.
 - d. (Descriptive ineffability).
 - e. Immediacy.
 - f. (Repeatability).

To start with, the property of independence covers that expressive content contributes a meaning dimension separate from the regular truth-conditional content, meaning that the descriptive, truth-conditional dimension imposes truth conditions, whereas the use-conditional dimension imposes felicity conditions that render an utterance felicitous or infelicitous instead of true or false. This can be tested by comparing a sentence with and without a purported use-conditional item and thus reflects the procedure in Pereira (2009). Omitting *lá_{UC}* should not affect the descriptive meaning of a sentence. Related to this, it should not be possible to negate *lá_{UC}* by ordinary negation (22-a-i), to directly deny it in dialogue (22-a-ii), nor to target it in an interrogative (22-a-iii). A way to object to content imposed by *lá_{UC}* is exemplified in (22-a-iv), where the *só que...* construction ('it is just that ...') provides an indirect way to refuse *lá_{UC}*. In line with Longhin-Thomazi (2003), *só que* represents a conjunctive periphrasis that adds to a preceding statement a new circumstance that breaks with some of the assumptions shared by the speakers, leading to meaning effects of refusal, surprise or counter argument. In (22-b), where the preceding directive lacks *lá_{UC}*, the *só que* construction is highly marked if not odd, given that it is unclear what exactly the objection is supposed to refer to.⁵

- (22) a. A: *Acaba lá de comer a sopa!* (with intensification reading).
'(Just) finish your soup!'
- (i) B: *#Não, mas já vou fazê-lo.*
'No, but I will do it.'
 - (ii) B: *#Já vou fazê-lo, mas não lá.*
'I will do it right away, but not there.'
 - (iii) B: *#Como/onde é que devo fazê-lo? Lá?*
'Where am I supposed to do it?'
 - (iv) B: *Está bem, só que vou fazê-lo com calma.*
'Alright, but I'm going to do it slowly.'
- b. A: *Acaba {ø} de comer a sopa!*
B: *?? Está bem, só que vou fazê-lo com calma.*
'A: (Just) finish your soup! B: Alright, but I'm going to do it slowly.'

The property of nondisplaceability states that expressives/use-conditional items predicate something of the utterance of the situation, meaning that they cannot be shifted away from the utterance situation. While descriptive, truth-conditional content can be easily dissociated from the situation of the actual utterance, e.g., by means of past or conditional operators, expressives are bound to the utterer and the time and place of the utterance (cf. Cruse 1986, p. 272). Yule (2006, p. 9) compares this immediacy of expressives to animal language, exemplified by a dog whose expressive interjection "Grrr" means "Grrr right now" and not "Grrr last night, over in the park". The property of nondisplaceability also

proves to be a good fit for the examples in (23), as the listed imperatives are strictly tied to the speakers *origo*. Trying to shift $lá_{UC}$ in a directive speech act to the past (23-a) or to enter it in the scope of a universal quantifier and the subjunctive mood (23-b) results in oddity. A repair strategy might be to force a reading, i.e., where $lá_{TC}$ adverbially modifies the location of the eating event.

- (23) a. *Eu lembro-me de quando me ligaste. Foi ontem ao meio dia quando eu pedi para o meu filho acabar [$*lá_{UC}/lá_{TC}$] a sopa.*
 ‘I remember when you called me. It was yesterday at noon when I asked my son to finish their soup.’
- b. *Sempre que peço ao meu filho que acabe [$*lá_{UC}/lá_{TC}$] de comer a sopa, ele come tudinho.*
 ‘Whenever I ask my son to finish their soup, he eats it all.’
- c. *Se alguém estivesse muito doente mas [$*lá_{UC}/lá_{TC}$] se sentasse à mesa, era surpreendente.*
 ‘If someone was very ill but sat down at the table, it was surprising.’

The property of perspective independence is quite similar to the nondisplaceability condition and states that expressive or use-conditional content is evaluated from a particular perspective. In general, the perspective is the speaker’s. Perspectival shifts, e.g., initiated by reported speech or attitude reporting verbs, as in (24-a) and (24-b), clash with the strong association of use-conditional content with the immediate speaker’s *origo*, although there can be deviations if conditions are right (cf. Amaral 2018, p. 329). In fact, (24-c) may mark such an exception, since an anonymous reviewer points out that it is acceptable in the $lá_{UC}$ reading. This does not exclude the possibility of a nonshifted interpretation though, where $lá_{UC}$ is still attributed to the speaker’s perspective. Concerning future studies, this suggests investigating if sentence mood impacts the shiftability of $lá_{UC}$, i.e., if it is shiftable indeed when occurring in declaratives.

- (24) a. *A tua tia é muito severa, sabias? Pede ao nosso filho que acabe [$*lá_{UC}/lá_{TC}$] de comer a sopa.*
 ‘Your aunt is very strict, you know. She asks our son to finish their soup.’
- b. [Mãe: *Acaba lá de comer a sopa!* Filho: *Não percebi, o rádio está muito alto.* Pai: *A tua mãe quer que acabes [$*lá_{UC}/lá_{TC}$] de comer a tua sopa.*
 ‘Finish your soup! Son: I do not understand, the radio is too loud. Dad: Your mother wants you to finish your soup.’
- c. *O João tem estado muito doente, mas o Pedro disse que ontem [$lá_{UC}/lá_{TC}$] se sentou à mesa.*
 ‘João has been very ill, but Pedro said he sat at the table yesterday.’

The property of descriptive ineffability states that “speakers are never fully satisfied when they paraphrase expressive content using descriptive, i.e., non-expressive, terms” (Gutzmann 2013, p. 42). In fact, this criterion has received a lot of criticism. An important issue raised by Geurts (2007, p. 210) is that there are also truth-conditional items with meaning content “suffering” from descriptive ineffability, such as “the” or “because” (cf. also Amaral 2018, p. 328); i.e., this cannot be unique to expressive/use-conditional meaning. In fact, Pereira (2009) actually does provide paraphrases for some of the examples discussed above. However, it is easy to notice that these ponderous explanations, complemented by further clarifications concerning the context, confirm that they are not to be understood as semantically equivalent to what is conveyed by $lá_{UC}$.

The last criterion, the one of immediacy, covers that expressives/use-conditional items, “like performatives [...], achieve their intended act simply by being uttered” (Gutzmann 2013, p. 44). Put differently, while truth-conditional meaning can be revised in certain ways, use-conditional meaning can hardly be taken back, and its effect certainly evolves once the underlying use-conditional item has been uttered. As alluded to by (Potts 2007, p. 180),

immediacy is related to the non-deniability in discourse discussed above and also applies to $lá_{UC}$.

In sum, the discussed data support the claim that non-adverbial $lá$ can be understood in terms of the use-conditional item $lá_{UC}$.

3.3. A Note on Meta-Linguistic Negation

As mentioned above, here is a brief note on certain uses of $lá$ that are excluded from the analysis.

(25) (Marques and Duarte 2015, p. 118).

- a. *Sei lá.*
'I do not know.'
- b. *Quero lá saber.*
'I do not want to know.'

Regarding $lá$'s meaning contribution in these examples, (Marques and Duarte 2015, p. 118) allude to its negating function, which is also conveyed by the English translations 'I do not know' and 'I do not want to know', complemented furthermore by an axiological distance between the speaker and content, i.e., again by illocutionary modification as discussed in the previous subsection. In fact, this might actually suggest a status of a mixed item in terms of Gutzmann (2015), featuring both truth-conditional (negation) and use-conditional content (illocutionary modification):

$$(26) \quad Sei\ lá = \frac{\emptyset}{sei} + \frac{\text{illocutionary modification}}{\neg}$$

$$= \frac{\text{illocutionary modification}}{\neg sei}$$

In the sense of (Martins 2014, p. 638), these uses are described as giving rise to metalinguistic negation on merely pragmatic grounds in the tradition of Horn (1989):

(27) A: *Estás um pouco preocupado?*—B: *Estou lá um pouco preocupado, estou morto de preocupação.*
'A: Are you a little worried?—B: I'm not a little worried, I am worried sick'

(Horn 1989, p. 363) defines metalinguistic negation as an objection of a previous utterance by not focusing on its truth conditions but on its assertability. In this vein, $lá$ in (27) does not negate the truth conditions of the previous utterance but objects its understatement, as it prepares an utterance that is actually meant to reinforce it. According to this view, $lá$ does clearly not carry a negation operator in its truth-conditional meaning component. Instead, the negating effect is a product of a pragmatic inference that furthermore does not operate on the descriptive meaning. Conceiving of this implicature as a conventional one, metalinguistic negation might be understood as an own use-conditional meaning component as shown in (28).

$$(28) \quad estou\ lá\ um\ pouco\ preocupado = \frac{\emptyset}{estou\ um\ pouco\ preocupado} + \frac{\text{metalinguistic negation}}{\emptyset}$$

$$= \frac{\text{metalinguistic negation}}{estou\ um\ pouco\ preocupado}$$

Another option though would be to understand metalinguistic negation as just another pragmatic outcome of illocutionary modification (29). Based on the conception of emotional deixis, $lá$ accordingly creates a very strong mitigation effect of subjective distance. This could be understood as leading to a tremendously low degree of assertivity/commitment to what is said, creating thus metalinguistic negation pragmatically.

$$(29) \quad estou\ lá\ um\ pouco\ preocupado = \frac{\emptyset}{estou\ um\ pouco\ preocupado} + \frac{\text{illocutionary modification}}{\emptyset}$$

$$= \frac{\text{illocutionary modification}}{estou\ um\ pouco\ preocupado}$$

Assuming (29) would mean that there is nothing conventional about metalinguistic negation (the conventional underlying meaning would be again illocutionary modification). In fact, this is supported by (30), where an interpretation of ‘God does not know’ is clearly ruled out. This suggests a merely pragmatic nature of metalinguistic negation indeed, derived on the grounds of the semantically stored illocutionary modification.

- (30) *Sabe lá Deus se chegamos amanhã.* (Pereira 2009, p. 94)
 ‘God [knows/*does not know] if we’ll arrive tomorrow.’

In sum, it seems plausible to treat metalinguistic negation as an implicature derived from $lá_{UC}$ ’s use-conditional meaning of illocutionary modification. Since this assumption calls for more substantiating research though, e.g., also concerning the exact type of implicature, I leave the intriguing *desideratum* of a proper multidimensional treatment of metalinguistic negation for future research.⁶

4. *Lá* in Diachrony

4.1. A Case of Pragmaticalization: From $lá_{TC}$ to $lá_{UC}$

In this section, the study of *lá* is approached from a diachronic perspective, based on the working hypothesis that $lá_{UC}$ historically evolved from $lá_{TC}$. In fact, non-adverbial uses of *lá* have also been discussed in the previous literature as a result of a derivation process. Though not explicitly stated, they seem to be understood as a product of a synchronic, ad hoc creation made in conversation. Put differently, the assumption is that there is only one lexical entry for *lá* such that $lá_{UC}$ is conceived of as a result of an implicature-based inference. What drives this process is assumed to be an inferred transfer of $lá_{TC}$ ’s spatial distance to the domain of subjective distance, an operation whose result is also known under the general label of emotional deixis (cf. Lakoff 1974; Potts and Schwarz 2010; Aguiar and Barbosa 2023, cf. also Lopes 2019 and the discussion of further Portuguese examples of emotional deictics, like *cá* (‘here’) and *aí* (‘there’)). In a broader research context, assuming one and only lexical entry for *lá* relates to the monosemy vs. polysemy (or minimalist vs. maximalist) problem (cf. e.g., Hansen 1998, p. 85 and Waltereit 2006, p. 8 in the context of discourse/modal particles). This minimalist view seems to be particularly in line with the high degree of ambiguity obtained in many examples of *lá*, which oftentimes makes it also hard to distinguish $lá_{UC}$ from $lá_{TC}$ in linguistic analysis for native speakers (cf. Marques and Duarte 2015, p. 121). On the other hand, there is a great body of research on the diachronic rise of discourse/modal particles (cf. e.g., Detges and Waltereit 2007) assuming a historic process leading up to synchronic polysemy, i.e., a reflection of diachronic change. A guiding principle in this process has been identified in terms of the subjectification thesis and the related concept of pragmaticalization (Traugott 1989, cf. also Diewald 2011).⁷ In (Traugott 1989, pp. 34–35), three general tendencies of subjectification are listed. Among these are the typical process from external (description of situation) to internal (evaluative/perceptual/cognitive) and the development of expressing “the speaker’s subjective belief state/attitude toward the proposition”, that qualify to also apply to a diachronic formation of $lá_{UC}$. Within multidimensional semantics, pragmaticalization processes are assumed to typically evolve from descriptive/truth-conditional to expressive/use-conditional content (cf. Davis and Gutzmann 2015; Amaral 2018, p. 331), which the present analysis is built on. Accordingly, the process of pragmaticalization is understood as a diachronic type shift from truth to use conditions, illustrated in (31) with examples in English and German.

- (31) Pragmaticalization as a type shift (Davis and Gutzmann 2015, p. 203).
- a. $A : \langle \sigma, t \rangle > A_{ex} : \langle \sigma, u \rangle$.
 - b. Descriptive nouns > expressives (Davis and Gutzmann 2015, p. 198).
 - (i) Boor ‘countryman, farmer’ > ‘crude person’ (Engl.).
 - (ii) Wip ‘woman’ > weib ‘woman.PEJ’ (Germ.).

- c. Adverbs/adjectives > modal particles (Germ., Davis and Gutzmann 2015, p. 198).
 - (i) Eben ‘flat’ > ‘just, exactly’.
 - (ii) Schon ‘already’ > ‘somewhat’.

As a fundamental building block in the approach, a two-step pattern is assumed. This means that there is an intermediate stage in the derivation of *boor*_{EXP} or *eben*_{MP}, quite similar to bridging contexts in the overlap model known from grammaticalization theory (cf. e.g., Heine 2003). The diamond operator (\diamond) in (32) therefore indicates a mixed use-conditional item that contains both a truth- and a use-conditional meaning dimension that mediates between the truth-conditional origin and the purely use-conditional destination:

- (32) Two-step pragmaticalization.
- a. $A > A \diamond A_{ex} > A_{ex}$.
 - b. Boor: $\langle e, t \rangle > boor \langle e, t \rangle \diamond boor_{ex} : \langle e, u \rangle > boor_{ex} \langle e, u \rangle$.

The intermediate stage is typically associated with ambiguous contexts oscillating between an old and a new meaning, before the new one gets stored as conventionalized. In more detail, this process is understood as follows. A purely descriptive/truth-conditional item starts to generate expressive meaning at some point by means of a conversational implicature that is generated in particular contexts and situations (Stage I). Once there is enough of a regularity, the conversational implicature turns into a conventionalized one, stored now as conventional use-conditional content that complements the truth-conditional dimension (Stage II). Finally, the truth-conditional meaning component may be lost, creating a purely use-conditional item, as is the case with, e.g., “boor” (Stage III). However, in case of what the authors call pragmatic fission (33), the third stage may also split into a purely truth-conditional item restored along with the use-conditional item, creating lexical ambiguity.

- (33) Pragmatic fission (Davis and Gutzmann 2015, p. 204).

$$A > A \diamond A_{ex} > \begin{cases} A \\ A_{ex} \end{cases}$$

Applied to the present study, this is a promising scenario to understand the process that *lá* has passed through diachronically and that we keep as a working hypothesis for the corpus study discussed in the next section.

- (34) Possible pragmaticalization path of *lá*.

$$lá > lá_{TC} \diamond lá_{UC} > \begin{cases} lá_{TC} \\ lá_{UC} \end{cases}$$

4.2. Corpus Study

For the corpus query, the historical section of the *Corpus do Português* was consulted Davies (2006). Based on the working hypothesis that *lá*’s pragmaticalization path can be described with (34), I particularly hypothesized the following stages.

- Stage I:
The original morpheme *lá* carries only truth-conditional content and serves to modify the location of a predicate. A conversational implicature of illocutionary modification may be generated in certain contexts, leading to mitigation or reinforcement effects as discussed in Section 3.2.
- Stage II:
In certain contexts, the implicature of illocutionary modification is conventionalized, leading to a hybrid, i.e., two-dimensional expression.

- Stage III:
The hybrid expression splits into two homophonous expressions, one encoding original truth-conditional content and another encoding use-conditional content.

The related research questions that the query aims at can be formulated as follows.

- What are the time frames corresponding to the stages formulated above?
- What are typical onset contexts for (some of) the stages, i.e., potential linguistic properties promoting the development?

Regarding the corpus queries, the procedure was as follows. First, a general overview of the frequencies of *lá* in diachrony was generated, which is illustrated in Table 2. It shows the raw frequencies of the lexical item *lá* per century. Given that the subcorpora for each century are not balanced in terms of their size, I calculated ratios by dividing the number of tokens by the corpus size in order to compare them.⁸ The subcorpus for the 20th century is furthermore annotated according to diatopic variation, based on the distinction of European (EP) and Brazilian Portuguese (BP), as well as for register, based on the distinction of academic, newspaper and fictional texts, as well as spoken language. This means that the subcorpora for the 16th–19th century contain mixed data in terms of variation, which impedes filtering exclusively for EP data.⁹ In order to maintain the comparability across the columns, the query for the 20th century was not restricted either. In this vein, the data presented in Table 2 provide a general overview of the most important trends in the life cycle of *lá* in Portuguese.¹⁰

Table 2. Total occurrences of *lá* in the *Corpus do Português* per century (Davies 2006).

	16th	17th	18th	19th	20th
Tokens	1644	977	305	10,499	16,170
Size	4,435,031	3,407,741	2,234,951	10,008,622	20,777,725
Tokens/size (*1000)	0.37	0.29	0.03	0.17	1.05

Before describing the findings, another note of clarification is needed. The search of instances of *lá* is based on the spelling *lá*. This is not trivial, as many occurrences in older texts are still spelled *la*, i.e., without an acute accent, due to the diachronic evolution of the feminine definite article of Lt. *illa(m) > la > a* (Azevedo 2005, p. 158). I refrained from including the spelling *la* in the query due to the problem of the massive quantity of definite articles in the data. In this sense, the query is based on a sample. This should not be problematic, since corpus data always represent samples.¹¹ As can be seen in Table 2, the sample is big enough for the purposes of the study, allowing for generalizations.

Considering the data presented in Table 2, there is a decline to be noted leading from 0.37 in the 16th to a low point at 0.03 in the 18th century, before the numbers rise again, until they explode in the 20th century (1.05). This suggests an important turning point around the 18th/19th century, serving as a preparatory stage for the upcoming rise of *lá*. Potentially, it was not until then that speakers started to conventionally add a use-conditional meaning component to *lá* (Stage II). Going in that direction, the oldest examples suggesting use-conditional uses were of interest in the next step. Therefore, the search focussed on *lá* occurring in singular imperatives, both in formal and informal ones (e.g., *tome lá!/toma lá!* '(please) take!'). The imperative construction was chosen for two reasons. First, this again meant sampling, i.e., filtering for the sentence mood of the imperative to improve the overview of the data. Second, imperatives are described to be the central sentence mood in terms of frequencies that *lá_{UC}* occurs in (Marques and Duarte 2015, p. 123), which suggests a high degree of pragmaticalization. Accordingly, imperatives should be amenable to implicatures of illocutionary modification from very early on already. For the formal imperatives, the search was based on the keyword [VS* 1á] to find the most frequent constructions, complemented by manual searches for the informal imperative of these. Note that both forms are equivalent with other ones: the informal *toma*

coheres with the third person singular indicative present tense and the formal *tome* with the third person singular subjunctive present tense, meaning that qualitative inspections of relevant examples are needed.

An overview of the results of the described query is shown in Table 3. Similarly to Table 2, the 18th/19th century seems to be an important turning point, as there is a rise to be registered from the 19th century onwards. Frequency has actually been shown to be an important catalyst in pragmaticalization (Bybee 2003). With an increase with respect to both type and token frequency, conditions are good for a new function to be consolidated in the language system or, to put it in individual-cognitive terms, conditions are good for a new function to be entrenched, before *lá_{UC}* receives its own lexical entry at a later stage. Furthermore, the data suggest a strong focus on oral and fictional registers in the 20th century, thus confirming previous studies. The striking role of fictional data in this regard can be explained by a likely high amount of staged orality, i.e., of imitations of spoken language or language of proximity in the sense of Koch and Oesterreicher (1985).

Another crucial insight revealed when comparing the two tables lies in that the data are compatible with the idea of a gradual pragmaticalization of *lá* over the centuries. While the general frequencies of *lá* developed from 0.37 to 0.03 to 1.05 (a curve like a checkmark), this trend is clearly not reflected in Table 3 (explosion from the 18th to the 19th centuries). Crucially, this suggests ruling out the possibility that *lá_{TC}* has been available ever since as a synchronically available, secondary meaning component in the sense of the minimalist view discussed above. Otherwise, the data in Table 3 should much more resemble the trend depicted in Table 2. Instead, the use-conditional meaning component of *lá* must have been added in the course of the time as a result of a process of meaning expansion.

Table 3. Singular imperative + *lá*, CDP.

	16th	17th	18th	19th	20th	EP	BP	ACA	NEWS	FICT	ORAL
VEJA LÁ		1	1	62	94	86	8		7	44	43
VÊ LÁ			1	83	85	67	18	1	1	59	24
DIGA LÁ				36	51	45	6		3	23	25
OLHE LÁ				37	40	23	17		2	20	18
OLHA LÁ		1		47	62	48	14		5	33	24
DEIXE LÁ			1	42	32	29	3			21	11
DEIXA LÁ				45	43	37	6	1	4	29	9
ANDE LÁ			1	32	6	2	4			6	
ANDA LÁ	1	2		38	24	21	3		1	20	3
TOME LÁ				30	9	7	2		1	6	2
TOMA LÁ	1			36	23	21	2		5	15	3
ESPERE LÁ				17	14	14			1		13
ESPERA LÁ				10	11	11				5	6
OUÇA LÁ				7	18	18			1	1	16
OUVE LÁ				20	23	20	3			19	14
CONTE LÁ				7	14	13	1			4	10
CONTA LÁ	1			10	16	14	2			15	1
DIGA-ME LÁ				4	11	11				1	10
DIZ-ME LÁ					5	5				3	2
DESCULPE LÁ					9	9				1	8
DESCULPA LÁ					2	2			1	1	
EH LÁ				5	3	2	1			3	

As a next step, I checked each of the examples occurring prior to the 19th century for uses of *lá_{UC}*, i.e., featuring illocutionary modification or at least a flavor thereof, suggesting either a status of Stage I, i.e., of a conversational implicature of illocutionary modification (35), or of Stage II, i.e., of a conventional implicature of illocutionary modification (36). Examples hosting clear uses of *lá_{TC}* are not listed here, but they can easily be accessed online using the platform.

- (35) Examples compatible with Stage I (conversational implicature of illocutionary modification).
- a. Manuel Bernardes, “*Nova floresta*”, 1710.
Eis aí te fez Cristo digno de levares a sua cruz e segui-lo; vê lá nao faças pé atrás e acudas mais pela tua carne que pela tua alma e te furte o diabo o reino da glória.
 ‘Behold, Christ has made you worthy to take up your cross and follow him. Do not stand back and fight more for your flesh than for your soul, and let the devil rob you of the kingdom of glory.’
 - b. Francisco Manuel de Melo, “*Apolo*”, 17th century.
Mulher, olha lá como amaldiçoas: não toques no campanário!
 ‘Woman, look how you curse: do not touch the bell tower!’
 - c. Candido Lusitanoazendo, “*Carta a um filólogo de Espanha*”, 1750.
Como v.m. o tem, veja lá se gosta daquelle timbre das armas, e timbre daquelles, por quem chora o Tejo....
 ‘Since you have it, see if you like that timbre of the arms, and timbre of those for whom the Tagus weeps....’
 - d. Manuel da Costa, “*Arte de Furtar*”, 1645.
se algum não tiver isto por factivel, veja lá não lhe provêm, que lhe succedeo a elle.
 ‘If anyone does not take this for granted, look at what happened to him.’

In each of these examples, *lá* is used together with predicates amenable to a locative semantic role, i.e., to local specification. In (35-a) and (35-b), *ver* and *olhar* (‘look, see’) are easily amenable to a specification of where to look. Given that the examples furthermore connect with the idea of a subjective meaning component, the presence of a conversational implicature seems to be plausible. In (35-a), the reinforcement effect goes hand in hand with the seriousness of the religious task of not standing back. In (35-b), there is expressivity triggered by *almadiçoar*, also suggesting a reinforcing effect. In (35-c), the opposite function seems to in place, i.e., a mitigating effect, suggested by the socially much higher addressee of *vossa mercê* (‘your grace’). Example (35-d) is a bit out of competition, as it is part of a monologue, i.e., there is no direct addressee in “if somebody [...]”, and it is embedded under a conditional operator *se*, which might be problematic with expressive/use-conditional items as discussed in Section 3. In fact, (Sambrana 2021, p. 94) analyzes this very example as a first occurrence of *veja lá* as a discourse marker, which suggests that it not necessarily about illocutionary modification but more about discourse organization.

Concerning the candidates for Stage II listed in (36), there are now predicates involved that do not actually call for a locative semantic role, like *tomar* and *contar* (‘(to) take, (to) tell’). As indicated by a reviewer, there is actually also a locative component contained in the semantics of *tomar*, as there is a movement of an item away from the speaker origo towards a goal. However, the goal of the location is exactly defined by the receiver, so there is no need for further specification, which marks the contrast with *olhar* and *ver* as discussed above in (35).

- (36) Examples compatible with Stage II (conventional implicature of illocutionary modification).
- a. Gil Vicente, *Farsa do juiz da beira*, 1525.
Vai-se e vem um Escudeiro com um seu Moço e diz: Toma lá esse sombreiro.
 ‘A squire comes along with a young man and says, “Take that parasol.”’
 - b. Gil Vicente, *Auto da barca do purgatório*, 1518.
Senhor tartarugo[,] digo que mentis como bestigo, salvaror. Fala em tua menencória e não fales em passar e conta lá outra história porque em festa de tal glória não hás ninguém de levar.

‘Mr. Tortoise, I say you lie like a beast, savior. Speak in your menagerie and do

not talk about passing and tell another story because you have no one to take to such a glorious feast.'

Given that these are nevertheless combined with *lá* seems now to be possible because of the conventionalization of the implicature of illocutionary modification, which had been previously generated as a conversational one. Although declaratives were not explicitly searched, there is one in (36-b), featuring a clearly evaluating function, i.e., reducing the degree of assertivity as discussed in Section 3.2. Future studies may thus follow up on whether *lá*_{UC} with declaratives developed after *lá*_{UC} was conventionalized with imperatives indeed. That would mean that at first, conversational implicatures (Stage I) were generated with imperatives, before in Stage II, where the conventionalization already enabled a broader range of sentence mood. Related to this, the first instances of the purely exclamative *eh lá* are also indicated in Table 3. They trace back to the 19th century and also suggest an emergence posterior to imperatives.

It may seem a bit counterintuitive that the discussed examples for Stages I and II do not perfectly mirror their actual chronological occurrences. In this vein, Gil Vicente appears as an innovative promoter, as he is the only author that already very early on uses *lá* with rather advanced functions in terms of its pragmatization path. Since processes of change are multilayered developments that may be promoted to different degrees in different speech communities at different times before conventionalization in the system is accomplished, this does not directly contradict the discussed assumptions. In particular, the change process from conventionalizing a conversational implicature is not to be understood as an abrupt and clearcut process that speakers suddenly obey. As noted by a reviewer, the leading role of Vicente's theater plays may be particularly due to the contained passages featuring informality and language close to orality and the fact that the number of theater plays in the corpus decreases over the centuries. In fact, a more balanced corpus in terms of register would certainly improve the whole picture, particularly regarding the frequencies and distributions of innovative uses to see if the conventionalization proceeded to different extents in different registers at different times. In (37), examples from the 19th century are listed. They suggest a status of either Stage II or even III. Note that in the historic corpus-based methodology, it is hardly possible to identify when exactly the transfer from Stage II to III must have taken place. Using synchronic experimental methods, this could be approached by testing to what extent the typical properties of expressive/use-conditional items discussed in Section 3.2 apply to *lá* at a given stage. Given that there is no operationalization suggested in Davis and Gutzmann (2015) either, I leave it for further research to identify if there are other ways to distinguish Stage II and III based on diachronic corpus data. In both examples in (37), it is obvious though that there is not a potential but a certain reinforcement effect associated with *lá* and thus a strong subjective note of involvement by the speaker, emphasized by their impatience and joy. In sum, the 19th century clearly started to host a great number of uses of pertaining to at least Stage II, whereas prior occurrences were still highly sporadic.

(37) Stage II (and III?).

a. Eça de Queirós, "O Crime do Padre Amaro", 1875.

E eu por mim o que fazia, para calar toda essa gente, era casar-me já. Eu bem sei que tu não morres por ele, bem sei. Deixa lá! Isso vem depois. O João é bom rapaz, vai ter o emprego.

'And what I'd do, to shut all those people up, is get married now. I know you will not die for him, I know that. Never mind! That'll come later. João is a good boy, he'll get the job.'

b. Almeida Garrett, "Falar verdade e mentir", 1845.

Vamos fala, conta-me lá como isso foi, quero saber tudo por meudo.

'Come on, tell me how it happened, I want to know everything.'

- c. Camilo Castelo Branco, “*Maria Moisés*”, 1876.
Vá, mano, conte lá a história.
 ‘Come on, man, tell the story.’

4.3. Discussion

The results of the corpus analysis turn out to be compatible with the hypotheses formulated above, crucially with the claim that *lá* passed through a diachronic process of pragmaticalization of $lá_{TC} > lá_{UC}$. With respect to the first research question of the corpus study, the data show first instances of conversational and also conventionalized implicatures from the 15th century onwards and suggest a stronger conventionalization from the 19th century onwards. According to the discussed account, Stage III must have been approached at around this time. Concerning the transition from Stage II to III, it was not possible to identify a specific period using the current method, mainly due to limitations of the historical corpus approach. Regarding the second research question, the amenability of predicates to the semantic role of location turned out to serve as an indicator for the degree of pragmaticalization at a certain stage. In this sense, the analysis suggests that, at first, predicates prone to call for a locative semantic role allowed for the generation of conversational implicatures of illocutionary modification before there was an extension to predicates less amenable to specifying a location.

5. Conclusions

In this contribution, I analyzed European Portuguese *lá* based on a multidimensional semantic framework in the sense of Gutzmann (2013, 2015). Crucially, I showed a synchronic split into two conventionalized meaning configurations leading up to polysemy, i.e., to a truth-conditional and a use-conditional *lá*. While $lá_{TC}$ is well known to contribute to the descriptive meaning of an utterance by adverbially modifying a location, $lá_{UC}$ expresses use-conditions (or felicity conditions) that apply to particular contexts (instead of worlds). From a synchronic point of view, this claim was substantiated based on a range of diagnostic tests for use-conditional meaning discussed in the literature. Concerning the diachronic dimension, data from the *Corpus do Português* were presented, supporting the assumption of a two-step pragmaticalization path of $lá > lá \diamond lá_{UC} > lá_{TC} / lá_{UC}$ and resulting in pragmatic fission in the sense of Davis and Gutzmann (2015).

Regarding previous research on *lá*, I would like to highlight the following three achievements. First, it was possible to reduce a great part of the inventory of functions assumed for non-adverbial *lá* to the one of use-conditional meaning of illocutionary modification, which also holds across sentence moods. As discussed in Section 3.3, the analysis might also be compatible with metalinguistic negation (in the sense of Martins 2014), which, however, remains to be investigated in detail. As pointed out by a reviewer, another function of *lá* not treated in the analysis refers to it occurring in particular information structural constructions, e.g., in topicalizations. Again, it remains to be studied if examples like *Lá de Jorge nada temas* (‘With respect to Jorge, you do not fear anything.’) are also compatible with the analysis or not. Second, the semantic meaning contribution of $lá_{UC}$ was highlighted and delimited from the pragmatic potential associated with it. Third, it was possible to explain the formation of $lá_{UC}$ based on an pragmaticalization approach featuring the idea of a stepwise conventionalization.

Besides these achievements, there is a range of open questions remaining for future work. Next to the ones mentioned above, another one certainly pertains to whether there are the same synchronic effects and diachronic processes to be observed with *lá* in other varieties of Portuguese, as well as with other spatial deictics such as *cá* (‘here’) and *aí* (‘there’).

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: I am very grateful to the anonymous reviewers for their comments and suggestions, as well as to Patrícia Amaral, who invited me to contribute to this Special Issue. Furthermore, I wish to thank Beatriz de Medeiros Silva for discussions on the constructed Portuguese examples in Section 3.2.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

$lá_{TC}$ truth-conditional *lá*
 $lá_{UC}$ use-conditional *lá*

Notes

- ¹ Examples in (1) taken from <https://dicionario.acad-ciencias.pt/pesquisa/?word=I%C3%A1> (accessed on 10 May 2024).
- ² *Aí* is also well known for its non-adverbial function(s) in Brazilian Portuguese, cf. e.g., Johnen (1997); Aquino and Kahil (2022). As pointed out by a reviewer, another spatial deictic that is not listed in (7) is *acolá*, cf. e.g., Teixeira (2005).
- ³ The use-conditional dimension in (26) is left empty here for the sake of simplicity. (Gutzmann 2015, p. 202) claims that sentence mood is a use-conditional item which is thus generally stored in the use-conditional dimension of a sentence.
- ⁴ The listed types of sentence mood are not meant to entail their illocutionary counterpart (cf. e.g., indirect speech acts).
- ⁵ An anonymous reviewer points out that (22-b) is actually acceptable. This means that a revolting refusal of the illocutionary force associated with an order like *Acaba de comer a sopa!* is always possible. It remains an open question though whether (22-a-iv) has better acceptability ratings than (22-b), which would mean that the *só que* construction works slightly better with an antecedent indeed.
- ⁶ According to (Pereira 2011, p. 64), negation phenomena with *lá* are also present in BP, though they are strictly restricted to rhetorical questions, as opposed to EP.
- ⁷ In the literature, the rise of discourse/modal particles has also been analyzed as a case of grammaticalization, cf. e.g., Wegener (1998) and the discussion in Liesbeth and Jacqueline (2017) and Detges and Waltereit (2016).
- ⁸ The reason for the smaller size of the subcorpora for the 17–18th century might be, among other factors, the great earthquake in Lisbon in 1755 that destroyed nearly the whole city, including lots of archived documents and texts, cf. (Teixeira 2015, p. 122) and the reference therein to Matos (2001).
- ⁹ A full list of the approximately 57,000 texts included in the corpus is provided here: <https://www.corpusdoportugues.org/hist-gen/help/cdp.xls> (accessed on 10 May 2024).
- ¹⁰ In fact, it seems plausible that *lá* in BP might have undergone a similar evolution as in EP, since it synchronically disposes of similar uses (cf. Pereira 2011) besides the very frequent *aí*.
- ¹¹ A reviewer wondered if there might be a relevant amount of $lá_{TC}$ (or even $lá_{UC}$) written without an accent in the earlier centuries that were lost in the query. A check using the same formula used in the generation of Table 3 without accent reveals a very small number of data points featuring deictic *la*. This suggests that the trend in the evolution depicted in Table 2, based on the spelling *lá*, is representative.

References

- Aguiar, Joana, and Pilar Barbosa. 2023. Emotional deixis in online hate speech. In *Hate Speech in Social Media: Linguistic Approaches*. Berlin and Heidelberg: Springer, pp. 139–64.
- Ali, Manuel Saïd. 1971. *Meios de expressão e alterações semânticas*. Rio de Janeiro: Francisco Alves. First published 1930.
- Amaral, Patrícia. 2018. Expressive meaning. In *Handbuch Pragmatik*. Edited by Frank Liedtke and Astrid Tuchen. Berlin and Heidelberg: Springer, pp. 325–33.
- Aquino, Marcell Chérchiglia, and Tamires Arnal Kahil. 2022. As partículas modais *mas* e *aí* pela perspectiva de falantes do português brasileiro: Uma investigação da linguagem em uso. *Confluência* 62: 172–98. [CrossRef]
- Azevedo, Milton M. 2005. *Portuguese: A Linguistic Introduction*. Cambridge: Cambridge University Press.
- Briz, Antonio, and Marta Albelda. 2013. Una propuesta teórica y metodológica para el análisis de la atenuación lingüística en español y portugués. la base de un proyecto en común (es. por. atenuación). *Onomázein* 28: 288–319.
- Brown, Penelope, and Stephen C. Levinson. 1978. Universals in language usage: Politeness phenomena. In *Questions and Politeness*. Edited by Esther N. Goody. Cambridge: Cambridge University Press, pp. 56–289.

- Bybee, Joan. 2003. Mechanisms of change in grammaticization: The role of frequency. In *The Handbook of Historical Linguistics*. Oxford: Blackwell, pp. 602–23.
- Costa, João, and Ana Maria Martins. 2010. Middle scrambling with deictic locatives in european portuguese. In *Romance Languages and Linguistic Theory 2*. Amsterdam: John Benjamins, pp. 59–76.
- Cruse, D. Alan. 1986. *Lexical Semantics*. Cambridge: Cambridge University Press.
- Davies, Mark. 2006. Corpus do Português: 45 Million Words, 1300s–1900s. Available online: <http://www.corpusdoportugues.org/hist-gen/> (accessed on 9 May 2024).
- Davis, Christopher, and Daniel Gutzmann. 2015. Use-conditional meaning and the semantics of pragmaticalization. In *Proceedings of Sinn und Bedeutung*. Edited by Eva Csipak and Hedde Zeijlstra. Göttingen: University of Göttingen, vol. 19, pp. 197–213.
- Detges, Ulrich, and Richard Waltereit. 2007. Different functions, different histories: Modal particles and discourse markers from a diachronic point of view. *Catalan Journal of Linguistics* 6: 61–80.
- Detges, Ulrich, and Richard Waltereit. 2016. Grammaticalization and pragmaticalization. In *Manual of Grammatical Interfaces in Romance*. Berlin and Boston: De Gruyter, pp. 635–58.
- Diewald, Gabriele. 2011. Pragmaticalization (defined) as grammaticalization of discourse functions. *Linguistics* 49: 365–390. [CrossRef]
- Duarte, Isabel Margarida. 1989. Alguns operadores de agulhagem comunicativa (na prosa narrativa de Eça de Queirós e José Cardoso Pires). Master's thesis, Faculdade de Letras, Universidade do Porto, Porto, Portugal.
- Duarte, Isabel Margarida. 2010. La dimension modale de “că” et “lá” en portugais. *Studii și Cercetări Lingvistice* LX: 179–95.
- Favaro, Marco. 2023. *Modal Particles in Italian. Adverbs of Illocutionary Modification and Sociolinguistic Variation*. Berlin: Language Science Press.
- Franco, António. 1988. Partículas modais da língua portuguesa: Relances contrastivos com as partículas alemãs. *Revista da Faculdade de Letras do Porto Línguas e Literatura* 5: 137–56.
- Franco, António. 1991. *Descrição Linguística das Partículas Modais no Português e no Alemão*. Coimbra: Coimbra Editora.
- Geurts, Bart. 2007. Really fucking brilliant. *Theoretical Linguistics* 33: 209–14. [CrossRef]
- Grewendorf, Günther, and Dietmar Zaefferer. 1991. Theorien der Satzmodi. In *Semantik: Ein internationales Handbuch der zeitgenössischen Forschung; Semantics: An international Handbook of Contemporary Research*. Edited by Armin von Stechow and Dieter Wunderlich. Berlin and New York: De Gruyter, pp. 270–86.
- Grice, H. Paul. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*. Edited by Peter Cole and Jerry L. Morgan. New York: Academic Press, pp. 41–58.
- Gutzmann, Daniel. 2013. Expressives and beyond: An introduction to varieties of use-conditional meaning. In *Beyond Expressives: Explorations in Use-Conditional Meaning*. Edited by Daniel Gutzmann and Hans-Martin Gärtner. Leiden: Brill, pp. 1–58.
- Gutzmann, Daniel. 2015. *Use-Conditional Meaning: Studies in Multidimensional Semantics*. Oxford Linguistics. Oxford: Oxford University Press.
- Hansen, Maj-Britt Mosegaard. 1998. *The Function of Discourse Particles: A Study with Special Reference to Spoken Standard French*. Amsterdam: John Benjamins.
- Heine, Bernd. 2003. Grammaticalization. In *The Handbook of Historical Linguistics*. Oxford: Blackwell, pp. 575–601.
- Hengeveld, Kees. 2004. Illocution, mood, and modality. In *Ein Internationales Handbuch zur Flexion und Wortbildung*. Edited by Geert E. Booij, Christian Lehmann, Joachim Mugdan and Stavros Skopeteas. Berlin and New York: De Gruyter Mouton, pp. 1190–201. [CrossRef]
- Holmes, Janet. 1984. Modifying illocutionary force. *Journal of Pragmatics* 8: 345–65. [CrossRef]
- Horn, Laurence Robert. 1989. *A Natural History of Negation*. Chicago: University of Chicago Press.
- Johnen, Thomas. 1997. *Aí* como partícula modal do português. In *Comunicações. Edição informatizada*. Salvador: Congresso Internacional da Associação Brasileira de Linguística.
- Kaplan, David. 1999. The meaning of ouch and oops: Explorations in the theory of meaning as use. Manuscript, UCLA [not at hand].
- Koch, Peter, and Wulf Oesterreicher. 1985. Sprache der Nähe—Sprache der Distanz. *Romanistisches Jahrbuch* 36: 15–43. [CrossRef]
- Lakoff, Robin. 1974. Remarks on ‘this’ and ‘that’. In *Proceedings of the Chicago Linguistics Society* 10. Chicago: Chicago Linguistic Society, pp. 345–56.
- Leech, Geoffrey. 1983. *Principles of Pragmatics*. London and New York: Longman.
- Liesbeth, Degand, and Evers-Vermeul Jacqueline. 2017. Grammaticalization or pragmaticalization of discourse markers? *Journal of Historical Pragmatics* 16: 59–85. [CrossRef]
- Longhin-Thomazi, Sanderléia Roberta. 2003. A perífrase conjuncional só que: Invariância e variantes. *ALFA: Revista de Linguística* 47: 139–52.
- Lopes, Ana Cristina Macário. 2019. A deixis emocional em português europeu contemporâneo: Alguns contributos. *Revista da Associação Portuguesa de Linguística* 5: 225–35. [CrossRef]
- Marques, Aldina, and Isabel Margarida Duarte. 2015. Cá e lá: Atenuação, reforço e outros valores modais em PE. *Acta Semiotica et Linguística* 20: 115–28.
- Marques, Maria Aldina, and Isabel Margarida Duarte. 2017. Lá, atenuador em interações informais do português europeu. *Studia Universitatis Babeş-Bolyai, Philologia* 62: 17–34. [CrossRef]
- Martins, Ana Maria. 2014. How much syntax is there in metalinguistic negation? *Natural Language & Linguistic Theory* 32: 635–72.

- Martins, Ana Maria. 2021. Syntactic aspects of metalinguistic negation. *Revue Roumaine De Linguistique-Romanian Review of Linguistics* 66: 175–97.
- Matos, Artur Teodoro. 2001. *Diário do Conde de Sarzedas, Vice-Rei do Estado da Índia (1655–56)*. Lisbon: Comissão Nacional para as Comemorações dos Descobrimientos Portugueses.
- Meisnitzer, Benjamin. 2012. Modality in the Romance languages: Modal verbs and modal particles. In *Modality and Theory of Mind Elements across Languages*. Edited by Werner Abraham and Elisabeth Leiss. Berlin and Boston: De Gruyter Mouton, pp. 335–60.
- Meisnitzer, Benjamin, and Lukas Müller. 2024. Partículas modais e partículas modalizadoras no Português Europeu numa perspetiva sincrónica e diacrónica. Unpublished manuscript.
- Pereira, Bruna Karla. 2011. A sintaxe cartográfica de ‘lá’ no Português Brasileiro: Um estudo da periferia esquerda. Ph.D. thesis, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
- Pereira, Mário Crescêncio de Jesus. 2009. Aspectos Semânticos e Pragmáticos de Aqui, aí, ali, cá e lá em Português Europeu. Master’s thesis, Faculdade de Letras, Universidade do Porto, Porto, Portugal.
- Potts, Christopher. 2005. *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.
- Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33: 165–97. [CrossRef]
- Potts, Christopher, and Florian Schwarz. 2010. Affective ‘this’. *Linguistic Issues in Language Technology* 3: 1–30. [CrossRef]
- Recanati, François. 2004. Pragmatics and semantics. In *The Handbook of Pragmatics*. Edited by Laurence R. Horn and Gregory Ward. Oxford: Blackwell, pp. 442–62.
- Sambrana, Vania Rosana Mattos. 2021. Construcionalização de marcadores discursivos formados por olhar e ver no português. Ph.D. thesis, Universidade Federal Fluminense, Niterói, Brazil.
- Sbisà, Marina. 2001. Illocutionary force and degrees of strength in language use. *Journal of Pragmatics* 33: 1791–814. [CrossRef]
- Schneider, Stefan. 2010. Mitigation. In *Interpersonal Pragmatics*. Edited by Miriam A. Locher and Sage L. Graham. Berlin and New York: De Gruyter Mouton, pp. 253–69. [CrossRef]
- Schwenter, Scott A. 2003. No and tampoco: A pragmatic distinction in Spanish negation. *Journal of Pragmatics* 35: 999–1030. [CrossRef]
- Schwenter, Scott A. 2016. Some issues in negation in Portuguese. In *The Handbook of Portuguese Linguistics*. Malden: Wiley Online Library.
- Shapiro, Lionel. 2020. Commitment accounts of assertion. In *The Oxford Handbook of Assertion*. Oxford: Oxford University Press. [CrossRef]
- Teixeira, Ana Cláudia Machado. 2015. A construção verbal marcadora discursiva VLocmd: Uma análise funcional centrada no uso. Ph.D. thesis, Universidade Federal Fluminense, Niterói, Brazil.
- Teixeira, José. 2005. De cá para lá e de aqui para aí: Rede de valores semânticos dos marcadores espaciais cá/lá/(acolá) e aqui/aí/ali. In *Estudos em homenagem ao Professor Doutor Mário Vilela*. Edited by Graça Maria de Oliveira e Silva Rio-Torto, Olívia Maria Figueiredo and Fátima Silva. Porto: Faculdade de Letras da Universidade do Porto, pp. 449–60.
- Traugott, Elizabeth. 1989. On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language* 65: 31–55. [CrossRef]
- Waltereit, Richard. 2006. *Abtönung. Zur Pragmatik und historischen Semantik von Modalpartikeln und ihren Funktionalen Äquivalenten in romanischen Sprachen*. Berlin and Boston: Max Niemeyer Verlag.
- Wegener, Heide. 1998. *Zur Grammatikalisierung von Modalpartikeln*. Berlin and New York: Max Niemeyer Verlag, pp. 37–56.
- Wittgenstein, Ludwig. 1922. *Tractatus Logico-Philosophicus*. International Library of Psychology, Philosophy, and Scientific Method. Frankfurt: Suhrkamp.
- Yule, George. 2006. *The Study of Language*, 3rd ed. Cambridge: Cambridge University Press.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

On the Emergence of Portuguese FCI *qualquer*: A Diachronic Perspective

Clara Pinto

Linguistics Research Centre of NOVA University (CLUNL), NOVA University, 1099-032 Lisbon, Portugal; clarapinto@fcs.unl.pt

Abstract: Romance free choice items (FCIs) are frequently pointed out as resulting from the grammaticalization of the relative determiner *qual* ‘which’ and an element derived from a volition verb, such as *querer* ‘want’. Contrary to other Romance FCIs, Portuguese *qualquer* ‘any’ remains understudied, therefore motivating the current research. In this article, I investigate the syntax and semantics of *qualquer*, from a diachronic perspective, based on examples extracted from 13th and 14th century texts. Analysis of contexts of occurrence of *qualquer* showed that, in Old Portuguese, the elements *qual* and *quer* could combine in different configurations, corresponding to different structures. On the one hand, the relative determiner *qual* could combine with a form of the volition verb in *ever* free relative clauses. On the other hand, *qual* and *quer* were also combined in appositive relative clauses, which seem to be at the core of postnominal *qualquer*. However, similar to what is argued for Old Spanish, *qualquer* was also a quantifier-like element, occurring in prenominal position and giving rise to universal interpretations. The different origins of prenominal and postnominal *qualquer* may help explain the different readings in contemporary data.

Keywords: free choice item; indefinite compounds; Old Portuguese; relative clauses; *qualquer*

1. Introduction

The term free choice item (FCI) was first coined by Vendler (1967) to refer to a particular property of the English item *any*: its freedom of choice. The term has been used to refer to items that can express both quantification and indetermination, giving rise to universal and existential readings, as illustrated with the English FCI *any* in (1) and (2), respectively:

- (1) Any student will pass the exam.
- (2) Take any apple from the basket.

In (1), *any student* can be considered equivalent to *every student*, therefore conveying a universal reading, while in (2), *any apple* carries an existential reading. This seems to be a feature of FCIs in general and not exclusive of English *any* and has been a central topic of debate within the semantic analysis of FCIs (cf. Giannakidou 2001).¹

The fact that these items allow both universal and existential readings poses the problem of knowing whether they should be considered universal quantifiers (cf. Dayal 2004) or existential indefinites (cf. Giannakidou 2001).

Despite the considerable amount of literature on the topic, FCIs have been mainly studied from a synchronic point of view and the diachronic perspective is still roughly explored. Nevertheless, the origin and syntactic/semantic features of FCIs in old stages of a language can help shed some light on their synchronic interpretation.

The comparison of some FCIs in Romance languages shows that these items share a common origin: they frequently result from the combination of a relative determiner or pronoun with a verbal form, most likely a volition or a copula verb, in the early stages of the language (cf. Lombard 1938; Haspelmath 1995). This is the case of Spanish items *cualquier(a)* ‘whatever’, *quienquiera* ‘whoever’, and *dondequiera* ‘wherever’ (cf. Rivero 1986,

Citation: Pinto, Clara. 2023. On the Emergence of Portuguese FCI *qualquer*: A Diachronic Perspective. *Languages* 8: 290. <https://doi.org/10.3390/languages8040290>

Academic Editor: Patrícia Amaral

Received: 17 July 2023

Revised: 23 November 2023

Accepted: 28 November 2023

Published: 18 December 2023



Copyright: © 2023 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1988; Company Company and Pozas Loyo 2009; Company Company 2009; Mackenzie 2019; Elvira 2020); Italian *qualsiasi* and *qualunque* ‘whatever’ (cf. Degano and Aloni 2021; Kellert 2021); and Catalan *qualsevol* (cf. Colomina I Castanyer 2002) and Galician *calquer* ‘whatever’ (cf. Ferreiro 1999), among others.

Old Portuguese data show that the relative determiner *qual* ‘which’ was frequently combined with a form of the volition verb *querer* ‘want’ in *ever* free relative clauses. Nevertheless, I argue that this is probably not the direct source of Portuguese FCI *qualquer*.

This paper aims to (i) provide empirical data on early uses of the FCI *qualquer*, while offering a syntactic and semantic description of its properties; (ii) argue against the idea that Portuguese *qualquer* directly results from *ever* free relative clauses with an additional internal head; (iii) put forth the hypothesis that prenominal and postnominal uses of *qualquer* have different origins and emerge in different chronological periods, resulting in the existence of a prenominal *qualquer* with quantifier-like properties and a postnominal *qualquer* with adjectival properties in Old Portuguese.

FCI *qualquer* in Synchrony

Before looking at the properties of *qualquer* in Old Portuguese, I briefly refer to the values and uses associated with the contemporary item *qualquer*, in order to highlight the possible differences regarding diachronic uses of the item.

First of all, it is worth mentioning that contrary to the vast literature surrounding the English FCI *any*, there are not many studies on Portuguese *qualquer*. We highlight the work by Mória (1992a), Peres (1987, 2013), and Moreno (2009), which offer mainly a semantic description of the item; and the works by Pires de Oliveira (2005) and Medeiros (2022), which analyze *qualquer* in the Brazilian variety. All of these works have in common the need to account for the several different interpretations displayed by *qualquer*, according to its position regarding the noun and the combination with other indefinite elements.

Qualquer associates with nouns and is traditionally paired either with indefinite pronouns or with quantifiers due to the different interpretations it may trigger. Nevertheless, as Peres (2013) observes, none of the classifications seems to totally translate the behaviour of *qualquer*.

Let us look at examples (3) and (4):

(3) *Qualquer* criança faz birra quando lhe dizem ‘não’
 any child do.^{3SG.PRES.} tantrum when her.^{3SG.DAT} say.^{3PL.PRES.} ‘no’
 ‘Any child will make a tantrum when told ‘no’.’

(4) Não devias conduzir tão depressa. *Qualquer* dia apanhas
 NEG should.^{3SG.IMP.} drive so fast any day catch.^{2SG.PRES.}
 um susto.
 a fright
 ‘You shouldn’t drive so fast. One of these days you will be given a fright.’

As can be seen from the comparison between the two contexts, *qualquer* can trigger a universal reading as in (3), being interpreted as *every child*. On the other hand, it can also convey an existential reading, as in (4), where it is equivalent to *a day, does not matter which*. This duality of meanings is usually a feature associated with other FCIs, as we have mentioned before.

Apart from the universal and the existential readings, *qualquer* may also trigger other values and combine with other elements within the determiner phrase (DP). Peres (2013) accounts for three different values associated with *qualquer*, namely ‘equivalence’, ‘unknown’, and ‘restriction’ values,² as exemplified below:

- (5) Será que ele tem *quaisquer* hipóteses de vencer (por poucas que sejam)?
 ‘I wonder if he has any chances of winning (no matter how few).’
- (6) Eu já li *qualquer* livro desse autor (não sei qual).
 ‘I already read some book by this author (I don’t know which).’

- (7) O presidente não recebe *qualquer* pessoa.
 ‘The president will not receive anyone.’

(Peres 2013, p. 798)³

The examples above all display *qualquer* in the prenominal position. However, contrary to FCIs such as *any*, *qualquer* may also occur in the postnominal position. In this last configuration, there is usually the presence of the indefinite determiner *um* ‘a’ before the noun and a tendency to favour readings with depreciative flavour as in (8).⁴

- (8) Ela não era uma rapariga *qualquer*.
 ‘She was not an ordinary girl.’

The combination with the indefinite determiner *um* ‘a’ and the values assumed by *qualquer* under such a configuration have motivated the proposal by Mória (1992a), with the distinction between three values for *qualquer*: universal, existential, and cardinal.

Qualquer also combines with the indefinite *outro* ‘other’, allowing different word orders, as illustrated from (9) to (11):

- (9) *Qualquer* outra pessoa teria sido mais simpática.
 ‘Any other person would have been nicer (apart from this one).’
- (10) Outra *qualquer* pessoa teria sido mais simpática.
 ‘Any other person would have been nicer.’⁵
- (11) Outra pessoa *qualquer* teria sido mais simpática.
 ‘Another person, no matter who, would have been nicer.’

The major difference in meaning is found between (9) and (11), showing that prenominal and postnominal *qualquer* do not always produce the same interpretation. While in (9), prenominal *qualquer* can refer to every single person, except a particular one, in (11), the existence of at least one person apart from the one at stake is presupposed. I will not elaborate on the issue here (but cf. Peres 1987 and Mória 1992a for a detailed description of contemporary data).

Qualquer does not occur with absolute pronominal reading, as, for instance, the indefinites *alguém* ‘someone’ or *ninguém* ‘anyone/no one’. Instances such as (12) are considered ungrammatical⁶ in contemporary data (agrammaticality is indicated by *), even though they are registered in 13th and 14th century texts. It can, however, occur in partitive constructions both in prenominal and postnominal positions as in (13) and (14), respectively:

- (12) **Qualquer* que seja corajoso vencerá a batalha
 ‘Whoever that is brave will win the battle.’
- (13) *Qualquer* (um) dos vestidos te fica bem.
 ‘All of the dresses suit you well.’
- (14) Um vestido *qualquer* dos que compraste ontem fica-te bem.
 ‘Any random dress from the ones you bought yesterday suits you well.’

In (13) and (14), *qualquer* occurs with a partitive prepositional phrase (PP). According to Pires de Oliveira (2005), the presence of the partitive construction determines that the set of alternatives underlying the freedom of choice of *qualquer* must be known to the speaker. This type of context lacks investigation and raises several questions regarding the indefinite or quantificational nature of *qualquer*.

Despite the presence of the partitive PP in (13) and (14), there are crucial differences, resulting from the position occupied by *qualquer*. First of all, only in (13) is the presence

of the indefinite *um* ‘a’ optional. Secondly, only in (13) is the partitive PP being directly selected by *qualquer*. Partitive complements are traditionally selected by quantifiers, which would position prenominal *qualquer* as a quantifier-like element.

As far as sentence (14) is concerned, the quantificational reading of *qualquer* in such contexts is frequently associated with the presence of *um*. However, the exact nature of *um* remains undetermined, since we can be in the presence of the indefinite determiner or the cardinal numeral. I am inclined to consider that the element *um* that combines with *qualquer* is an indefinite determiner instead of the cardinal element. The reason why I argue in this direction is related to the agrammaticality of contexts such as (16) and (17) where the adverb of exclusion *só* ‘only’ or the (prenominal) adjective *único* ‘single’ force a cardinal interpretation of *um*.⁷ If *um* was interpreted as a cardinal, we would expect these contexts to be felicitous, but they are not, suggesting that *um* is the indefinite determiner. Furthermore, cardinal *um* is incompatible with the idea of freedom of choice, since it is impossible to choose if the set only contains one element.

- | | | | | |
|------|---------------------------------|----------|-----------------|--------------------------|
| (15) | Escolhe
Choose | uma
a | maçã
apple | <i>qualquer</i> .
any |
| | ‘Choose any apple.’ | | | |
| (16) | *Escolhe
Choose | uma
a | só
only | maçã
apple |
| | *‘Choose just one any apple.’ | | | |
| (17) | *Escolhe
Choose | uma
a | única
single | maçã
apple |
| | *‘Choose one single any apple.’ | | | |

Finally, one last note on *qualquer* is related to the existence of the plural form *quaisquer*. The morpheme -s marking plural is still added after *qual* and not at the end of the word.⁸ This fact is still a reminder of the compositional nature of *qualquer*, as will be shown in the next sections.

2. Materials and Methods

In this section, I present a few considerations concerning the sources and methodology on which the present work relies.

The data under analysis are circumscribed to the chronological period corresponding to Old Portuguese, which comprehends the 13th and 14th centuries, roughly following the periodization proposal for Portuguese by Cintra (cf. Castro 1999).⁹ This short timespan seems to be crucial for the development of the FCI *qualquer*, since it is only during this period that we find different configurations of the construction involving the relative *qual* and a form of the volition verb *querer* ‘want’.

In order to constitute a sample corpus, searches were performed semi-automatically, by searching the word *qual* and extracting only the relevant examples. All sentences containing a form of *qualquer* were then inserted in a database, using the program *FileMaker Pro 12 Advanced*,¹⁰ and they were annotated with relevant information, such as the order of the elements in the compound and their position in relation to the nominal element; the presence of modifiers; tense of the verbal form *querer* ‘want’; and other relevant features. The encoding of the examples and the annotation of relevant parameters allowed an easier comparison of the contexts.

As far as the textual sources are concerned, for the 13th century sample, I have considered the following texts:

Demanda do Santo Graal (DSG)—the full version of the edition by Piel and Nunes (1988), in an electronic format;

Foro Real (FR)—the full version of the edition by Ferreira (1987), available online through the corpus CIPM (cf. Xavier 1993–2003);¹¹

Legal documents edited by Martins (2001) in Documentos Portugueses do Noroeste e da Região de Lisboa (DPNRL);

Medieval Galician-Portuguese poetry (GP-poetry), in the edition compiled by Brea (1996), and available through the TMILG¹² corpus platform (cf. Varela Barreiro 2004).

For the 14th century sample, I have chosen the sources below:

Crónica Geral de Espanha (CGE)—the full version of the editions by Pedrosa (2012) and Miranda (2013), as part of their masters’ thesis;

Diálogos de São Gregório (DG)—the full text of the electronic edition by Machado Filho (2013);

Dos Costumes de Santarém (DCS)—the texts written between 1340–1360, in the edition by Rodrigues (1992), available online through the corpus CIPM.

Due to the scarcity of sources for Old Portuguese, I have considered some texts which have been transmitted by later copies. That is the case of *Demanda do Santo Graal* (DSG), *Crónica Geral de Espanha* (CGE), and *Medieval Portuguese-Galician poetry* (GP-poetry) on which some clarifications should be added.

Starting with the DSG text, it corresponds to a 15th century copy of an allegedly early 13th century translation from French. Despite the dating issues, I have considered it to be representative of 13th century Portuguese, based on the works by Castro (1993), Toledo Neto (2012), Martins (2013), and Pinto (2021), among others.

As far as the CGE text is concerned, the edition used is based on manuscript L, which is from the first quarter of the 15th century (cf. Cintra 1951–1990) and closer to the original text from 1344 (the original manuscript, called manuscript Y by Cintra (1951–1990), was lost).

Finally, Medieval Galician-Portuguese poetry is transmitted by three manuscripts, two of which are from the end of the 15th or beginning of the 16th century (*Cancioneiro da Vaticana* and *Cancioneiro da Biblioteca Nacional*) and one from the end of the 13th century (*Cancioneiro da Ajuda*). The edition used by the corpus TMILG, and which we have consulted here, is based on the three manuscripts. Despite the chronology of the manuscript, they are said to reflect 13th century Portuguese.

Table 1 shows the estimated number of words contained in each text and the total number of words per century.

Table 1. Number of words per century and text.

Text	13th Century					14th Century			Total	
	DSG	GP-Poetry	DPNRL	FR	Total	CGE	DG	DCS		
Number of words	212,145	?	29,847	51,022	293,014	403,580	106,166	33,690	543,436	836,450
Forms of <i>qualquer</i> (in percentage)	0.008%	-	0.04%	0.098%		0.008%	0.014%	0.08%		

As one can see, the total number of words is higher in the 14th century, due to the nature of the sample, which contains a very long text (the CGE text). On the other hand, textual sources for the 13th century are not abundant, resulting in a lower number of words. I was not able to determine the number of words for the 13th century part constituted by Galician-Portuguese poetry. Searches were performed using the corpus TMILG, which contains the edition compiled by Brea (1996), but to which there is no indication of the exact number of words.

I have collected the occurrences where *qualquer* takes the exact same form as in contemporary Portuguese,¹³ but I have also considered the occurrences displaying (i) the elements *qual* and *quer* in adjacency but graphically separated (that is *qual quer*); (ii) the elements *qual* and *quer* separated by a lexical element (as in *qual X quer*); and (iii) the elements *qual* and *quer* with the previous two configurations, but with the verbal element displaying variable inflection (as in *qual quiser* or *qual X quiser*).

As Table 1 also shows, the FCI *qualquer* is not a frequent item in any of the texts, being, however, more expressive in the text of *Foro Real* (FR), representing 0.098% of all the words in the document.

The corpus has a total of 166 occurrences of *qualquer*, with 90 belonging to the 13th and 76 to the 14th century. It should be noticed that, despite having a higher number of total words in the 14th century, the number of occurrences found for *qualquer* is lower than the one for the 13th century¹⁴.

In the next sections, I present the data collected for medieval *qualquer*. The description of some particular syntactic features of *qualquer* is made under a generative grammar perspective. I very briefly refer to classical projections, such as determiner phrase (DP) (cf. Abney 1987), complementizer phrase (CP) (cf. Rizzi 1997), and quantifier phrase (QP) (Cardinaletti and Giusti 1992).

3. *Qualquer* in Medieval Portuguese

3.1. General Distribution and Patterns of Occurrence

The data collected shows that medieval *qualquer* displayed different behaviour from the contemporary item, being able to occur in some syntactic configurations that seem to have been lost after the 14th century.¹⁵

Looking at the data, we identify three main configurations for *qualquer*, in terms of word order. The first one corresponds to *qualquer* preceding a nominal element and which we call prenominal. The second configuration presents *qualquer* following a nominal element, therefore in postnominal position. The third configuration presents the two elements *qual* and *quer* separated by a lexical item, which, in most cases, is a noun. These cases are illustrated from (18) to (20), respectively. To these three patterns, we add a pronominal use, therefore without the presence of any nominal element, as in (21).

- (18) [...] e rogamos a qualquer Tabellion que esta carta ujr
 and ask.^{1PL.PRES} to any notary that this letter see
 que faça ende a carta da dita partiçõ.
 that do.^{3SG.PRES.SUBJ} of.that the letter of.the said division
 ‘and we ask any notary who sees this legal document that writes the legal document of the aforementioned division.’
 (DPNRL)
- (19) [...] que nenhua das pessoas sobredictas
 defendemos
 prohibit.^{1PL.PRES} that none of.the persons aforementioned
 nõ possa meter a juyzo nenhua villa
 NEG can.^{3SG.PRES.SUBJ} put to judgement none village
 nen castello nen outro herdamẽto qualquer
 nor castle nor other property any
 ‘we prohibit that none of the aforementioned people can put under trial any village, nor castle, nor any other inheritance.’
 (*Foro Real*)
- (20) [...] e manda seu cavallo a qual parte quer
 and send.^{3SG.PRES} your horse to which part want
 pello freo e o faz star quando quer
 by.the bridle and it make.^{3SG.PRES} be when want.^{3SG.PRES}
 ‘and sends his horse where he wants by the bridle and makes it stand when he wants.’
 (*Demanda do Santo Graal*)

(21) E por esto maldicto he qualquer que
 and by this cursed be.^{3SG.PRES} anyone that
 treiçom faz, ca des ally adiante nunca
 betrayal do.^{3SG.PRES} because since there forward never
 se nê huu quer chamar do seu linhagem,
 se.^{REFLX} no one want.^{3SG.PRES} call of.the his lineage
 assy como foy deste.
 this.way as be.^{3SG.PAST} of.this
 ‘And for this reason, anyone who commits betrayal is cursed, because from that moment afterwards, no one wants to be called from his lineage, as it happened with him.’

(Crónica Geral de Espanha)

These configurations reflect different syntactic structures and may be assigned to three different groups, which are adopted from this moment on. I refer to prenominal *qualquer*, where pronominal uses are included; postnominal *qualquer*, which includes instances of an already lexicalized *qualquer* modifying a nominal element; and finally, I refer to relative *qualquer* to account for the cases where the underlying structure is still a relative clause. This group includes all the examples of discontinuous *qual* and a verbal form of *querer* (which we identify as *ever* free relative clauses and appositive relative clauses). Table 2 presents the distribution of each pattern in terms of number of occurrences, as well as in percentage.

Table 2. Distribution of occurrences of *qualquer* by source and century.

	13th Century									
	DSG	%	GP-Poetry	%	DPNRL	%	FR	%	TOTAL	%
postnominal <i>qualquer</i>	0	0	2	17	0	0	19	38	21	23
prenominal <i>qualquer</i>	5	31	2	17	12	100	19	38	38	42
relative <i>qualquer</i>	11	69	8	67	0	0	12	24	31	34
Total	16		12		12		50		90	
	14th Century							TOTAL		
	CGE	%	DG	%	DCS	%	TOTAL	%	13th + 14th	%
postnominal <i>qualquer</i>	1	3	8	53	12	44	21	28	42	25
prenominal <i>qualquer</i>	30	88	4	27	11	41	45	59	83	50
relative <i>qualquer</i>	3	9	3	20	4	15	10	13	41	25
Total	34		15		27		76		166	

When we compare the frequency of each pattern in the two centuries, we see that there are some changes from the 13th to the 14th century, as illustrated in Figure 1.

Figure 1 shows that prenominal *qualquer* was the most frequent configuration in 13th century data, followed by relative clauses with the elements *qual* and *quer*. However, while occurrences of prenominal *qualquer* continued to increase in the following century, the relative clause configuration decreased. Finally, there was also an increase in postnominal *qualquer* (N *qualquer*) from the 13th to the 14th century, although it is not as accentuated as in the previous configurations.

The frequencies presented above show that relative clauses with *qual* and *quer* started declining after the 13th century and *ever* free relatives disappeared from the language after the medieval period. On the other hand, prenominal and postnominal occurrences become the widespread patterns, probably filling in the gap left by the disappearance of the relative clause pattern.

In the following sections, I argue in favour of the existence of two items *qualquer* in Old Portuguese. There was a use of *qual* and *quer* associated with relative clauses (both *ever*

free and appositive relatives) and displaying different levels of grammaticalization; and there was also a specifier *qualquer*, which already behaved as an independent constituent (even though it may have originated in a relative clause in a much earlier period).

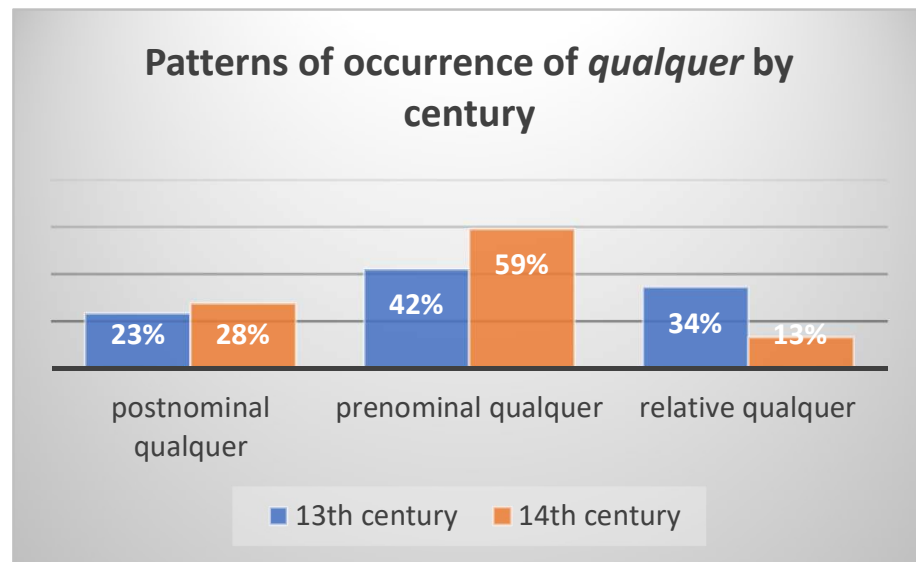


Figure 1. Patterns of occurrence of *qualquer* by century.

3.2. Relative *Qualquer*

Medieval Romance FCIs are said to originate in relative clauses (cf. Rivero 1984, 1988; Haspelmath 1995; Company Company 2009). Old Portuguese also displayed relative clauses involving the relative *qual* and a verbal form of *querer*, both in adjacency and in a discontinuous configuration. In this section, we look at occurrences of *qualquer* that correspond to instances of relative clauses.

The relative determiner *qual* is said to originate in the Latin form *QUALIS*, which was used as a *wh*-element in interrogative and exclamative clauses, but also participated in correlative constructions with the form *QUALIS...TALIS* (cf. Ernout and Thomas 1972, p. 156). Its use as a relative element is not registered in Latin, though. It has, therefore, been considered a Romance innovation, but this is a question still open to debate, since some authors situate its emergence already in Latin (cf. Ramat 2005).

In Old Portuguese, *qual* is registered in the corresponding contexts listed for Latin *QUALIS* and as a relative determiner/pronoun, introducing relative clauses.

Data from our corpus attests the possibility in Old Portuguese, with *qual* being a relative determiner in combination with a form of the volition verb *querer* ‘want’, as illustrated in (22) and (23) below:

(22)	«Certas certainly que that mim me comecei start. ^{1SG.PAST}	gram great seja», be. ^{3SG.PRES.SUBJ} teer have	folia amusement disse say. ^{3SG.PAST} me me	buscades.» search. ^{2PL.PRES} Ivam Ivam convem suit. ^{3SG.PRES}	« <i>Qual</i> which o the pois because	folia amusement Bastardo, Illegitimate que that	<i>quer</i> want. ^{3SG.PRES} «a to o it
------	---	---	---	--	---	--	---

‘You certainly search for great amusement. Whichever amusement that is, said Ivam, the Illegitimate, having it suits me since I have started it.’

(*Demanda do Santo Graal*)

(23)	mais but quiseres want. ^{2SG.FUT.SUBJ} 'but choose any death you want and we will give it to you'	escolhi choose:2SG.IMP e and	tu you darch'a-emos give. ^{1PL.FUT_it.ACC}	huma a	morte death	qual which
------	--	---------------------------------------	--	-----------	----------------	---------------

(Diálogos de São Gregório)

In (22), the element *qual* introduces a relative clause and combines with a verbal form *quer*, but the two items are separated by lexical material, namely the noun *folia* 'amusement'. In fact, *qual folia* 'which amusement' corresponds to the head (that is, a head and an additional internal head) of the relative clause, while *quer* is the verbal form, initially selecting a clausal complement introduced by *que* 'that'. This relative clause seems to be a free relative clause since there is no lexical antecedent.

On the other hand, in (23), the relative determiner *qual* is immediately followed by the verbal form *quer*, but contrary to (22), there is a nominal antecedent, which indicates that this is an appositive relative clause, modifying the noun. The two examples show us that *qual* and *quer* could combine in two types of relative clauses—free relatives and appositives.

Let us start by looking at free relative clauses. There are two different types of free relatives described in the literature: plain free relatives and *ever* free relatives. While it is generally assumed that plain free relatives have a definite interpretation, *ever* free relatives are associated with universal readings (cf. Dayal 1997). Examples (24) and (25), taken from Dayal (1997, p. 99), illustrate the two types of free relatives:

- (24) I ordered what he ordered for dessert. (=the thing he ordered for dessert);
- (25) John will read whatever Bill assigns. (=everything/anything Bill assigns).

The plain free relative in (24) produces an interpretation similar to a definite determiner phrase (DP), while the *ever* free relative in (25) has a prototypical universal reading.¹⁶

It is not my goal to investigate the semantics of plain and *ever* free relatives here (cf. Šimík 2020 for a semantic analysis), but a few comparative considerations should be made regarding *ever* free relatives, due to their parallel with free relatives with volition verbs in Romance languages, such as Portuguese. In Portuguese, the equivalent structure of *ever* free relative clauses involves the presence of an element originating from a volition verb meaning *want*,¹⁷ that is, the case of the element *quer*, which results from the third person singular form of the verb *querer* 'want' in the Simple Present Indicative, as in (26b). The pairs in (26) and (27) illustrate the differences between a plain free relative and a relative with *quer* in Portuguese.

(26)	a.	Quem who	vier come: ^{3SG.FUT-SUBJ}	à to.the	festa party	irá will	divertir-se have.fun		
		'Who comes to the party will have fun.'							
	b.	Quem who	<i>quer</i> <i>ever</i>	que that	venha come. ^{3SG.PRES.SUBJ}	à to.the	festa party	irá will	divertir-se have.fun
		'Whoever comes to the party will have fun.'							
(27)	a.	Vou go. ^{1SG.PRES}	contigo with.you	onde where	fores. go. ^{2SG.FUT}				
		'I will go with you where you go.'							
	b.	Vou go. ^{1SG.PRES}	contigo with.you	onde where	<i>quer</i> <i>ever</i>	que that	vás. go. ^{2SG.PRES.SUBJ}		
		'I will go with you wherever you go.'							

Although the relative clauses in each pair (26) and (27) may refer to the same entity/place, there are differences in meaning, as well as in the syntax, with free relatives with *quer* being modified by a restrictive relative clause with the subjunctive mood.

Ever free relatives can be considered semantically equivalent to free relatives with *quer* 'ever' in Portuguese, as seems clear from the comparison between the pairs in (28) and (29).

- (28) a. I will do whatever the teacher asks me.
 b. I will do anything/everything the teacher asks me.

- (29) a. Farei o que quer que a professora peça.
 do.^{1SG.FUT} the what ever that the teacher ask.^{3SG.PRES.SUBJ}
- b. Farei qualquer coisa/tudo o que a professora peça.
 do.^{1SG.FUT} any thing/ everything the that the teacher ask.^{3SG.PRES.SUBJ}

As shown in (28b) and (29b), it is possible to assume both a universal and an existential reading for the two sentences since the relevant strings ‘whatever’/‘o que quer’ can be replaced by *anything/qualquer coisa*, activating a free choice reading or by *everything/tudo*, which corresponds to universal quantification.

This seems to show that free relatives with *quer* are parallel to *ever* free relatives. Therefore, I adopt the term *ever* free relative to refer to free relatives combining the relative determiner *qual* and the particle *quer* resulting from a volition verb, whenever there is no nominal antecedent. This type of free relative clause is also known as a non-specific free relative (cf. Haspelmath 1995).

3.2.1. Ever Free Relative Clauses in Old Portuguese

Ever free relative clauses with a volition verb are frequently found in Old Portuguese texts as a strategy to introduce non-specific or indefinite references.¹⁸ Relative elements in *ever* free relatives can refer to [+/-human] or [+/-animate] entities, but they can also have a [+locative], [+temporal], or a [+manner] reading. Examples (30) to (35) illustrate these possibilities, with the following relative elements: *quẽ* ‘who’, *que* ‘what’, *u* ‘where’, *quando* ‘when’, *como* ‘how’, and *qual* ‘which’ in Old Portuguese:

- (30) E *quẽ* *quer* que contra isto ueer ou
 and who want.^{3SG.PRES} that against this see.^{3SG.FUT.SUBJ} or
 fazer algũa cousa moyra porende e nõ seya
 do some thing die for.that and NEG be.^{3SG.PRES.SUBJ}
 leyxado uiuo.
 left alive
 ‘And whoever sees or does something against this, must die for it and not be left alive.’

(*Foro Real*)

- (31) Mais nom me chal, que quer que me
 more NEG me.^{1SG.DAT} heat what want.^{3SG.PRES} that me.^{1SG.DAT}
 avenha desta batalha, ca ataa aqui ouve
 come.^{3SG.PRES.SUBJ} of.this battle because until here have.^{3SG.PAST}
 ende a honra e vos a desonra.
 of.that the honour and you the the dishonour
 ‘But it doesn’t matter whatever comes to me from that battle because until now I only had the honour and you the dishonour.’

(*Demanda do Santo Graal*)

- (32) «Nom», disse el, «mas Deos os guarde
 No say.^{3SG.PAST} he but God them protect.^{3SG.PRES.SUBJ}
 todos, u quer que elles sejam!»
 all where want.^{3SG.PRES.} that they be.^{3PL.PRES.SUBJ.}
 ‘No, he said, but God protects them all, wherever they are!’

(*Demanda do Santo Graal*)

- (33) E todos (co)munalmente seyã teodos de fazerlhy
 and all communally be.^{3PL.PRES.SUBJ} have.^{Past.Part.} of do.him.^{3SG.DAT}
 menagẽ a el ou a quẽ el mandar
 homage to he or to who he send.^{3SG.FUT.SUBJ}
 en seu logo quando quer que mãe.
 in his place when want.^{3SG.PRES.} that order.^{3SG.PRES.SUBJ}
 ‘And all should pay him homage, to him or to whom he sends on his behalf, whenever he orders.’

(Foro Real)

- (34) [...] a [ey]greia receba todo o seu como
 the church receive.^{3SG.PRES.SUBJ} all the his as
 quer que seya achado
 want.^{3SG.PRES.} that be.^{3PL.PRES.SUBJ} found.
 ‘may the church receive all that belongs to it, however it is found.’

(Foro Real)

- (35) [...] outorga-me que a minha alma seja
 grant.me.^{1SG.DAT} that the my soul be.^{3SG.PRES.SUBJ}
 com a sua de pos minha morte
 with the your of after my death
 e de pos a sua em qual
 and of after the your in which
 lugar quer que el seja
 place want.^{3SG.PRES.} that it be.^{3SG.PRES.SUBJ}
 ‘and grant me that my soul be with hers after our deaths, whichever place it might be.’

(Demanda do Santo Graal)

As is visible in the examples above, these *ever* free relatives display a universal/existential interpretation due to the maximality effect observed for free relatives (cf. Jakobson 1995).

Ever free relatives with *querer* were frequent in Old Portuguese and were kept in the language with all relative items (*quem, onde, quando, o que*), as attested by examples (36) to (40).¹⁹ The exception is the relative *qual*, which is ungrammatical in constructions such as (41a), which were attested in Old Portuguese. This is so because in CEP, *qual* cannot occur as a relative determiner taking an internal head.²⁰ It is, however, possible to have a context such as (41b), but in this case, the relevant constituent is no longer a relative element, but the FCI *qualquer*.

- (36) Quem quer que use este vestido ficará ridículo
 who want.^{3SG.PRES} that wear.^{3SG.PRES.SUBJ} this dress be.^{3SG.FUT} ridiculous
 ‘Whoever wears this dress will be ridiculous.’
- (37) Onde quer que vás, irei contigo.
 where want.^{3SG.PRES} that go.^{2SG.PRES.SUBJ} go.^{1SG.FUT} with.you
 ‘Wherever you go, I will go with you’
- (38) Esperarei por ti, quando quer que venhas.
 wait.^{1SG.FUT} for you when want.^{3SG.PRES} that come.^{2sg.Pres.Subj}
 ‘I will wait for you whenever you come.’
- (39) O que quer que digam não é verdade.
 the what want.^{3SG.PRES} that say.^{3PL.PRES.SUBJ} NEG be.^{3SG.PRES} truth
 ‘Whatever they say, it is not truth.’
- (40) Venderei o carro como quer que esteja.
 sell.^{1SG.FUT} the car how want.^{3SG.PRES} that be.^{3SG.SUBJ}
 ‘I will sell the car how ever it is’.

- (41a) *Qual problema *quer* que seja, será resolvido
 what problem want.^{3SG.PRES} that be.^{3SG.PRES.SUBJ} be.^{3SG.FUT} solved
 'Whatever problem it is, it will be solved.'
- (41b) Qualquer que seja o problema, será resolvido
 whatever that be.^{3SG.PRES.SUBJ} the problem be.^{3SG.FUT} solved
 'Whatever the problem is, it will be solved.'

An anonymous reviewer called attention to the possibility of *ever* free relatives with *qual* involving a referentially vague noun to have competed with other relative items. Although we find different nouns in *ever* free relatives, Table 3 shows that some nouns with generic interpretation appeared more often.

Table 3. Nouns occurring between *qual* and *quer* in *ever* free relative.

andança 'wandering'	clérigo 'cleric'	cousa 'thing'	desventura 'misfortune'	dona 'lady'	feito 'deed'	folia 'amusement'	guisa 'manner'	homem 'man'	hora 'hour'	judeu 'jew'	juiz 'judge'	justiça 'justice'	lugar 'place'	ordem 'order'	parte 'part'	pecador 'sinner'	tempo 'time'	terra 'land'	vilania 'vilany'	total
1	1	4	1	1	1	1	2	1	2	1	2	1	1	2	1	1	3	2	1	30

For instance, occurrences of *qual coisa quer* 'which thing want' or *qual tempo quer* 'which time want' may be considered equivalent to *o que quer* 'what (you) want' and *quando quer* 'when you want', respectively.

Although there are no grammaticalized forms involving *ever* free relatives headed by other relatives in Portuguese, in such contexts, the volition verb is not interpreted as a full lexical verb anymore. It seems to correspond to what Haspelmath (1995) called an *indefiniteness marker* since it occurs under the frozen form *quer* at all times and it is emptied from its original lexical meaning.²¹ Contrary to Portuguese, in other Romance languages, such as Spanish, we find grammaticalized forms such as *cualquier*, as well as also *quienquier*.

In fact, Romance FCIs from the WH-*quer* series have similarities with WH-*ever* FCIs in English, or with WH-*immer* constructions in German, showing that the emergence of FCIs from relative constructions is a much broader phenomenon. However, unlike Portuguese *qualquer* and its Romance cognates, English WH-*ever* FCIs keep their clausal status, not being able to take an NP argument (cf. Giannakidou and Cheng 2006).²²

3.2.2. Ever Free Relative Clauses with *Qual* and *Quer*

In (42), we find *qual* introducing an *ever* free relative clause, in association with a form of the volition verb *querer* 'want',

- (42) [...] devo encobrir a todo meu poder minha
 should hide to all my power my
 catividade, *qual* pecador *quer* que eu seja.
 captivity which sinner want.^{3SG.PRES} that I be.^{1SG.PRES.SUBJ}
 'I should hide my captivity by all means, whichever sinner I may be.'

(*Demanda do Santo Graal*)

Ever free relative clauses with *qual* distinguished themselves from similar relatives headed by other elements due to the possibility of selecting a nominal additional internal head. Free relative clauses are usually considered headless relatives since they do not have a lexical antecedent.²³ In an *ever* free relative like (42) above, *qual* is a relative determiner

followed by a nominal internal head—the noun *pecador* ‘sinner’. The sequence *qual pecador* ‘which sinner’ can then be considered a *wh*-phrase, in the sense of Caponigro (2019).

Sentences such as (43) below indicate that variable verbal inflection was possible in earlier uses of the construction, confirming its clausal status.

- (43) -Vai per teu conto a qual terra quiseres [...]
 go.^{2SG.IMP} by your tale to which land want.^{2SG.FUT.SUBJ}
 ‘Go by your means to whatever land you want’
 (Diálogos de São Gregório)

Despite the high frequency of *ever* free relative clauses in 13th century corpus, data already point to the ongoing grammaticalization of the volition verb *querer* into an indefiniteness marker (cf. Haspelmath 1995). In (43), the volition verb is inflected in the second person, Future Subjunctive,²⁴ but the majority of the examples in the corpus already display the fixed form *quer*, which may have been ambiguous during this period between a verbal form and non-verbal marker expressing indefiniteness.

For instance, in (44), the form *quer* that follows the noun *hora* ‘hour’ can no longer be interpreted as the lexical verb. *Hora* ‘hour’ cannot be the subject or object of *quer*; in other words, there are no arguments of *quer* in this sentence. In (44), it seems that *qual hora quer* is being interpreted as a nominal constituent with a free choice reading, followed by a restrictive relative clause. The form *quer* functions as an indefiniteness marker, rather than a full lexical verb, not selecting a complement.²⁵

- (44) Mays qual hora quer que sabhia dalguu
 but which hour want.^{3SG.PRES} that know.^{3SG:PRES.SUBJ} of.some
 erege logo o faça a saber ao
 heretic soon it.^{ACC} do.^{3SG.PRES.SUBJ} to know to.the
 bispo da terra
 bishop of.the land
 ‘but at whatever time you learn of a heretic, tell it to the bishop of that land right away.’
 (Foro Real)

This takes us to another particular feature of *ever* free relative clauses with *qual* and the volition verb—the frequent presence of *que* clauses after the verb, as in (45):

- (45) «qual vilani(a) quer que eu faza i
 which villainy want.^{3SG.PRES} that I do.^{1SG.PRES.SUBJ} there
 contra vos, a justar vos convem ou
 against you.^{2PL-ACC} to fight you.^{2PL-DAT} suit.^{3SG.PRES.} or
 queirades ou nom.»
 want.^{2PL.PRES.SUBJ} or NEG
 ‘Whatever villainy I do against you, you should fight it, whether you want it or not.’
 (Demanda do Santo Graal)

The nature of these *que* clauses is not consensual due to the initial ambiguity between a complement clause of the volition verb and a restrictive relative clause introduced by a relative pronoun. In early examples, the exact nature of the element *quer* (still a verb or already an indefiniteness marker) may not be straightforward.

Based on Old Spanish data, both Rivero (1988) and Company Company (2009) consider these *que* clauses to be dominantly restrictive relative clauses already in the medieval period. However, a different analysis is suggested by Mackenzie (2019), with the interpretation of *que* clauses still as complement clauses selected by the volition verb. Mackenzie (2019, p. 195) considers that contexts with a *que* clause represent «a violation of Chomsky and Lasnik’s (1977) ‘doubly-filled Comp’ filter, a constraint requiring the complementizer to be silent if an overt *wh*-word is also present in the same area of clause structure». I will comment on this later.

Even though *que* clauses have allegedly started as complement clauses of the volition verb, data from our corpus fails to attest this construction, since sentences displaying the volition verb with different inflection from the ambiguous third person singular Present tense *quer* do not occur with a *que* clause.

Additionally, the large majority, if not all, of the examples with a *que* clause seem to favour its interpretation as a restrictive relative. This aligns with the idea that the verbal form *quer* was losing its lexical properties and becoming an indefiniteness marker. Sentences such as (46) below rule out the complement clause interpretation.

(46)	Porque	os	comendadores de	qual	ordī	<i>quer</i>
	because	the	commanders of	which	order	want. ^{3SG.PRES}
	que	sō	postos enas	baylias	nō	poden
	that	be. ^{3PL.PRES.IND}	put in.the	<i>bailia</i> ²⁶	NEG	can. ^{3PL.PRES}
	auer	seus	mayores pera	demandar	seus	dereytos
	have	their	superiors to	demand	their	rights
	sobellas	cousas	que perteeçen	as	baylias	
	over.the	things	that belong	to.the	bailias	

‘Because the commanders of whatever order who are assigned for the *bailias* cannot have their superiors to demand their rights over things belonging to the *bailias*.’

(*Foro Real*)

In (46), the *que* clause is a restrictive relative clause modifying the noun *comendadores* ‘commanders’ (i.e., it restricts the set of commanders to the subset of those who are assigned to the *bailias*). In this particular context, the form *quer* is no longer interpreted as the volition verb, therefore not selecting a complement anymore. Furthermore, if the *que* clause was a complement clause of *querer*, we would expect the main verb to be in the subjunctive mood, as in sentence (45) above. However, ‘*sō postos*’ displays an indicative mood.

It seems that in the 13th century, *ever* free relative clauses were no longer unambiguously clausal instances since in most, if not all, of the examples, the volition verb is not fully behaving as a lexical verb anymore. The grammaticalization of the volition verb into an indefiniteness marker could have been the trigger for the reanalysis of *ever* free relative clauses such as the FCI *qualquer*. However, this proposal faces some challenges as far as Portuguese data are concerned.

In their analysis of Old Spanish data, Company Company and Pozas Loyo (2009) propose a three-step grammaticalization path for FCI *cualquier*. The authors consider that the first stage consisted of an *ever* free relative clause with an additional internal head, like the one presented in (46) for Portuguese. In a second stage, the free relative would occupy a prenominal position, but with the nominal element remaining *in situ* (as in *cual quiera castigo* ‘which want.^{3SG.PRES.SUBJ} punishment’), until it was reanalyzed as a non-clausal element, therefore reaching the third stage.

Also referring to Old Spanish, we find the proposal by Mackenzie (2019), who gives as an example of an *ever* free relative clause with *qual* and a form of the volition verb, the context presented in (47).

(47)	que por [CP de quales quier malas costumbres que ell omne sea].
	‘... that whatever bad habits a man might be prone to [. . .]’

(Mackenzie 2019, p. 194, *General estoria* I, fol. 272v)

At this point, I would like to address the problem of the violation of the ‘doubly-filled Comp’ filter introduced by Mackenzie (2019) that has been referred to earlier in this section. The context presented below in (47) is given by Mackenzie (2019) as an example of the violation of the ‘doubly-filled Comp’ filter involving *ever* free relatives with *qual* and a volition verb in Old Spanish. I consider that there is no violation of the ‘doubly-filled Comp’ filter here since *quales quier* does not correspond to an *ever* free relative but to a specifier of the noun, as we will see later on.

The example in (47) raises, however, a crucial question relative to the emergence of an independent item *qualquer*. Is it possible that *ever* free relative clauses with *qual* and *quer* taking an additional nominal head gave place to both prenominal and postnominal *qualquer*? This has been the evolution initially proposed by Cuervo (1893) and which has been followed by some authors (cf. Company Company and Pozas Loyo 2009) but rejected by others (cf. Elvira 2020) for Old Spanish.

Let us assume that the starting point of *qualquer* in (47) was an *ever* free relative clause like (46), with *qual* selecting an additional internal head. This configuration would determine that two relevant elements—*qual* and *quer*—would have first been separated by a nominal item, as in *qual maneyra quer* ‘which manner wants’. The presence of the nominal element between the relative determiner and the verbal form/indefiniteness marker would block the adjacency required for reanalysis. The nominal element could not be interpreted as the additional internal head anymore since it would occur after the verbal form, and therefore already under inflection phrase (IP), as the hypothetical representation in (48) illustrates.

(48) [DP [CP [C *qual* [IP *quer* [NP *maneyra*]]]]

Following Company Company and Pozas Loyo (2009), Kellert (2021, p. 17) considers that a configuration such as (48) corresponds to a relative clause with the NP *in situ*, which would be the second stage of grammaticalization of Spanish *cualquier(a)* and Italian *qualunque*. Although such a hypothesis should not be ruled out for Portuguese, I found no empirical evidence in the data to sustain such a stage.²⁷ Even though split DPs are registered in Old Portuguese (cf. Martins 2004), this configuration seems to apply mainly to modifiers and not to the splitting of the relative determiner and the additional internal head (cf. Cardoso 2011), as would be the case for (48). Furthermore, for the same chronological period, I did not find cases of NP *in situ* with the only other relative determiner taking an additional internal head: the relative *quanto(s)* ‘how.many/much’. Finally, *ever* free relative clauses with the NP *in situ* seem incompatible with the cases where *qualquer* combines with the indefinite element *outro* ‘other’ to its left as in *outro qualquer N*.²⁸ We look at these examples further on.

Germanic constructions with WH-*immer* seem to parallel *ever* free relatives with *qual*, due to the presence of an additional internal head, despite the non-verbal origin of *immer* ‘ever’. According to Bossuyt and Leuschner (2020, p. 207), WH-*immer* constructions in German are still not grammaticalized due to the impossibility of splitting the complex WH formed by the relative *welcher* ‘which’ and the nominal element (*welches *immer Buch* ‘whichever book’), similar to what we saw in *ever* free relatives with *qual*.

So far, we have argued that a merge of *qual* and *quer* seems unlikely due to the presence of an additional internal head. However, it is also relevant that *ever* free relative clauses with *qual* and *quer* ceased to be available after the 14th century. Elvira (2020) claims that the relative *qual* disappears from Old Spanish and that is the reason why *ever* free relatives with *qual* cease to occur. The same explanation fits the Portuguese case. As we have seen previously, *ever* free relative clauses existed in Old Portuguese introduced by different relative items. They all continue to occur in CEP, except for the ones introduced by *qual*. As such, it is not the case that the paradigm of *ever* free relatives disappeared or changed, but only that *qual* stopped being available. In fact, all instances of bare *qual* as a relative element have disappeared from the language. Only the relative pronoun *o-qual* is kept, but contrary to what was verified in Old Portuguese, it stops occurring with an additional internal head (cf. Cardoso 2008, 2011)²⁹.

In the next subsection, we look at appositive relative clauses, which were another clausal context of occurrence of *qual* and *quer*. Appositive relative clauses seem more likely to have favoured the reanalysis of the two elements.

3.2.3. Appositive Relative Clauses with *Qual* and *Quer*

Apart from *ever* free relative clauses, the relative determiner *qual* also combines with a form of the volition verb *querer* in appositive relative clauses as the one illustrated in (49):

(49)	Custume	h(e)	do	peõ	q(ue)	uêde	o
	Custom	be. ^{3SG.PRES}	of.the	peasant	that	sell. ^{3SG.PRES}	the
	vio	da	jugada	q(ue)	deue	a	el
	wine	of.the	tax	that	owe. ^{3SG.PRES}	to	the
	Rey	a	dar	q(ue)	en	pod(er)	seia
	king	to	give	that	in	power	be. ^{3SG.PRES.SUBJ}
	do	íugadeyro	de	demãdar	o	vinho	ou
	of.the	land.owner	to	demand	the	wine	or
	os	dín(hei)r(o)s	<i>qual</i>	<i>quís(er)</i> .			
	the	money	which	want. ^{3SG.FUT.SUBJ}			

'If the peasant sells the wine with which he would pay his tax to the king, may the land owner have the right to demand the wine or the money, whichever he wants.'

(*Dos Costumes de Santarém*)

It is quite clear that in (49), we are in the presence of a clause since we still find the volition verb inflected—in this case, in the future subjunctive. Furthermore, despite the presence of a coordinated DP with the last DP in the chain being plural, there is no number agreement between the DP and *qual*, which points to a parenthetical status of the clause. In this particular example, the sequence *qual quís(er)* corresponds to an appositive relative clause (or a parenthetical clause), which I argue to be a relevant context for the emergence of postnominal *qualquer*.

We find in the corpus only five examples where the verb did not correspond to the form *quer* but displayed different tense/mood (subjunctive mood, either future or imperfect tenses) and person/number inflection, as in (23), repeated below as (50).

(50)	[. . .] mais	escolhi	tu	huma	morte	qual	quiseres [. . .]
	but	choose. ^{2SG.IMP}	you	a	death	which	want. ^{2SG.FUT.SUBJ}

'but choose a death, whatever you want.'

(*Diálogos de São Gregório*)

Examples (49) and (50) correspond to additional information on the nominal item on the left and they usually constitute a comment on the existing freedom to choose any element from a list presented before.

Appositive relative clauses could also display discontinuity between the relative determiner *qual* and the volition verb *querer*, as in (51) and (52), but differently from *ever* free relatives, the element in between does not correspond to an additional internal head selected by the determiner. Both in (51) and (52), it is the subject of the clause, which could be lexically empty. As expected, appositive clauses always associate with a nominal element, which they modify.

For example, in (52), the relative appositive clause introduces a comment regarding the set of possible drinking choices presented before, reinforcing the freedom of choice.

Occurrences of *qualquer* as a postnominal modifier of the noun may have originated in appositive relative constructions as the ones presented in (50), with a null subject.

At this point, we hypothesize that instances of *qualquer* in postnominal position may have first originated in appositive relative clauses, instead of *ever* free relatives. Contrary to what we saw for *ever* free relatives, there is no additional internal head, leaving space for the reanalysis of the two elements as a non-clausal adjectival modifier of the noun.

(51)	e	que	eu	mande	lavarar	moeda	<i>qual</i>	eu	<i>quiser</i> .
	and	that	I	order. ^{1SG.PRES.SUBJ}	mint	coin	which	I	want. ^{1SG.FUT.SUBJ}

'and that I order to mint coin, whatever I want.'

(*Crónica Geral de Espanha*)

(52)	E,	cō	cada	huu,	devē	dar	ao
	and	with	each	one	should. ^{3PL.PRES}	give	to.the
	retador	cavallo	e	armas	e	de	comer
	challenger	horse	and	weapons	and	of	eat
	e	de	bever	vinho	ou	auga	qual
	and	of	drink	wine	or	water	which
	elle	<i>quisesse.</i>					
	he	want. ^{3SG.FUT.SUBJ}					

‘And with each other, you should give the challenger horse and weapons and food and drink, wine or water, whichever he wants.’

(Crónica Geral de Espanha)

Although the scarcity of examples does not allow one to present solid arguments for this hypothesis, the comparison with other Romance FCIs may add some strength to the discussion. The Italian FCI *qualsiasi* contains the pronoun *si* in postverbal position. However, according to (Kellert 2021, p. 17), it originated in the relative clause *qual si sia*, with the pronoun *si* between the relative determiner and the verb.³⁰ Degano and Aloni (2021) also argue in the same direction, stating that the forms *qualsiasi* and *qualsivoglia* occurred in medieval Italian as relative clauses, before being reanalyzed as independent items. The presence of the pronoun between *qual* and the verb in the first stage shows that, at least for Italian, the origin of the two FCIs could not have been an *ever* free relative clause with an additional internal head since the reflexive pronoun does not correspond to the nominal internal head.

Examples with a pronoun appearing between *qual* and *quer*, as exemplified in (53), are rare in Portuguese data, showing that this was not a productive construction.

(53)	Mais	en	grave	dia	naci, l	se	Deus
	but	in	unhappy	day	be.born. ^{1SG.PAST}	if	God
	conselho	non	m’	i	der’; l	ca	
	advice	NEG	me. ^{1SG.ACC}	here	give	because	
	d’estas	coitas	<i>qual-xe-quer</i> l		m’	ca	
	of.these	pains	which.SE.Expl.want. ^{3SG.PRES}		me. ^{1SG.ACC}	because	
	é	min	mui	grave	d’endurar	min	
	be. ^{3SG.PRES}	me. ^{1SG.DAT}	very	hard	of.endure	me. ^{1SG.DAT}	

‘But I was born in an unhappy day if God does not give me here advice, because of these pains, no matter which, are very hard for me to endure.’

(Galician-Portuguese poetry, TMILG)

The pronoun *xe* (SE) is usually associated with an expletive use or is interpreted as an ethic dative (Huber 1933, p. 176). It does not correspond to an additional internal head of the relative clause but seems to correspond to an expletive item. What is interesting here is the fact that, contrary to Portuguese, other Romance languages present grammaticalized FCIs that contain the expletive pronoun, as is the case of Old Italian.

Example (53) is the only occurrence found in the sample corpus, though. Due to the scarcity of examples, we may assume that this was not a frequent construction in Old Portuguese and, therefore, it seems logical that the grammaticalized form of the FCI does not preserve the pronoun in its interior. This does not invalidate the emergence of postnominal *qualquer* from appositive relative clauses, similar to what is argued for Italian by Degano and Aloni (2021).

In any case, unambiguous occurrences of *qual quer* as an appositive relative clause are not frequent in the corpus. Most cases of *qualquer* in the postnominal position can already be interpreted as non-clausal, resulting from the merge between the relative determiner *qual* and *quer*, which was already an indefiniteness marker. However, assuming that postnominal *qualquer* originates from reanalysis of appositive relative clauses poses a problem to prenominal occurrences of *qualquer*. We could consider that, after lexicalizing as an independent item, postnominal *qualquer* starts to occur in prenominal position. Under

this hypothesis, we would expect to find a higher frequency of postnominal *qualquer* in the 13th century, but what we find is prenominal occurrences as the most widespread pattern. In the next section, I investigate prenominal and postnominal occurrences of lexicalized *qualquer* and I argue in favour of the existence of two different items: a prenominal *qualquer* that was a quantifier-like item and postnominal *qualquer*, functioning as an adjectival-like modifier.

3.3. Prenominal and Postnominal Instances of *Qualquer*

In this section, I look at prenominal and postnominal occurrences of *qualquer* as a lexicalized item. I argue that prenominal *qualquer* was already a quantifier-like element in 13th century texts, while postnominal *qualquer* behaved as an adjectival element, resulting from the reanalysis of appositive relative clauses, as we have seen before.

Examples (54) and (55) show the most common patterns of occurrence of *qualquer* in prenominal position³¹, where it is the only specifier for the nominal element.

- (54) E ella disse que ante *queria* morrer de
 and she say.^{3SG.PAST} that before want.^{3SG.IMPERF} die of
qual *quer* morte ante que seer christãa.
 which ever death before than be Christian
 ‘And she said that she would rather die of any death rather than being a Christian.’
(Demanda do Santo Graal)

- (55) [...] e rogamos a *qualquer* Tabellion que esta
 and ask.^{1PL.PRES} to whichever notary that this
 carta uj’r que faça ende a carta
 letter see.^{3SG.FUT.SUBJ} that do.^{3SG.PRES.SUBJ} of.this the letter
 da dita partiçõ
 of.the said partition
 ‘and we ask any notary who sees this letter that he makes the letter of the aforementioned partition.’
(DPRNL)

Prenominal *qualquer* also occurs more than half of the times with a *que* clause, but the *que* clause in question is always a restrictive relative clause that modifies the DP in which *qualquer* occurs, as in (56).

- (56) Mas, per qual *quer* maneyra que elle morresse,
 but by which ever manner that he die.^{3SG.IMPERF.SUBJ}
 ouve o poboo delle grande perda.
 have.^{3SG.PAST} the people of.he great loss
 ‘But, by whatever manner he died, the people suffered a great loss.’
(Crónica Geral de Espanha)

The examples above seem to indicate that 13th century prenominal instances of *qualquer* were independent of any possible clausal origin.

Prenominal *qualquer* seems to behave as other quantifier-like elements, as I try to demonstrate in the next paragraphs.

While in prenominal position, *qualquer* usually precedes a noun, as in (57), or a prepositional phrase, as in (58).

- (57) E ella disse que ante *queria* morrer de
 and she say.^{3SG.PAST} that before want.^{3SG.IMPERF} die of
qual *quer* morte ante que seer christãa.
 which ever death before than be Christian
 ‘And she said that she would rather die of any death rather than being a Christian.’
(Demanda do Santo Graal)

a [+human] feature. This possibility was lost until the end of the 16th century. Although I cannot indicate the exact period in which *qualquer* stops occurring as a bare quantifier, the loss of the [+human] feature can likely be paired with the same event affecting *algum* and *nenhum*.³³ Galician *calquera* and Spanish *cualquiera* have kept that possibility though (cf. Álvarez and Xove 2002, p. 5005; Company Company and Pozas Loyo 2009).

- (63) [. . .] e começou de ferir dhua e da
and start.^{3SG.PAST} of hurt of.one and of.the
outra parte, de tal guisa que *qualquer*
other part in such manner that anyone
a que elle dava hua paancada nō
to who he give.^{3SG.IMPERF} one hit NEG
avia mester mais ferida.
have.^{3SG.IMPERF} need more wound
‘and he started to attack in all directions in such a way that anyone whom he hit would be knocked out’
(Crónica Geral de Espanha)

Occurrences of *qualquer* as a bare quantifier are, however, always followed by a *que* clause, which is unambiguously a restrictive relative clause. The presence of the preposition *a* ‘to’ between *qualquer* and the element *que* in (63) indicates that it is a relative clause due to preposition pied-piping (the preposition introduces the dative selected by the verbal form *dava* ‘give.^{3SG.IMPERF}’). In these cases, the relative clause seems to set the domain of restriction of the quantifier since *qualquer* appears in a bare configuration.

The examples presented so far seem to point to the existence of a quantifier-like element *qualquer*, which could be paired with other quantifiers/indefinites available in Old Portuguese and which seems to be independent from the clausal instances we have seen before.

Although quantifier *qualquer* occurs mostly as the only specifier of the nominal element, there are a few examples that require some clarification. I refer to the cases where prenominal *qualquer* coocurs with the indefinite *outro* ‘other’ before the noun (cf. Brugè 2018; Brugè and Giusti 2021). There is a total of nine occurrences of prenominal *qualquer* with *outro* ‘other’ in the corpus (it represents 20% of all prenominal entries). In two examples, *qualquer* precedes *outro* ‘other’, while in the remaining entries, *qualquer* appears after *outro* ‘other’, as illustrated in (64) and (65), respectively.

- (64) [. . .] e carta ou cartas ende fazer pelos tabellioes de Lixbõa
and letter or letters of.this do by.the notaries of Lisbon
ou de *qualquer* *outro* logar
or of whichever other place
‘and make letter or letters of this through the notaries of Lisbon or any other place’
(DPRNP)

- (65) [. . .] ou desse alguu aver por alguu beneficio
or give.^{3SG.IMPERF.SUBJ} some good by some benefit
da Sancta Igreja ao rey ou ao
of.the Saint Church to.the king or to.the
prellado ou a *outro* *qualquer* padroeiro, asi
prelate or to other whichever patron like.this
eclesiastico como sagral
ecclesiastic as sacred
‘or give any good by any benefit from the Holy Church to the king or the prelate or any other patron as ecclesiastic as sacred.’
(Crónica Geral de Espanha)

The combination with *outro* is usually found as a way to introduce the last nominal item of a coordination displaying two or more nouns in alternative. Although the sequence

in (64) does not pose problems to a quantifier-like interpretation, sentences such as (65) seem incompatible with *qualquer* being the head of QP because we have *outro*, an adjectival element, preceding the quantifier. This may be a reflex of the different nature of *qualquer* in sentences (64) and (65)³⁴. I return to this point in Section 4.

Let us now look at occurrences of lexicalized *qualquer* in the postnominal position.

While in the postnominal position, *qualquer* is almost always part of an indefinite DP. The most frequent pattern (it represents 66% of the occurrences) is the one containing the indefinite *outro* ‘other’ in the prenominal position, as illustrated in (66). Other elements such as *um* ‘a’ or *algum* ‘some’ can also be found, as in examples (67) and (68), respectively, but they are infrequent. There are only three occurrences of *um* ‘a’ and one of *algum* ‘some’.

- | | | | | | | | |
|------|------|---------------------------|--------------------------------|-----------------------------|--------------------------------|------|---------|
| (66) | E | se | iustiça | fezer, | aya | a | pêa |
| | and | if | justice | do. ^{3SG.FUT.SUBJ} | have. ^{3SG.PRES.SUBJ} | the | penalty |
| | que | auerya | outro | ome | qual | quer | que |
| | that | have. ^{3SG.COND} | other | man | whichever | | that |
| | tal | feyto | fezesse. | | | | |
| | such | deed | do. ^{3SG.IMPERF.SUBJ} | | | | |
- ‘And if justice is made, may he have the same penalty any other man would for such a deed.’
- (Foro Real)

- | | | | | | | |
|------|----------|-------------------------|--------------------------|----------------------------------|---------------------------|--------|
| (67) | [...] nō | lhe | daram | mayor | corrigimento | q(ue) |
| | NEG | him. ^{3SG.DAT} | give. ^{3PL.FUT} | bigger | correction | than |
| | dua | pínquada | que | lhe | dem | nos |
| | of.one | stroke | that | him. ^{3SG.DAT} | give. ^{3PL.PRES} | in.the |
| | narizes | de | que | saya | sangue | ou |
| | noses | of | that | go.out. ^{3SG.PRES.SUBJ} | blood | or |
| | dua | chaga | símplx | qual | q(ue)r | |
| | of.one | wound | simple | whichever | | |
- ‘they will not give him any correction other than a stroke in the nose, from where blood runs or from any simple wound.’
- (Dos Costumes de Santarém)

- | | | | | | | | | |
|------|----------|----------|--------|--------------------------|--------------------------|-------|---------|-------|
| (68) | E | vós | mentes | non | metedes, l | se | ela | filho |
| | and | you | minds | NEG | put. ^{2PL.PRES} | if | she | son |
| | fezer, l | andando, | como | veedes, l | con | algun | peon | qual |
| | make | walking | as | see. ^{2PL.PRES} | with | some | peasant | which |
- quer
ever
- ‘And can you not notice that she may get pregnant, going out, as you see, with some ordinary peasant.’
- (Galician-Portuguese poetry)

Postnominal *qualquer* also occurs with modification using a restrictive relative clause with the subjunctive mood in 23% of the contexts. The relative clause frequently displays the copula *ser* ‘be’, as in (69).

- | | | | | | | | |
|------|------|-----------|-----------|--------------------------------|-------------------------------|----------|------------------------------|
| (69) | Os | scriuaans | publicos | tenhã | as | notas | primeyras |
| | the | scribes | public | have. ^{3PL.PRES.SUBJ} | the | notes | first |
| | de | todalhas | cartas | que | fezerẽ, | assy | |
| | of | all.the | letters | that | make. ^{3PL.FUT.SUBJ} | this.way | |
| | as | dos | juyzos | coma | das | uendas | |
| | the | of.the | judgments | as | of.the | sales | |
| | come | doutro | preyto | qual | quer | que | seya |
| | as | of.other | contract | whichever | | that | be. ^{3SG.PRES.SUBJ} |
- ‘The public scribes must have the first notes of all letters they do, the ones from the judgments, as well as the ones from the sales or from other contract, whatever may be.’
- (Foro Real)

Apart from the examples above, postnominal *qualquer* is also registered in what we can call a bare noun configuration, without the presence of any prenominal element, as illustrated in (70). This pattern corresponds to approximately 17% of the occurrences of postnominal *qualquer*.

(70)	Se	alguu	ome	fezer	demanda	a	outro	
	if	some	man	do. ^{3SG.FUT.SUBJ}	demand	to	other	
	sobre	casa	ou	uinha	ou	herdade	<i>qualquer</i> ,	
	over	house	or	vineyard	or	property	whichever	
	((demande))	an(te)	aquel	alcayde	u	é		morador
	demand. ^{3SG.PRES.SUBJ}	before	that	mayor	where	be. ^{3SG.PRES}		resident
	aquel	a	quē	demandē				
	that	to	whom	demand. ^{3PL.PRES.SUBJ}				

‘If any man makes a demand to other over a house or vineyard or any property, ask before the mayor where lives the man to whom demands are made.’

(*Foro Real*)

Sentences such as the one above are ruled out in contemporary Portuguese due to the fact that CEP does not allow singular bare nouns, except in very specific contexts. Singular bare nouns (and bare nouns in general) have been pointed out as occurring more freely in earlier stages of the language, which can explain examples such as (70). The combination with bare nouns is also registered for Old Italian *qualunque*, with this pattern being more frequent than the one including an indefinite determiner (cf. Kellert 2021). In any case, bare nouns are usually associated with an indefinite or generic interpretation.

4. On the Emergence of Two *Qualquer*: General Discussion

Up to this point, I have argued that postnominal instances of *qualquer* most likely originated in appositive relative clauses. I have therefore rejected the assumption that they resulted from *ever* free relative clauses with the nominal additional internal head *in situ*.

On the other hand, I have claimed that postnominal *qualquer* seems to behave as an adjective-like modifier of the noun, while prenominal *qualquer* is close to a quantifier-like item. However, it is yet to be explained how prenominal *qualquer* emerged.

I follow the insights by Rivero (1984, 1988) for Old Spanish. The author suggests that the relative *qual* had “a double lexical classification” in medieval Spanish, still occurring in *ever* free relative clauses with the volition verb, but also occurring as part of the word *qualquer*, a member of the quantifiers’ paradigm. To *ever* free relatives, I also add the intervention of relative *qual* and the volition verb in appositive relative clauses.

Nevertheless, the ambiguity of relative *qual* is not enough to explain the emergence of prenominal *qualquer*. Data from medieval Italian *qualsisia* show us a context of use that we did not find in Portuguese data, namely the occurrence of a relative clause in prenominal position, as illustrated in (71) and (72), to which Degano and Aloni (2021) associate different interpretations: a *no matter* (71) and an *FC indefinite* interpretation (72).

(71)	Qual	si	sia	la	cagione,	oggi	poche	o	non
	what	CLITIC	is.SUBJ	the	reason	today	few	or	not
	niuna	donna	rimasa	ci	è	la	qual...		
	no-one	women	left	to-us	is	the	who		

‘Whatever the cause is, today few or no women felt is such that ...’

(Boccaccio, *Decameron* VI: 1–10, 1353; apud Degano and Aloni 2021, p. 464)

(72)	i	quali	sì	timorosamente	mostrano	di	dare	le	openioni	sopra
	the	WHO	so	timidly	show	to	say	the	opinions	on
	qual	si	sia	proposta.						
	What	clitic	is:SUBJ	proposal						

‘who so timidly show that they say their opinions on any proposal’

(Della Casa, *Galateo ovvero de’ costume*, 1558; apud Degano and Aloni 2021, p. 464)

I hypothesize that Old Portuguese may have had similar structures, involving the volition verb *querer*, and those may have been reanalyzed as independent prenominal *qualquer*. In *Foros de Castelo Rodrigo*, a collection of local laws written in the first half of the 13th century, Cintra (1984) accounts for the occurrence of a construction with the form *qual que quer eglesia* ‘what that want church’, with the relative clause preceding the nominal element. In any case, if this type of relative is at the core of prenominal *qualquer*, reanalysis must have occurred very early in the language, prior to the reanalysis of postnominal *qualquer*.

Finally, one last question that needs to be addressed is related to differences in meaning associated with the prenominal or postnominal position of *qualquer*. Even though prenominal and postnominal *qualquer* seem to originate in a relative clause, the clauses were different, and the chronology of the reanalysis was also different. This may have had some implications for the meanings associated with *qualquer*, depending on its position in relation to the noun. In CEP, when occurring in the postnominal position, *qualquer* never displays a universal reading, while in the prenominal position, it is usually interpreted as a universal quantifier, but it can also be associated with an existential interpretation.

I start by recalling that prenominal *qualquer* was paired with quantifiers, while postnominal *qualquer* was associated with an adjectival-like nature. However, examples such as (73) were considered problematic for a quantifier-like status due to the presence of the indefinite *outro* to the left of the quantifier.

(73)	E	guareceu	daquella	gordura	tam	bem	que
	and	protect. ^{3SG.PAST}	of.that	grease	so	well	that
	tornou	a	seer	assi	delgado	como	outro
	become:3SG.PAST	to	be	this.way	thin	as	other
	qualquer	homen	que	delgado	fosse		
	whichever	man	who	thin	be. ^{3SG.IMP-SUBJ}		

‘And he avoided that grease so well that he became thin again, as any other thin man.’

(Crónica Geral de Espanha)

I hypothesize that in contexts such as (73), *qualquer* does not correspond to a quantifier, but is adjectival, just like when it occurs in the postnominal position. I agree with the insights by Elvira (2020), who claims that prenominal uses of *cualquiera* result from its reinterpretation from an adjective into a determiner, similar to what had occurred with other adjectives, such as *cierto* ‘certain’. However, contrary to Elvira (2020), I consider that *qualquer* is not exactly reinterpreted as a determiner. Similar to other adjectives, it starts occupying two different positions within the syntactic structure: one which is the basic position for adjectives (under the scope of noun phrase (NP), for Portuguese, as argued by Gonzaga 2004) and another which is dedicated to adjectives with quantificational interpretation. This last position is said to host adjectives such as *certo* ‘certain’, which assume different meanings, when appearing before or after the noun (cf. Gonzaga 2004), as illustrated in (74) with *várias* ‘several/diverse’.³⁵

(74)	a.	Várias	peessoas	estiveram	presentes	na	festa.
		several	persons	be. ^{3PL.PAST}	present	in.the	party
		‘Several people were at the party’					
	b.	Pessoas	várias	estiveram	presentes	na	festa
		persons	diverse	be. ^{3PL.PAST}	present	in.the	party
		‘Different people were at the party’					

Brito and Lopes (2016) suggest a position higher than NP, as specifiers of a number phrase (NumP) projection,³⁶ as the hosting site for these adjectives.³⁷ Following both Gonzaga (2004) and Brito and Lopes (2016), we can assume three different positions for adjectives, as illustrated in (74):

(75) [DP [D°] [NumP [AP₃] [Num' [Num°] [NP [AP₂][N' [N°] [AP₁]]]]]]

The positions AP1 and AP2 correspond to the basic position. Following Gonzaga (2004, p. 24), «adjectives that directly change the noun are base generated inside the NP». They can be specifiers or complements of NP.³⁸ As for the position AP3, it corresponds to a position dedicated to the aforementioned adjectives with quantificational interpretation. The architecture in (75) can host the orders *um qualquer* N, as well as *um N qualquer*, while the structure previously presented in (59) accounts for the order *qualquer* N, where *qualquer* is a quantifier.

This hypothesis would, in principle, allow one to explain the fact that postnominal *qualquer* is never assigned a universal interpretation, while *qualquer* may display a universal reading, but also an existential reading, if in the prenominal position. The universal reading would be conveyed by instances of *qualquer* as a quantifier, while existential readings would result from adjectival *qualquer* in the prenominal position, sitting in a position dedicated to adjectives with quantificational reading. An anonymous reviewer draws attention to the fact that other quantifiers also present variation relative to prenominal and postnominal positions. However, only *qualquer* requires the presence of a determiner in the postnominal position (cf. contrast between 76d and 76e), but rules it out in the prenominal position (cf. 76), pairing with other quantifiers. This suggests that there are two different instances of *qualquer* at stake.

(76)	(a)	Qualquer Algum Nenhum <i>Any/none</i>	rapaz <i>boy</i>	
	(b)	*Qualquer *Algum *Nenhum <i>Any/none</i>	um <i>a (DET)</i>	rapaz <i>boy</i>
	(c)	Um <i>a (DET)</i>	qualquer *nenhum *algum <i>any/none</i>	rapaz <i>boy</i>
	(d)	Rapaz <i>boy</i>	*qualquer algum ³⁹ nenhum <i>any/none</i>	
	(e)	Um <i>a (DET)</i>	rapaz <i>boy</i>	qualquer *algum *nenhum <i>any/none</i>

Unfortunately, such a theory needs empirical validation that I am not in a position to offer here. I postpone such an analysis to future research.

5. Conclusions

This paper has presented the development of the FCI *qualquer* in Portuguese, constituting, to the best of my knowledge, the first in-depth study on the diachrony of *qualquer*.

Based on texts of the medieval period, I have argued in favour of two different origins for *qualquer*. Following the proposal by Rivero (1986, 1988) for Old Spanish, I have defended the existence of a quantifier *qualquer*, occurring only in prenominal position and associated with universal interpretations, already in the 13th century. On the other hand, *qualquer*

also develops as an adjectival-like element. I have suggested that this use originated in appositive relative clauses through merge of the relative determiner *qual* ‘which’ and the indefiniteness marker *quer*, a former form of the volition verb *querer* ‘want’.

In this paper, I have also put forth the idea that different diachronic origins for the two instances of *qualquer* may be at the core of different interpretations in contemporary uses. I have hypothesized that universal readings are conveyed by the quantifier, which can only occur in the pronominal position. On the other hand, instances of adjectival *qualquer* usually occur in the postnominal position. Nevertheless, an existential interpretation can be assigned to adjectival *qualquer* when it occupies a position dedicated to adjectives with a quantifier or determiner-like interpretation, as proposed by Brito and Lopes (2016). This idea, however, lacks empirical evidence, and I will, therefore, postpone it to future work.

This paper offered an important contribution to the study of FCIs by presenting a detailed description of Portuguese *qualquer* and enabling the comparison with other similar items in other Romance languages. It also tried to establish a relation between the diachronic development of *qualquer* and its synchronic behaviour, contributing to the general understanding of polarity items in contemporary data.

Funding: This investigation has been funded by the FCT—*Fundação para a Ciência e a Tecnologia, I.P.*, through the projects UIDB/03213/2020 and UIDP/03213/2020—Centro de Linguística da Universidade NOVA de Lisboa (CLUNL).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The author declares no conflict of interest.

Notes

¹ The English FCI *any* is problematic, though, because it can also be a negative polarity item (NPI). The literature on the nature of *any* is vast and some authors advocate in favour of two items *any* (cf. Dayal 2004), while others have tried to pursue a unifying approach (cf. Kadmon and Landman 1993; Horn 2005). I will not enter this discussion here.

² Peres (2013) considers as cases of ‘equivalence’ the contexts where *qualquer* displays a free choice reading. This means that all the elements in the set are seen as equivalent. The ‘unknown’ value refers to interpretations where it is not possible to determine the identity of the element being referred to by *qualquer*. Finally, the ‘restriction’ value applies when *qualquer* is used to isolate a subset within the set of elements introduced by the noun.

³ Examples are from Peres (2013), the translation is mine.

⁴ This interpretation seems to be recent in the language and resembles what is described by Kellert (2021) as an evaluative meaning for Italian postnominal *qualunque*.

⁵ An anonymous reviewer has drawn attention to the fact that the word order ‘outro+qualquer+N’ does not sound natural. There is probably variation among speakers in the acceptance of such order. A brief search in *Corpus do Português* (cf. Davies 2006) returns several entries with such word order pattern, as example (i).

(i) Amanhã poder em Espanha, Portugal ou em outro qualquer país europeu
tomorrow can be in Spain Portugal or in other any country European
‘Tomorrow it can be in Spain, Portugal or any other European country’
Corpus do Português, <https://www.corpusdoportugues.org/web-dial/> (consulted on 2 October 2023)

⁶ According to Company Company and Pozas Loyo (2009), the Spanish equivalent is still possible as a pronominal.

⁷ The close relation established by the adjective *single* and the numeral has been noticed by Kayne (2020) for English. Kayne (2020, p. 343) suggests that the numeral *one* is accompanied by a «silent adjective corresponding to single». For Portuguese data, Pinto (2021) suggests a similar relation between the adjective *único* ‘single’ and the cardinal *um* ‘one’.

⁸ Given the fact that *qualquer* is a fully grammaticalized item, we would expect the plural morpheme *-s* to be added at the end of the word, but the form *quaisquer* is still preserved, mostly due to scholarisation. However, in informal oral speech, we frequently find the form *quaisques*, with two *-s* morphemes. The internal *-s* after *qual* is most likely not being interpreted as a plural morpheme anymore, and the form *quaisquer* is seen as a fully grammaticalized word. Interestingly, Modern Galician does not allow number variation of *calquer(a)*, despite deriving from the same medieval form *qual quer* (cf. Ferreiro 1999).

⁹ Cintra considers as Old Portuguese the period between 1214 and 1420.

¹⁰ Available at <https://www.claris.com/filemaker/> (accessed on 5 June 2023).

- 11 Full reference in the *References* section of this paper.
- 12 *Idem.*
- 13 This includes both singular *qualquer* and plural *quaisquer*.
- 14 This could simply be a consequence of text typology, but it is probably a reflection of the progressive loss of the relative pattern. While in the 13th century, there were three different possible patterns of occurrence of *qualquer*, by the end of the 14th century, only two were possible.
- 15 A brief search in two texts from the 15th century—*Crónica de D. Pedro de Menezes* (cf. Brocardo 1997) and *Orto do Esposo* (cf. corpus CIPM)—did not return occurrences of *qual. . .quer* as an *ever* free relative or appositive relative.
- 16 Dayal (1997, p. 99) draws attention to the fact that these values can be reversed in specific contexts.
- 17 Portuguese also displays an equivalent strategy involving a relative clause and the copula *ser* ‘to be’, as illustrated in sentence (c). Both the *ever* free relative in (b) and the relative clause with *ser* are equivalent to (a).
- (a)—*Whoever that is, do not open the door.*
 (b)—*Quem quer que seja, não abras a porta.*
 (c)—*Seja quem for, não abras a porta.*
- 18 Huber (1933) refers to them as generalizer relative pronouns.
- 19 It should be mentioned that *ever* free relatives with bare *que* ‘what’ are no longer possible. They were replaced by the relative compound *o-que* ‘the.what’.
- 20 Only the compound *o-qual* is possible in CEP and it is said to only marginally allow internal heads (cf. Cardoso 2011). In Bechara ([1961] 2001) we still find examples of relative clauses introduced by *o-qual* and exhibiting an internal additional head. On the other hand, Brito (1991) presents these constructions as marginal.
- 21 The form *quer*, descendant from the volition verb *querer* ‘want’, is also present in the formation of the adverbial item *sequer* (from Latin *SI QUAERIT*) and the correlative disjunctive conjunction *quer. . .quer*.
- 22 On *wh-ever* words in English, see Larrivé and Duffley (2020).
- 23 This is not such a straightforward assumption, though. The existence of a lexically null antecedent has been proposed by several authors (Brito 1991; Mória 1992b; van Riemsdijk 2006). For instance, Mória (1992b) uses the term *relatives without an expressed antecedent* to refer to free relatives.
- 24 In the cases where there is variable verb inflection, the person/number does not always match the person/number of the verb in the main clause.
- 25 For sake of simplicity and to avoid attributing a chronology to the grammaticalization of *quer* as an independent marker, I gloss it as a verbal form regardless of its status in examples from medieval Portuguese.
- 26 A *baillia* was the jurisdiction given to clergy members or knights from military orders.
- 27 As stated by Kellert (2021, p. 17) the proposed grammaticalization path “has not been empirically testified for Spanish *cualquiera* due to the fact that the first Old Spanish data already contain the lexicalised item *cualquiera*”. This puts Spanish *cualquiera* in the same situation as Portuguese *qualquer*.
- 28 In case we assumed the NP to be *in situ* and the relative *qual* to be in Spec, CP, this would require the indefinite *outro* ‘other’ to have moved from the DP (and under what motivation?) and occupy a position higher than Spec, CP to respect word order. Also, there are no registers in the corpus of *ever* free relatives preceded by *outro* (as in *outro qual N quer*).
- 29 CEP still allows relative clauses headed by *wh*-phrases. It is the case of *quanto(s)* ‘how.many/much’, which can still occur with an additional internal head, as illustrated in (i):
- (i) O Paulo bebeu quanto leite quis (beber). (Cardoso 2008, p. 88)
 ‘Paul drunk as much milk as he wanted.’
- 30 Palomo (1934) and Rivero (1986) also account for the existence of the form *qualsequier* in Old Spanish data.
- 31 It is worth mentioning that the writing of prenominal *qualquer* separately (63) or as one single graphic unit (64) does not determine any differences in terms of syntactic/semantic properties.
- 32 Whether *qualquer* takes a covert DP or no complement at all is a question I cannot address here.
- 33 On the [+ human] feature of the indefinites *algum* ‘someone’ and *nenhum* ‘no.one’ in Old Portuguese, see Martins (2016) and also Pinto (2021).
- 34 Sequences of *qualquer* + *OUTRO* (as *Q > altri*) may, eventually, correspond to complex quantifiers, following the recent proposal by (Brugè and Giusti 2021).
- 35 For instance, the adjective *certo* ‘right/certain’ triggers different readings depending on its position within the DP structure. In the postnominal position, it is a qualitative adjective (i), while in the prenominal position it activates an existential meaning (ii).
- (i) *Devas valorizar mais as pessoas certas.*
 ‘You should give more value to the right people.’
- (ii) *Certas pessoas não merecem a nossa atenção.*
 ‘Certain people do not deserve our attention.’

- 36 Number Phrase (NumP) is a projection between D and N, which was first proposed by Ritter (1991) to encode number features. After Déprez (2005), it has also been argued to be the host position for cardinals and weak quantifiers.
- 37 Bernstein (1993) advocates for a position higher than NumP for such adjectives.
- 38 It is assumed that attributive adjectives are universally inserted in a prenominal position, and in languages with surface order ‘noun-adjective’, it is derived by cyclic leftward movement of the noun to a higher functional head in the nominal spine (cf. Cinque 1994, a.o.).
- 39 It should be mentioned that postnominal uses of *algum* ‘some’ produce negative readings due to nominal negative inversion (cf. Martins 2015).

References

Primary Source

- Davies, Mark. 2006. *Corpus do Português*. Available online: <https://www.corpusdoportugues.org/> (accessed on 15 May 2023).
- Varela Barreiro, Xavier. 2004. *Tesouro Medieval Informatizado da Língua Galega*. Santiago de Compostela: Instituto da Língua Galega. Available online: <https://ilg.usc.es/tmilg/> (accessed on 20 May 2023).
- Xavier, Maria Francisca. 1993–2003. *Digital Corpus of Medieval Portuguese (CIPM—Corpus Informatizado do Português Medieval)*. Lisboa: Faculdade de Ciências Sociais e Humanas da Universidade Nova de Lisboa. Available online: <http://cipm.fcsh.unl.pt/> (accessed on 20 May 2023).

Secondary Source

- Abney, Steven. 1987. *The English Noun Phrase in Its Sentential Aspect*. Ph.D. thesis, MIT, Cambridge, MA, USA.
- Álvarez, Rosario, and Xosé Xove. 2002. *Gramática da Língua Galega*. Vigo: Editorial Galaxia.
- Bechara, Evanildo. 2001. *Moderna Gramática Portuguesa*, 37th ed. Rio de Janeiro: Companhia Editora Nacional/Editora Lucerna. First published 1961.
- Bernstein, Judy. 1993. *Topics in the Syntax of Nominal Structure across Romance and Germanic Languages*. Ph.D. thesis, City University of New York, New York, NY, USA.
- Bossuyt, Tom, and Torsten Leuschner. 2020. WH-ever in German, Dutch and English: A contrastive study showcasing the ConverGEN-Tiecorpus. In *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Edited by Renata Enghels, Bart Defrancq and Marlies Jansegers. Berlin and Boston: De Gruyter Mouton, pp. 183–220.
- Brea, Mercedes. 1996. *Lírica Profana Galego-Portuguesa. Corpus Completo das Cantigas Medievais, con Estudio Biográfico, Análise Retórica e Bibliográfica Específica*. Santiago de Compostela: Centro de Investigacións Lingüísticas e Literarias Ramón Piñeiro.
- Brito, Ana Maria. 1991. *A Sintaxe das Orações Relativas em Português*. Lisboa: Instituto Nacional de Investigação Científica.
- Brito, Ana Maria, and Ruth Lopes. 2016. The structure of DPs. In *The Handbook of Portuguese Linguistics*. Edited by Willem Leo Wetzels, João Costa and Sergio Menuzzi. Malden: Wiley-Blackwell, pp. 254–74.
- Brocardo, Teresa. 1997. *Crónica do Conde D. Pedro de Meneses—Gomes Eanes de Zurara. Edição e Estudo*. Lisboa: FCG/JNICT.
- Brugè, Laura, and Giuliana Giusti. 2021. Some notes on outro in Portuguese. *Revista de Estudos Linguísticos da Universidade do Porto (Número especial de Homenagem à Professora Ana Maria Barros de Brito)* 441–71.
- Brugè, Laura. 2018. Otro: Consideraciones a favor de su no inclusión en la categoría Determinante. *Borealis* 7: 109–34. [CrossRef]
- Caponigro, Ivano. 2019. In Defense of What(ever) Free Relative Clauses They Dismiss: A Reply to Donati and Cecchetto (2011). *Linguistic Inquiry* 50: 356–71. [CrossRef]
- Cardinaletti, Anna, and Giuliana Giusti. 1992. Partitive *ne* and the QP-hypothesis. A Case study. Paper presented at the XVII Meeting of Generative Grammar, Trieste, Italy, February 22–24. Torino: Rosenberg & Sellier, pp. 121–41.
- Cardinaletti, Anna, and Giuliana Giusti. 2006. The Syntax of Quantified Phrases and Quantitative Clitics. In *The Blackwell Companion to Syntax*. Edited by Martin Everaert and Henk van Riemsdijk. Oxford: Blackwell, vol. 5, pp. 23–93.
- Cardoso, Adriana. 2008. Relativas com núcleo interno e relativo de ligação na história do português. In *Textos Seleccionados do XXIII Encontro Nacional da Associação Portuguesa de Linguística. Paper presented at the XXIII Meeting of APL, Évora, Portugal, October 1–3, 2007*. Lisboa: APL, pp. 77–92.
- Cardoso, Adriana. 2011. *Variation and Change in the Syntax of Relative Clauses: New Evidence from Portuguese*. Ph.D. thesis, Universidade de Lisboa, Lisboa, Portugal.
- Castro, Ivo. 1993. «Demanda do Santo Graal»; «Livro de José de Arimateia»; «Matéria da Bretanha»; «Merlim». In *Dicionário de Literatura Medieval Galega e Portuguesa*. Edited by Julia Lanciani and Giuseppe Tavani. Lisboa: Caminho.
- Castro, Ivo. 1999. O português médio segundo Cintra. In *Lindley Cintra: Homenagem ao Homem, ao Mestre e ao Cidadão*. Edited by Isabel Hub Faria. Lisboa: Cosmos & Faculdade de Letras da Universidade de Lisboa, pp. 367–70.
- Chomsky, Noam, and Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry* 8: 425–504.
- Cinque, Guglielmo. 1994. On the evidence of partial N-movement in the Romance DP. In *Paths towards a Universal Grammar*. Edited by Guglielmo Cinque, Jan Koster, Jean-Yves Pollock, Luigi Rizzi and Rafaella Zanuttini. Washington, DC: Georgetown University Press, pp. 85–110.
- Cintra, Luís Filipe Lindley. 1984. *A Linguagem dos Foros de Castelo Rodrigo*. Lisboa: INCM.

- Cintra, Luís Filipe Lindley. 1951–1990. *Crónica Geral de Espanha de 1344, Edição crítica do texto português*. 4 vols. Ed. Fac-similada. Lisboa: INCM.
- Colomina I Castanyer, Jordi. 2002. Paradigmes flectius de les altres classes nominals. In *Gramàtica del Català Contemporani, 2: Sintaxi*. Edited by Joan Solà, Maria-Rosa Lloret, Joan Mascaró and Manuel Perez Saldanya. Barcelona: Empúries, pp. 539–84.
- Company Company, Concepción. 2009. Parámetros de gramaticalización en los indefinidos compuestos del español. In *Romanística sin Complejos. Homenaje a Carmen Pensado*. Edited by Fernando Sánchez Miret. Berna: Peter Lang, pp. 71–104.
- Company Company, Concepción, and Julia Pozas Loyo. 2009. Los indefinidos compuestos y los pronombres genérico-impersonales omne y uno. In *Sintaxis histórica de la lengua española*. Edited by Concepción Company Company. México: FCE/UNAM, vol. 2, pp. 1073–1219.
- Cuervo, Rufino. 1893. *Diccionario de construcción y régimen de la lengua castellana, vol. I*. Santafé de Bogotá: Instituto Caro y Cuervo.
- Dayal, Veneeta. 1997. Free Relatives and Ever: Identity and Free Choice Readings. Paper presented at the VII SALT Conference, Stanford University, Stanford, CA, USA, March 21–23. Edited by Aaron Lawson. Ithaca: Cornell University, pp. 99–116.
- Dayal, Veneeta. 2004. The universal force of free choice any. *Linguistic Variation Yearbook* 4: 5–40. [CrossRef]
- Degano, Marco, and Maria Aloni. 2021. Indefinites and free choice: When the past matters. *Natural Language & Linguistic Theory* 40: 447–84.
- Déprez, Viviane. 2005. Morphological number, semantic number and bare nouns. *Lingua* 115: 857–83. [CrossRef]
- Elvira, Javier. 2020. Lexicalización primero, gramaticalización después. Aproximación a la génesis de cualquiera. In *El español y las lenguas peninsulares en su diacronía miradas sobre una historia compartida. Estudios dedicados a M.^a Teresa Echenique Elizondo*. Edited by María Alcade, Juan Sánchez Méndez, Francisco Grau, Mercedes Merín, Amparo Vidal, Adela Valle, Francisco Colomer and Santiago Llavata. Valencia and Neuchâtel: Tirant /Universidade de Neuchâtel, pp. 153–68.
- Ernout, Alfred, and François Thomas. 1972. *Syntaxe Latine*. Paris: Klincksieck.
- Ferreira, José de Azevedo. 1987. *Afonso X, Foro Real*. Lisboa: INIC, pp. 125–309.
- Ferreiro, Manuel. 1999. *Gramática Histórica Galega I: Fonética e Morfosintaxe*. Santiago de Compostela: Edicions Laiovento.
- Giannakidou, Anastasia. 2001. The Meaning of Free Choice. *Linguistics and Philosophy* 24: 659–735. [CrossRef]
- Giannakidou, Anastasia, and Lisa Lai-Shen Cheng. 2006. (In)definiteness, Polarity, and the Role of wh-morphology in Free Choice. *Journal of Semantics* 23: 135–83. [CrossRef]
- Gonzaga, Manuela. 2004. The Structure of DP in European Portuguese—Evidence from Adjectives and Possessives. *Harvard Working Papers in Linguistics* 10: 19–49.
- Haspelmath, Martin. 1995. Diachronic sources of “all” and “every”. In *Quantification in Natural Languages*. Edited by Emmon Bach, Eloise Jelinek, Angelica Kratzer and Barbara Partee. Dordrecht: Kluwer Academic Publishers, pp. 363–82.
- Horn, Larry. 2005. Airport ‘86 revisited: Toward a unified indefinite any. In *Reference and Quantification: The Partee Effect*. Edited by Gregory Carlson and Jeffry Pelletier. Stanford: CSLI Publications, pp. 179–205.
- Huber, Joseph. 1933. *Altportugiesisches Elementarbuch*. Heidelberg: Carl Winters, (Portuguese translation, 1986, Gramática do Português Antigo. Lisboa: Fundação Calouste Gulbenkian).
- Jacobson, Pauline. 1995. The quantificational force of English free relatives. In *Quantification in Natural Language*. Edited by Emmon Bach, Eloise Jelinek, Angelica Kratzer and Barbara Partee. Dordrecht: Kluwer Academic Publishers, pp. 451–86.
- Kadmon, Nirit, and Fred Landman. 1993. ‘Any’. *Linguistics and Philosophy* 16: 353–422. [CrossRef]
- Kayne, Richard. 2020. Some thoughts on one and two and other numerals. In *Linguistic Variation: Structure and Interpretation*. Edited by Ludovico Franco and Paolo Russo. Boston and Berlin: De Gruyter, pp. 335–56.
- Kellert, Olga. 2021. Free choice indefinites in old and modern Italian. *Journal of Historical Syntax* 5: 1–27.
- Larrivé, Pierre, and Patrick Duffley. 2020. Whatever floats your boat: A corpus-based investigation of definiteness, quantification, modality, presuppositional content, scalarity and epistemic stance with wh-ever words. *International Review of Pragmatics* 12: 206–45.
- Lombard, Alf. 1938. Une classe spéciale de termes indéfinis dans les langues romanes. *Studia Neophilologica* 11: 186–209. [CrossRef]
- Machado Filho, Américo. 2013. *Diálogos de São Gregório. Edição Semidiplomática*. Bahia: UFBA. Available online: <https://repositorio.ufba.br/handle/ri/17922> (accessed on 30 May 2023).
- Mackenzie, Ian. 2019. *Language Structure, Variation and Change: The Case of Old Spanish Syntax*. London: Palmgrave MacMillan.
- Martins, Ana Maria. 2001. *Documentos Portugueses do Noroeste e da Região de Lisboa*. Lisboa: Imprensa Nacional Casa da Moeda.
- Martins, Ana Maria. 2004. A Emergência do Português Escrito na Segunda Metade do século XII. In *(Dis)curso da Escrita: Estudos de Filologia Galega Ofrecidos em Memória de Fernando R. Tato Plaza*. Edited by Rosario Álvarez and Antón Santamarina. A Coruña: Fundación Pedro Barrié de la Maza, pp. 491–526.
- Martins, Ana Maria. 2013. Copiar o português duocentista: A Demanda e o Jose de Arimateia. In *Ao sabor do texto, Estudos dedicados a Ivo Castro*. Edited by Rosario Álvarez, Ana Maria Martins, Henrique Monteagudo and Ana Maria Ramos. Santiago de Compostela: Universidade de Santiago de Compostela, pp. 383–402.
- Martins, Ana Maria. 2015. Ordem de palavras e polaridade: Inversão nominal negativa com algum/alguno e nenhum. *Diacrítica* 29: 401–28.
- Martins, Ana Maria. 2016. Introdução: O português numa perspectiva diacrónica e comparativa. In *Manual de Linguística Portuguesa*. Edited by Ana Maria Martins and Ernestina Carrilho. Berlin/Boston: De Gruyter, pp. 1–39.

- Medeiros, Bruno. 2022. Tópicos sobre a semântica de ‘qualquer’. Bachelor thesis, Faculdade de Letras da Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.
- Miranda, Sílvia. 2013. Reconstituição do ms. L da Crónica Geral de Espanha de 1344 (2.^a parte). Master thesis, Universidade de Lisboa, Lisboa, Portugal.
- Móia, Telmo. 1992a. *Aspectos da semântica do operador qualquer*. *Cadernos de Semântica* 5. Lisboa: Faculdade de Letras da Universidade de Lisboa.
- Móia, Telmo. 1992b. A Sintaxe das Orações Relativas sem Antecedente Expresso do Português. Master thesis, Faculdade de Letras da Universidade de Lisboa, Lisboa, Portugal.
- Moreno, António. 2009. Para uma caracterização enunciativa do marcador qualquer. In *Textos Seleccionados do XXIV Encontro Nacional da Associação Portuguesa de Linguística. Paper presented at the XXIV Meeting of APL, Braga, Portugal, November 20–22, 2008*. Lisboa: APL, pp. 355–65.
- Palomo, José. 1934. The Relative Combined with Querer in Old Spanish. *Hispanic Review* 2: 51–64. [CrossRef]
- Pedrosa, Marta. 2012. Reconstituição do ms. L da Crónica Geral de Espanha de 1344 (1.^a parte). Master thesis, Universidade de Lisboa, Lisboa, Portugal.
- Peres, João. 1987. Para uma semântica formal da quantificação nominal não-massiva. Ph.D. thesis, Universidade de Lisboa, Lisboa, Portugal.
- Peres, João. 2013. Semântica do sintagma nominal. In *Gramática do Português*. Edited by Eduardo Paiva Raposo, Maria Fernanda Bacelar do Nascimento, Antónia Mota, Luísa Segura and Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, vol. 1, pp. 735–813.
- Piel, Joseph-Maria, and Irene Nunes. 1988. *Demanda do Santo Graal*. Lisboa: Imprensa Nacional—Casa da Moeda.
- Pinto, Clara. 2021. Minimizers and the Syntax of Negation: A Diachronic and Comparative Approach from European Portuguese. Ph.D. thesis, Universidade de Lisboa, Lisboa, Portugal.
- Pires de Oliveira, Roberta. 2005. Qualquer e o conceito de livre-escolha. *DELTA* 21: 251–77. [CrossRef]
- Ramat, Anna. 2005. Persistence and Renewal in the Relative Pronoun Paradigm: The Case of Italian. *Folia Linguistica Historica* 26: 115–38. [CrossRef]
- Ritter, Elizabeth. 1991. Two functional categories in noun phrases: Evidence from Modern Hebrew. In *Syntax and Semantics, Perspectives on Phrase Structure: Heads and Licensing*. Edited by Susan Rothstein. San Diego: Academic Press, vol. 25, pp. 37–62.
- Rivero, María-Luisa. 1984. Diachronic Syntax and Learnability: Free Relatives in Thirteenth-Century Spanish. *Journal of Linguistics* 20: 81–129. [CrossRef]
- Rivero, María-Luisa. 1986. Dialects and diachronic syntax: Free relatives in Old Spanish. *Journal of Linguistics* 22: 443–54. [CrossRef]
- Rivero, María-Luisa. 1988. La sintaxis de qual quiere y sus variantes en el español antiguo. *Nueva Revista de Filología Hispánica* 36: 47–73. [CrossRef]
- Rizzi, Luigi. 1997. The Fine Structure of the Left Periphery. In *Elements of Grammar. Handbook in Generative Syntax*. Edited by Liliane Haegeman. Dordrecht: Kluwer, pp. 281–337.
- Rodrigues, Maria Celeste. 1992. *Dos Costumes de Santarém*. Master thesis, Faculdade de Letras da Universidade de Lisboa, Lisboa, Portugal.
- Šimík, Radek. 2020. Free relatives. In *The Wiley-Blackwell Companion to Semantics*. Edited by Daniel Gutzmann, Lisa Matthewson, Cécile Meier, Hotze Rullmann and Thomas Ede Zimmermann. Hoboken: Wiley-Blackwell, pp. 1–38.
- Toledo Neto, Sílvio. 2012. Os testemunhos portugueses do Livro de José de Arimateia e o seu lugar na tradição da Estoire del Saint Graal: Colação de exemplos. In *De cavaleiros e cavalarias. Por terras de Europa e America*. Edited by Lenia Mongelli. São Paulo: Humanitas, pp. 579–89.
- van Riemsdijk, Henk. 2006. Free relatives. In *The Blackwell Companion to Syntax*. Edited by Martin Everaert and Henk Riemsdijk. Oxford: Blackwell, vol. 2, pp. 338–82.
- Vendler, Zeno. 1967. *Linguistics in Philosophy*. Ithaca and London: Cornell University Press.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

A Diachronic Overview of the Prepositional Accusative in Portuguese

Ana Regina Calindro

Department of Linguistics and Philology, Federal University of Rio de Janeiro, Rio de Janeiro 21941-901, Brazil; anacalindro@letras.ufrj.br

Abstract: One of the several differences between Modern European Portuguese (EP) and Modern Brazilian Portuguese (BP) is the prepositional expression of complements licensed by the preposition *a*. While in EP the preposition *a* occurs in several contexts, this element has been substituted by other strategies in BP, as is extensively discussed in the literature. The aim of this paper is to investigate the historical behavior of *a*-marked prepositional accusatives (PP-ACC) in Portuguese. In order to do so, a search was conducted for PP-ACCs in the *Historical Portuguese Corpus Tycho Brahe*. The results showed an increase of PP-ACCs in the 17th century, followed by a decrease in the 18th century. Thereafter, unmarked accusatives (NP-ACC) were analyzed in the corpus, which resulted in 7756 sentences, contrasting with 624 PP-ACCs in the same contexts. This result shows that the *a*-marked accusative is far less common than bare accusatives in Historical Portuguese. Psych verbs, however, behaved differently, showing a constant increase in PP-ACCs. In EP, the preposition *a* still introduces Experiencer arguments in structures with some psych verbs (*O vinho agradou ao João*—lit. ‘The wine pleased ‘to’ John’). In BP, the preposition *a* has disappeared in psych predicates (*O vinho agradou Ø o João*—‘The wine pleased John’). In both Modern EP and BP, most PP-ACCs have become typical unmarked direct objects. In the context of psych verbs, however, structural accusative assignment has shifted to structural dative Case in Modern EP, so as to ascertain the interpretation of the Experiencer in the internal argument via the preposition *a*. While in Modern BP, the argument is not overtly marked since it receives inherent accusative case in the derivation.

Keywords: Historical Portuguese; prepositional accusative; differential object marking; preposition *a*; psych verbs

Citation: Calindro, Ana Regina. 2024. A Diachronic Overview of the Prepositional Accusative in Portuguese. *Languages* 9: 194. <https://doi.org/10.3390/languages9060194>

Academic Editor: Patrícia Amaral

Received: 5 November 2023

Revised: 5 May 2024

Accepted: 6 May 2024

Published: 24 May 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The argument structures found in Modern European Portuguese (EP) and Modern Brazilian Portuguese (BP) have undergone several changes historically (cf. Galves 2001, 2007, 2020, a.o.), especially regarding the prepositional expression of complements licensed by the preposition *a*. In EP, the preposition *a* is used to introduce indirect dative arguments, as in (1a) and (2a) below. In BP, synchronic data shows that the preposition *a* found in EP has been replaced by other prepositions with a more transparent thematic role, such as *para* and *de* in several contexts ((1b) and (2b)) (Torres Morais and Salles 2010; Torres Morais and Berlinck 2018). Additionally, some predicates with psych verbs encode their Experiencer argument with the preposition *a* in EP, but the same complement is unmarked in BP (cf. 3):

(1)	a.	A	Maria	<u>enviou</u>	uma	carta	ao João	/enviou-lhe	uma	carta.
		The	Maria	sent	a	letter	P _{a(to)} the João	/sent-CL.DAT.3SG	a	letter.
	b.	A	Maria	<u>enviou</u>	uma	carta	ao	/para o João	/ele.	
		The	Maria	sent	a	letter	P _{a(to)}	/P _{para(to)} the João	/him.	
(2)	a.	A	Maria	<u>roubou</u>	o	relógio	ao João	/roubou-lhe	o relógio.	
		The	Maria	stole	the	watch	P _{a(to)} .the João	/stole-CL.DAT.3SG	the watch.	
	b.	A	Maria	<u>roubou</u>	o	relógio	do João	/dele.		
		The	Maria	stole	the	watch	P _{de(of)} .the João	/his.		
(3)	a.	O	vinho	<u>agradou</u>	aos convidados		/agradou-lhes.			
		The	wine	pleased	P _{a(to)} .the guests		/pleased- CL.DAT.3PL			
	b.	O	vinho	<u>agradou</u>	(*a) os convidados		/os agradou	/agradou eles.		
		The	wine	pleased	P _{a(to)} .the guests		/CL.ACC.M.3PLpleased	/pleased they.ACC ¹		

The comparison between both varieties is based on the fact that they share a common historical background, as the language that arrived in Brazil in the 16th century was Classical Portuguese (CP) (Galves 2007). Over the years, CP has become EP and BP, as these varieties have parted ways in several syntactic and morphological aspects, such as the different strategies used to encode their arguments (cf. Galves 2001). Therefore, the texts produced by Portuguese authors from the 16th century on are essential to the study of both Portuguese varieties.²

Specifically, the main goal of this paper is to investigate the contexts in which accusative arguments were *a*-marked historically, as example (4) from the 17th century illustrates. Synchronic data from both BP and EP show that accusative direct objects selected by dicendi verbs such as *chamar*, ‘call’, are never marked.³

(4) (...) chama **a** **Pedro e André**, e **aos filhos** **do** **Zebedeu**.
 call P_{a(to)} Pedro and André, and P_{a(to)}.the sons of.the Zebedeu
 ‘(...) call Pedro and André, and Zebedeu’s sons.’

17th century (V_004,202.1819)⁴

The literature on Portuguese variation and change highlights that the great historical modifications in the structure of BP began to be noticed in the 18th century; some became stable in the 19th and 20th centuries, while others are still in the process of variation in current BP (Kato et al. 2009; Torres Morais and Berlinck 2018; Galves 2020). Furthermore, Galves (2020) attests that even though there are several studies with an empirical basis about the 19th and 20th centuries, there are fewer works on the 17th and 18th centuries. For that reason, an investigation on Historical Portuguese from the 16th to the 19th century is very relevant to understanding the contexts and possible reasons for the changes attested in EP and BP. Accordingly, to achieve the main aim of this paper, 19 texts from the Historical Portuguese corpus, organized by the *Tycho Brahe Project* at Unicamp, were analyzed.⁵ These texts cover precisely the period between the 16th and 19th centuries. The data collected provided clear evidence, especially, for psych verbs, which showed a distinct behavior in licensing arguments with the preposition *a* throughout the centuries, as we will discuss in more detail in the results section.

As shown in examples (1) to (3), the preposition *a* introduces indirect dative arguments in Modern EP. This element has been analyzed as a dummy marker that licenses the dative argument. According to Torres Morais (2007), there are two pieces of evidence for the *a*-DP dative status in EP: firstly, the possibility that the argument will be displaced to the topic position, maintaining the preposition (*Ao João, a Maria enviou uma carta*, ‘To João, Maria sent a letter’). Secondly, the fact that the *a*-DP in EP always alternates with dative clitics (*lhe/lhes*), as attested to in the examples above.

In Modern BP, on the other hand, the dative clitic *lhe(s)* has been replaced by other strategies, such as 3rd person pronouns preceded by contentful prepositions, such as *de/para ele(s)/ela(s)*, ‘of/to him/her/them’, as exemplified above (Calindro 2015, 2016; Torres Morais and Berlinck 2018; Bazenga and Rodrigues 2019).⁶ This is evidence that case assignment in

BP is different from EP, since the morphological case in the form of the clitic *lhe* has been lost and the dummy preposition *a* has been replaced by contentful prepositions that assign an oblique case to the argument they introduce or has disappeared, as in the context of psych verbs (cf. 3b) (Calindro 2015, 2020).

As for psych verbs specifically, their structure is transitive (cf. 3), not ditransitive as in (1) and (2). The direct object in (3) is the Experiencer of the event, overtly marked in EP (3a) but not in BP (3b). Most of the literature on variation and change in Portuguese has addressed ditransitive verbs with Goal indirect objects introduced by the preposition *a* in EP and *para* in BP. Transitive contexts with prepositional accusatives seem to be less explored (Ramos 1992; Gibrail 2003; Calindro 2017). Even less discussed are syntactic issues related to the licensing of the preposition *a* in psych predicates with Experiencer arguments (cf. Figueiredo Silva 2007; Carneiro and Naves 2010). This fact, coupled with the corpus analysis, led us to a second goal, which is to analyze the reasons why EP and BP mark Experiencers in a distinct way. We are assuming that, with psych verbs, since the Experiencer argument is generated as an internal argument, a conflict is displayed between case and theta assignment. To solve this divergence, the argument is *a*-marked, in some contexts, in predicates with psych verbs in EP. In BP, on the other hand, the Experiencer argument enters the derivation with inherent accusative case, and the Theme, such as *vinho*, ‘wine’, in (3b) is raised to receive the nominative case (in Chomsky’s (1981) terms).⁷

Finally, in the literature, *a*-marked accusative arguments are sometimes addressed as PP-ACC and/or are treated as differential object marking (DOM), so, in the next section, we present an overall view of how these phenomena are related and addressed in the studies about Portuguese, followed by a discussion on psych verbs. In Section 3, the research methods are exposed. Section 4 shares the data results and a discussion on the categorical status of *a*-marked objects. Based on these considerations, our proposal is put forward in Section 5, followed by the final remarks in the last section.

2. PP-ACC and DOM in Historical Portuguese

It seems not to be clear in the literature how and why accusative *a*-marking/DOM appeared in Romance. According to Bossong (1991), many languages employ what he calls DOM in order to syntactically encode subsets of direct objects. This phenomenon is present in several Romance languages, such as Spanish, Catalan, Romanian, and Italo Romance varieties (Andriani 2015). With the exception of Romanian, the DOM marker is homophonous to the dative used to introduce indirect objects, as the preposition *a* in Portuguese. Therefore, the data presented previously indicate the possibility of *a*-marked objects in Portuguese being instances of DOM. On the surface, DOM seems to share features with prototypical prepositions, hence this phenomenon may also be found in the literature about Romance as PP-ACC. According to Fábregas (2013), DOM only partially shares features of prototypical prepositions. Additionally, Gerards (2020) points out that the diachronic data on DOM poses a problem, as it has been attested that *a*-marking first appeared with [+human] personal pronouns—*Isso agradou a ela*, ‘lit’. This pleased ‘to’ her (Döhla 2014); that is, specifically, one of the few elements that display overt case marking in all Romance languages and which, for that reason would not need additional Case marking.

Overt marking seems to emerge when the argument is autonomous from the verb, i.e., with self-constituent objects. In fact, Bossong (1991) discusses that being dependent or independent of the verbs is a factor to be taken into account when analyzing instances of DOM. This leads us to the type of verb the argument is associated with. For example, in Spanish, there are two verbs that always select a marked argument: *buscar* (to look for) and *querer* (to want) (Bossong 1991, pp. 158–59).⁸

Actually, Bossong (1991, p. 160) claims that DOM is related to three domains: i. the domain of inherence – [+deictic] < [+proper] < {[+parent]} < [+human] < {[+person]} < [+animated] < [+discrete]; ii. the domain of reference – [+individuality] < [+/-referentiality]/ [+/-definiteness]; and iii. the domain of constituency – [+/-independent existence]/ [+/-pragmatic constituency]. The author argues that it is common to use a mixture of

these dimensions when analyzing overtly marked arguments. For Spanish and Romanian, Bossong claims inherence is the dominant factor required for DOM to occur. In this paper, as further explained in the methods section, we analyzed the data via verb classes, so we will consider what the author calls *constituence*—the argument being dependent or not on the verb—the main factor for our analysis at this point of the research. López (2016) also claims that syntactic configuration is one of the aspects that trigger DOM in Romance, along with features of the verbs and features of the object.

Synchronic EP data shows overtly marked arguments in predicates with some psych verbs (cf. 3a). Hence, the main question that emerged from the examples presented in the introduction was when and how this marking appeared historically and whether structures with verbs other than psych verbs also present *a*-marked arguments. Therefore, as will be explained in more detail in the methodology and results sections, the first search in the corpus was for all the accusatives marked by the preposition *a* that had been labeled by the *Tycho Brahe* group as PP-ACCs. The results confirmed that psych verbs behaved differently from other verb groups. This was indeed a very relevant context to address in order to understand how *a*-marked arguments behave in Portuguese. Nevertheless, before addressing the types of verbs present in the corpus, another question required consideration—should we consider the *a*-marked arguments instances of PP-ACCs or DOM in Portuguese?

Given the DOM characteristics put forth by the aforementioned authors, PP-ACC seems to be part DOM, which can be seen as a wider phenomenon, as it is related to several factors. Gerards (2020, p. 1) observes that the Romance-specific term PP-ACC may be misleading and inferior, so the author suggests a “categorically agnostic label *flagging DOM*”. This seems to be a valid solution; however, it is not the aim of this paper to solve this issue. What concerns us at the moment is the historical change Portuguese has undergone, considering that our results showed a decrease in *a*-marked elements in general, which is in conformity with the synchronic Portuguese data, as the vast majority of direct objects are unmarked accusatives. Moreover, to address the specific differences regarding PP-ACC and DOM, it is important to analyze the features of the objects, as suggested by López (2016) and Bossong (1991), who consider the domains of inherence and reference when analyzing DOM arguments, as mentioned above. It is not in the scope of this paper, however, to analyze the specific features of the *a*-marked arguments, such as definiteness, referentiality, and animacy (to name a few features mentioned by Bossong (1991)). We intend to do so in future work.⁹

As shown in the introduction, argument marking changes depending on the verb class that the predicate belongs to. This first observation guided the research towards the analysis of all *a*-marked arguments, followed by the analysis of the different verb classes the marked object appeared with. Then, it led to the analysis of psych verb predicates specifically, since the pattern presented in this context differs from the other ones, as already explained. Therefore, alongside the disappearance of PP-ACCs in BP, one of the aims of this research was to understand the context in which the Experiencer argument ceased to be marked in all psych predicates in BP, as exemplified in (3b). In the following section, the properties of psych verbs will be addressed.

Returning to PP-ACC/DOM, historically, overtly marked arguments in Romance appeared with topicalized personal pronouns used anaphorically and emphatically (Döhla 2014, p. 272):

Proto-Romance—11th century

(5) **a** **mibe** **tu** no(n) querés.
 P_{a(to)}. me you neg love
 ‘(but) me you don’t love’

(adapted from Corriente 1997, p. 319 *apud* Döhla 2014, p. 272)

In Modern Spanish, where, differently from Portuguese, DOM is a core phenomenon, definite animate DPs, for instance, are always associated with DOM. Manzini and Franco

(2016) claim that all definite arguments display DOM in Spanish, as these arguments are attached VP-internally via DOM. On the other hand, DOM with indefinite animate arguments depends on the verb it is associated with, as some verbs prefer DOM. The authors argue that some action, psych, and perception predicates, which have non-affected objects, seem to have more transparent v-V structures. These are exactly the classes with more *a*-marked cases in the corpus analyzed here, which will be discussed in the results section.

In Spanish, the development of the preposition *a* as an accusative marker used to be seen in the literature as a singular historical event that demanded a specific historical explanation. Bossong (1991) points out, however, that this manifestation is one of the most common ways case marking changes in natural languages. According to the author, if subject and object are no longer distinguishable in the system, morphological marking is then substituted by a new differential system, which may in turn “become non-differential again by the continuous extension of the sphere of positive object marking; at this point of the evolution, the life cycle of case marking may start anew” (Bossong 1991, p. 152), which seems to be the case of the data analyzed in this paper. Therefore, *a*-marking was a way to overtly case-mark atypical direct objects, but as Portuguese progressed, this marking was no longer necessary, because SVO order became even more common than VOS, for instance, as we will discuss further in this paper.

Specifically, when it is not possible to distinguish between subject and object by formal means, two things may happen. First, such as what happened from Latin to Romance, when morphological marking is eliminated, it may just disappear, as in French, which has replaced morphological marking with structural Case marking (Bossong 1991, p. 146). The same is true for Nominative and Accusative in Portuguese in general, as we see further in this paper. The second option is to resort to what Bossong calls *a new grammemic marking*, such as *a*-marking.

According to Döhla (2014), DOM in Modern EP and BP is considered to be a marginal phenomenon. The author claims that DOM occurred in Old Portuguese due to three main reasons: i. parallelism (6)—when one *a*-marked object is followed by another one; ii. left dislocation—to emphasize the topicalized patient argument (7); and VSO order (8):

Old Portuguese

(6) (...) **tendes** em vossa ajuda muy certos **a** **mym** e **ao** **Conde d’Ourem.**
 (you) in yours help very sure P_{a(to)} me and P_{a(to)} the Count of Ourem
 have in yours help very sure P_{a(to)} me and P_{a(to)} the Count of Ourem
 ‘you have me and the Count of Ourem to help you.’

(7) **aos** **proues** e **mjnguados** **sostinha.**
 P_{a(to)}. the poor and neglected support
 ‘(he/she) supported the poor and the neglected.’

(8) **amando** **mais** **as** **maes** **a** **seos** **filhos.**
 loving more the mothers P_{a(to)} their children
 ‘the mothers loving more their children.’

(Delille 1970, pp. 36, 39, 42 *apud* Döhla 2014, pp. 274–75)

As Döhla, Ramos (1992), and Gibrail (2003) also verified, there was a higher occurrence of *a*-marked arguments with VSO in the Historical Portuguese data they analyzed.¹⁰ As mentioned before, there seems to be an overlap between what is understood as DOM and what is understood as PP-ACC. These authors, for example, apply the name PP-ACC to the same examples Döhla refers to as DOM.

The numbers presented by Ramos and Gibrail show an increase in PP-ACCs in the VSO context in the 18th century. The authors claim that object marking in VSO sentences aims to disambiguate the syntactic function of NPs. One hypothesis is that as Portuguese started to present a more fixed SVO order (Galves 2020), object marking ceased to be necessary.

As mentioned before, our intent in this section was to present previous works on *a*-marked arguments in Historical Portuguese. Since these works address *a*-marked direct objects both as PP-ACC and DOM, it was important to give an overall view of these phenomena. Given what was presented in this subsection, there is indeed an overlap between the concepts of PP-ACC and DOM, as both strategies are responsible for overtly case-marking direct objects and may be related to the syntactic configuration and features of the verbs (López 2016). So, as specific features of the *a*-marked NPs in the data analyzed will not be discussed in this paper, we will continue to address *a*-marking arguments as PP-ACC. We leave the discussion on DOM vs. PP-ACC for future work, otherwise there would be too many variables to be considered here.¹¹

Psych Verbs

In this paper, one of the main focuses is on structures with psych verbs where the Experiencer argument is still marked in Modern EP but unmarked in Modern BP.

In predicates with psych verbs, the Experiencer argument must be a person, hence [+animate] and [+human], who experiences a mental state (Belletti and Rizzi 1988). It can be either the subject (9) or the object (10) of the event. Additionally, the Theme argument, which is the content or the object of the mental state, may also be the subject (11a) or the object (11b). Even though these verbs fall into the same category, Belletti and Rizzi (1988) separate them into the three types, exemplified in (9) to (11), due to their distinct theta grids.

Italian

- (9) Gianni teme questo.
Gianni fears this
- (10) Questo preoccupa Gianni.
This worries Gianni
- (11) a. Questo piace a Gianni.
This pleases P_{a(to)} Gianni
- b. A Gianni piace questo.
P_{a(to)} Gianni pleases this

(Belletti and Rizzi 1988, pp. 291–92)

In Portuguese, sentence (12) is equivalent to (9). The examples in (13) show that there are also two possible mappings of the structure of some psych predicates (Carneiro and Naves 2010); (13a) is equivalent to (10), but the Theme argument can also be the content of the subject mental state, as we can see in (13b). In addition, the category of the verb *assustar*, ‘frighten’, also has an agentive component in Portuguese, since the [+human] subject can act intentionally, as in *Maria assustou João*, ‘Maria frightened John’ (Naves 2005, p. 122).

- (12) João_{EXP} teme a aranha_{THEME}.
João fears the spider
‘John fears the spider.’
- (13) a. A aranha_{AGENT} assusta o João_{EXP}.
The spider frightens the João
‘The spider frightens João.’
- b. O João_{EXP} se assusta com a aranha_{THEME}.
The João CL.3SG frightened with the spider
‘João gets frightened with the spider.’

Among the many issues related to psych verbs, recall that we are specifically interested in the category of psych verbs such as *agradar*, ‘please’ (cf. 3)¹², for three distinct reasons: firstly, the argument structure of these verbs includes object marking by the preposition *a* in EP, but it does not in modern BP; secondly, the data from the corpus analyzed showed that this context displays *a*-marked arguments throughout the four centuries; finally, the

constituent order in these predicates is different in EP and BP, as shown specifically in examples (15) and (18):

- | | | | | | | |
|------|--------------------|-------------|----------------|----------------------|--------------|------|
| (14) | O | filme | <u>agradou</u> | ao | João. | (EP) |
| | The | movie | pleased | P _{a(to)} | João | |
| (15) | Ao | João | <u>agradou</u> | o | filme. | (EP) |
| | P _{a(to)} | João | pleased | the | movie | |
| (16) | O | filme | <u>agradou</u> | -lhe. | | (EP) |
| | The | movie | pleased | -CL.DAT.3SG | | |
| (17) | O | filme | <u>agradou</u> | o | João. | (BP) |
| | The | movie | pleased | the | John | |
| (18) | *O | João | <u>agradou</u> | o | filme. | (BP) |
| | The | João | pleased | the | movie | |
| (19) | O | filme | <u>agradou</u> | -o. | | (BP) |
| | The | movie | pleased | -CL.ACC.3SG. | | |
| (20) | O | filme | <u>agradou</u> | ele. | | (BP) |
| | The | movie | pleased | he.ACC ¹³ | | |

Observe that the structures with *agradar* in EP, with a prepositional-marked Experiencer argument, are the same as with *piacere* in Italian (cf. 11). In BP, however, the Experience is unmarked in the structures with *agradar*, similarly to (10) in Italian. Moreover, if we substitute the DP Theme *filme*, ‘movie’, in (17) for a [+animate] DP, such as *João agrada Maria*, ‘João pleases Maria’, there is the possibility of an agentive interpretation, which is not possible in (17) with the Theme *filme*, ‘movie’, but possible in (13a) with *assustar*, ‘frighten’. Thus, *agradar* would be part of a fourth class that combines the characteristics of (13a) and (17), which does not exist in Italian. In addition, the argument in BP is pronominalized with accusative elements with the abstract accusative case (cf. 19 and 20), while EP shows a dative counterpart (16).

Having presented some considerations on PP-ACC, DOM, and psych verbs crosslinguistically, in the next section we present the methodology employed to deal with the historical data at hand.

3. Materials and Methods

In this section, the data collected from the *Tycho Brahe Corpus* is presented. The corpus consists of 88 texts written by Portuguese and Brazilian authors who were born between 1380 and 1978. Out of the 88 texts, 19 texts had already been syntactically annotated at the time of the data collection (today there are 27 texts annotated syntactically), which adds up to a total of 819,932 words (Galves et al. 2017).

The main purpose of analyzing Portuguese historical data was to examine the contexts in which *a*-marked direct objects (DOs) presented variation between the 16th and 19th centuries. At first, a search was carried out for all occurrences of prepositional accusatives (PP-ACC-marked DOs) in the corpus, in sentences such as (4) (renumbered as (21)), below:

- (21) (...) chama a Pedro e André, e aos filhos do Zebedeu.
 call P_{a(to)} Pedro and André, and P_{a(to)}. the sons of.the Zebedeu
 ‘(...) call Pedro and André, and Zebedeu’s sons.’

17th century (V_004,202.1819)

The relevant information about the texts examined in this paper is in Table 1:

Table 1. Analyzed Texts.

	Text Code	Author	Date of Birth	Text Title	Number of Words
16th Century	c_007	Diogo do Couto	1542	Décadas (livro de crônicas)	47,605
	g_001	Manuel de Galhegos	1597	Gazeta	28,839
	g_008	Pero Magalhães de Gandavo	1502	História da Província de Santa Cruz	22,944
	p_001	Fernão Mendes Pinto	1510	Periginação	47,580
	s_001	Luis de Sousa	1556	A vida de Frei Bertolameu dos Mártires	53,986
17th Century	b_001	André de Barros	1675	Vida do apostólico padre Antonio Vieira	52,055
	b_008	José da Cunha Brocado	1651	Cartas	35,058
	c_002	Maria do Céu	1658	Vida e Morte de Madre Helena da Cruz	27,419
	v_002	Padre Antonio Vieira	1608	Cartas	57,088
	v_004	Padre Antonio Vieira	1608	Sermões	53,855
18th Century	a_001	Matias Aires	1705	Reflexões sobre a vaidade dos homens	56,479
	a_004	Marquesa de Alorna	1750	Cartas	49,900
	c_001	Cavaleiro de Oliveira (Francisco Xavier)	1702	Cartas, Cavaleiro Oliveira	51,234
	c_005	José Daniel Rodrigues da Costa	1757	Entremezes de Cordel	24,252
19th Century	a_003	Marquês de Fronteira e d'Alorna	1802	Memórias do Marquês de Fronteira e d'Alorna	54,588
	b_005	Camilo Castelo Branco	1826	Maria Moisés	20,548
	g_004	Almeida Garret	1799	Teatro	44,252
	o_001	Ramalho Ortigão	1836	Cartas a Emília, Ramalho Ortigão	32,441

All authors mentioned above were Portuguese.¹⁴ The 19 annotated texts add up to 39,761 sentences distributed as follows: 8930 in the 16th century; 8948 in the 17th century; 10,967 in the 18th century; and 10,916 in the 19th century.

In addition to a large amount of data available in the *Tycho Brahe Corpus*, the syntactic annotation of the corpus makes it possible to use the Corpus Search computational tool (available on the project website), which allows searching for specific syntactic contexts through queries.¹⁵ In the 39,761 sentences from the corpus, 624 sentences exhibited PP-ACCs. The results showed a vast variety of verbs accompanied by an overtly marked accusative argument.

After looking for all PP-ACCs, the data were divided into different verb classes, based on studies on Spanish and Catalan (Pineda 2012, 2017) and Portuguese (Cançado 1996, 2013). Pineda's work, for instance, focuses on a similar ongoing change in Catalan, but from dative to accusative; these studies were used as reference in order to divide the data collected into different groups, so that it would be possible to run another search for NP-ACCs in the same type of predicates in which the PP-ACCs were found. This division resulted in 11 distinct verb classes: contact; dicendi; psychological; social interaction;¹⁶ transfer; transfer related to values; reverse transfer; transfer of knowledge; related to places; quantity; and other types of verbs. See some examples from the 17th century data below:

Contact Verbs

- (22) Para tirar toda a duvida, oiçamos **ao** **mesmo** **Christo**.
 To solve every the doubt listen.1PL P_{a(to)} the same Christ
 'To solve every doubt, let's listen to the same Christ.'
 (V_004,70.149)

Dicendi Verbs

- (23) mandou chamar **a** **Dom** **Duarte** **de** **Castellobranco** seu cunhado,
 asked call P_{a(to)} Dom Duarte de Castellobranco his brother-in-law,
 marido de sua segunda irmã Dona Luiza de Mendonça.
 husband of his second sister Dona Luiza de Mendonça
 '(someone) sent for Dom Duarte de Castellobranco his brother-in-law,
 husband of his second sister Dona Luiza de Mendonça.'
 (C_002,139.107)

Psych Verbs

- (24) Tambem lhe era duro deixar nos claustros adonde se criara,
 Too her was hard to leave in.the cloisters where CL.3rd was raised
 companheyras, e amigas de tantos anos, e na corte a uma irmã
 colleagues, and friends of many years, and in.the court P_{a(to)} a sister
a quem tanto amaua.
 P_{a(to)} whom so much loved
 'It also hard for her to leave colleagues and friends of many years in the
 cloisters where she was raised, and in the court a sister who she loved so much.'
 (C_002, 214.1039)

Verbs related to places

- (25) (...) a comunicação com Deos; o retiro de toda a creatura, alguma doente,
 the communication with God; the retreat of all the creature, some sick,
a quem visitaua precisada da charidade com as mãos e olhos prezos.
 P_{a(to)} whom visited in need of.the charity with the hands and eyes fixed.
 '(...) the communication with God; the retreat of all creatures, some sick,
 who (he/she) visited with the hands and fixed eyes.'
 (C_002,165.427)

Quantity related verbs

- (26) Resta nos o que **a** **tudo** excede a notícia das gentes,
 Remain us what P_{a(to)} everything exceeds the news of.the peoples,
 que habitam uma, e outra margem;
 that inhabit one, and other margin
 'It is left for us everything related to the news about the peoples that live one
 margin and the other.'
 (B_001_PSD, 88.700)

Verbs of Transfer

- (27) (...) mandou Dom Jorge Mascarenhas, Marquês de Montalvão, e Vice-Rei
 sent Dom Jorge Mascarenhas, Marquis of Montalvão, and Vice-King
 daquele Estado, **a seu filho Dom Fernando Mascarenhas** no ano de 1641
 of.that State, P_{a(to)} his son Dom Fernando Mascarenhas in.the year of 1641
 'Dom Jorge Mascarenhas, Marquis of Montalvão, and Vice-King of that
 State sent his son Dom Fernando Mascarenhas in the year of 1641.'
 (B_001_PSD,19.162)

Verbs related to transfer of values

- (28) Que esta Praça Diuina assim como alumea aos peccadores, premeya
 That this Square Diuina just as enlightes P_{a(to)} the sinners awards
aos Justos, a huns abrindo os olhos, **a outros** enchendo o coração.
 P_{a(to)}.the Just, P_{a(to)} some opening the eyes P_{a(to)} others fulfilling the heart
 'May this Divine Square just as it enlightes the sinners, awards the Just, opens
 the eyes of some and fulfills the heart of others.'
 (C_002,205.928)

Verbs of reverse transfer

(29) O Padre António Vieira com excessivas expressões recebeu nos braços
 The Priest António Vieira with excessive expressions received in.the arms
aos dois Padres, como a irmãos, como a filhos, e como
 P_{a(to)}.the two Priests, as P_{a(to)} brothers, as P_{a(to)} children, and as
a heróicos companheiros de sua glória;
 P_{a(to)} heroic fellows of his glory;
 ‘The Priest António Vieira with excessive expressions received two priests in the
 arms, as brothers, as children, and as heroic fellows of his glory.’

(B_001_PSD,193.1528)

Verbs of transfer of knowledge

(30) Com estes foi um Índio Cristão antigo, **a quem** instruíram
 with these went an Indian Christian ancient, P_{a(to)} whom instruct
 os Padres, e adestraram na forma do Baptismo, para que
 the Priests, and train in.the way of.the Baptism, in order that
 to
 nos casos precisos os instruisse, e baptizasse;
 in.the cases necessary them instruct, and baptized;
 ‘An ancient Christian Indian, who was instructed by the Priests, and trained in
 the ways of the Baptism, went with them, in order to instruct and baptize the
 cases needed.’

(B_001_PSD,196.1545)

Verbs of social interaction

(31) Teue grandissimos desejos da solidão, e de imitar nella
 There were enormous desires of.the loneliness, and of imitate in.her
aos antigos Annacoretas;
 P_{a(to)} old anchorites;
 ‘There were enormous loneliness desires, and of imitating the old anchorites.’

(C_002,206.949)

Other types of verbs

(32) Que esta Praça Diuina assim como alumea **aos peccadores,** premeya
 That this Square Diuina just as enlightens P_{a(to)}.the sinners awards
 aos Justos, a huns abrindo os olhos, a outros enchendo o coração.
 P_{a(to)}.the Just, P_{a(to)} some opening the eyes P_{a(to)} others fulfilling the heart
 ‘May this Divine Square just as it enlightens the sinners, awards the Just, opens the
 eyes of some and fulfills the heart of others.’

(C_002,205.928)

In the next section, the details of the data collected will be presented, in order to analyze the behavior and the contexts in which PP-ACC cases were found in Historical Portuguese.

4. Results and Discussion

The search conducted in the corpus resulted in a total of 624 sentences with PP-ACCs, distributed throughout the centuries as shown in Figure 1: ¹⁷

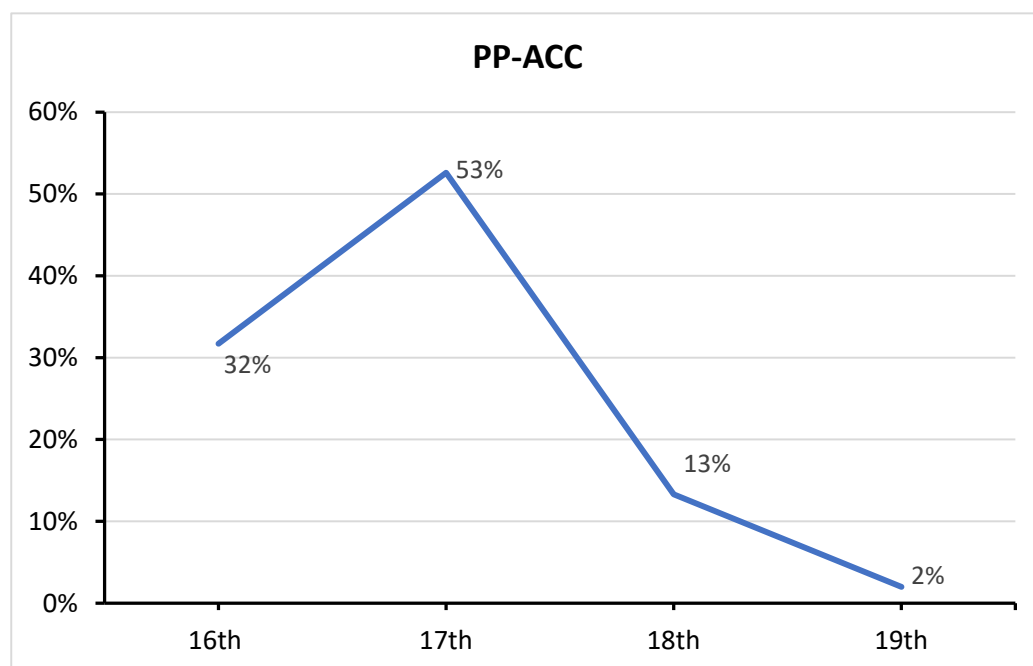


Figure 1. PP-ACC results.

The data show that PP-ACC occurrences increased in the 17th century; however, there followed an immediate fall in the 18th century (a preliminary analysis of this data was presented in Calindro 2017).

Gibrail (2003), who also analyzed the *Tycho Brahe Corpus*, attests to a high number of PP-ACCs with the NP ‘God’, as well as the combination *guardar a Deus*, ‘keep/save God’. In fact, in the data analyzed here, there were 96 examples of this structure in the 17th century, so these examples were excluded from this analysis. Moreover, it seems that such construction has become a crystallized expression even in Modern BP, in which sentences like *adorar a Deus*, lit. ‘worship ‘to’ God’, are still present in vernacular BP with an *a*-marked DO; hence, it does not show the variation and change process being addressed here.¹⁸

In the data we are concerned with, one explanation for the decrease in the PP-ACC cases in the 18th century may be the change in word order in Portuguese. Galves (2020, p. 21) attests to a rise in the VS order in the 17th century, followed by its fall in the next century. In fact, Ramos (1992) and Gibrail (2003) showed that PP-ACC was more productive in VS contexts.

Therefore, to verify whether word order influenced the occurrence of PP-ACCs in the corpus analyzed in this paper, the data was divided into the following contexts: SVO, VOS. Interestingly, the results showed that the order of the constituents does not seem to be relevant for *a*-marking the argument, as they presented very similar percentages, especially in the 16th (13% for both VS and SV) and the 17th century (15% for VS and 14% for SV). As mentioned before, we may assume that the accusatives ceased to be marked when the fixed SVO became more common in Portuguese, so we intend to look in more detail at the data collected in the future.

At the moment, two other facts are striking in the corpus analyzed: i. the quantity of PP-ACCs (624) in comparison with NP-ACCs (7756); ii. the increase in *a*-marked arguments in the 17th century, followed by an immediate fall in the 18th century, and the few cases found in the 19th century as well.

Ramos (1992), Gibrail (2003), and Döhla (2014) argue that Portuguese was greatly influenced by Spanish in the 17th century; therefore, these authors assume that this may explain why there was an increase in PP-ACCs in this particular century. Indeed, DOM is still a very productive phenomenon in Modern Spanish (López 2012, 2016; Fábregas 2013, a.o.). According to Gerards (2020, p. 2): “Spanish is the Romance language with the most

advanced (...) DOM system". Even Gil Vicente, a canonical Portuguese author, wrote plays in Spanish.

However, it does not mean this influence reached all the Portuguese population at that time. Galves (2020) observes that the written variety encountered in the sources available for diachronic empirical research corresponds to the written and spoken language of the dominant classes, who were probably the only ones who had access to literature. This register was presumably quite far from the oral language of the common people, who were mostly illiterate. Actually, the corpus we are dealing with is even further away from popular variety, insofar as it is composed of renowned authors, such as Camilo Castelo Branco and Almeida Garrett, but, unfortunately, this is the data we have available when it comes to Portuguese from the 16th to the 19th centuries.

Inasmuch as Spanish may have played a part in the increase of PP-ACCs in the 17th century, the smaller quantity of *a*-marked DOs in comparison to unmarked ones strongly suggests that there are language internal issues responsible for this marking.

Since the data was divided into 11 verb classes, as mentioned in the methods section, each context was analyzed separately. The data showed that the quantity of *a*-marked DOs with psych verbs and social interactional verbs has remained a dominant factor throughout the centuries when compared to all the other nine verb classes combined, as we can see in Figure 2 (see also Calindro 2017). From the 328 overtly marked arguments in the 17th century, 33% (109) are sentences with psych verbs, 17% (57) are sentences with social interaction verbs, and 49% (162) are the other verb classes. Hence, the two former verb classes seem to be more relevant contexts for *a*-marked arguments to occur in. This is compatible with the fact that in Modern EP the Experiencer argument of some psych verbs is still introduced by the preposition *a* (cf. 6).¹⁹

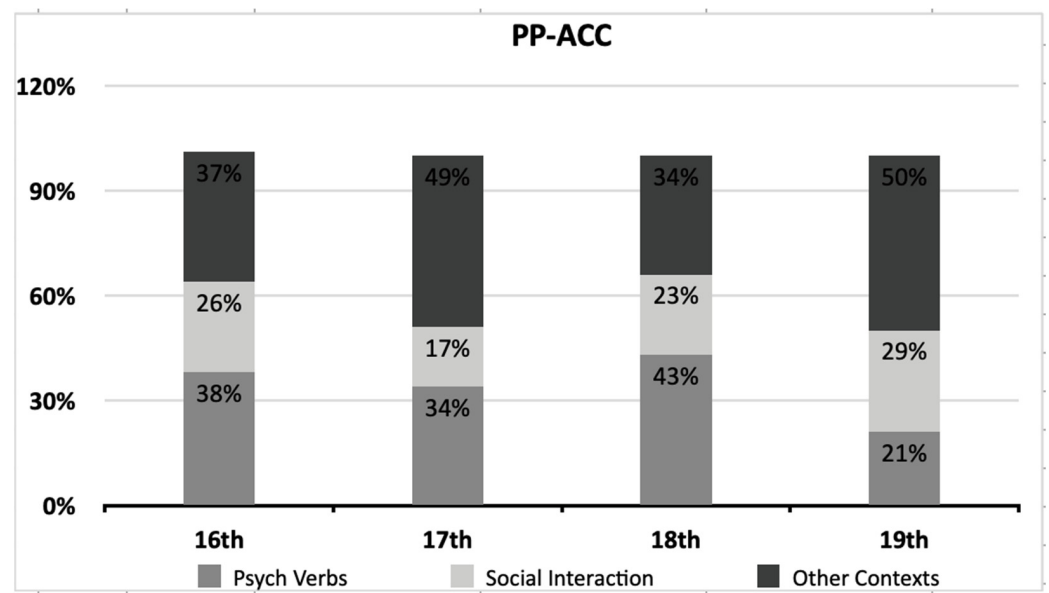


Figure 2. PP-ACC with psych verbs; social interaction verbs; other verbs.

When we analyze only the context of psych verbs, we notice a steady increase, proportionally, in PP-ACCs throughout the centuries, from 7.25% to 33.3% of the total data, as we can see in Table 2, when we compare psych verbs with PP-ACC and NP-ACC, followed by examples (33) and (34) from the 18th century, which illustrate PP-ACC and NP-ACC associated to psych verbs, respectively:²⁰

Table 2. PP-ACC/NP-ACC with psych verbs throughout the centuries.

	16th	17th	18th	19th
PP-ACC	7.2% (76)	15.8% (110)	28% (35)	33.3% (3)
NP-ACC	92.3% (975)	84.2% (586)	72% (90)	66.7% (6)
Total	1051	696	125	9

- (33) (...) (ela) poderá muito bem agradar **aos espectadores** com um bom quadro.
 may very well please P_{a(to)}.the viewers with a good painting.
 ‘(she) may very well please the viewers with a good painting.’
 (C_001, 140.1)
- (34) (...) e os discursos gerais não podem ofender **os particulares** que são discretos.
 and the speeches public not can offend the private citizens who are discreet
 ‘and the public speeches cannot offend the private citizens who are discreet.’
 (C_001, 20.18)

In the context of psych verbs, the Experiencer argument has referential properties typical of subjects. Thus, in order to guarantee its referential properties, *a*-marking was preserved in EP, while in BP it became unmarked. As psych verbs are two-place predicates, due to economy reasons, the fixed SVO order, attested by Galves (2020), seemed to be enough for the Experiencer to be read as accusative and not nominative in BP, so *a*-marking was not necessary anymore. We will return to this discussion in the next section.

5. Proposal

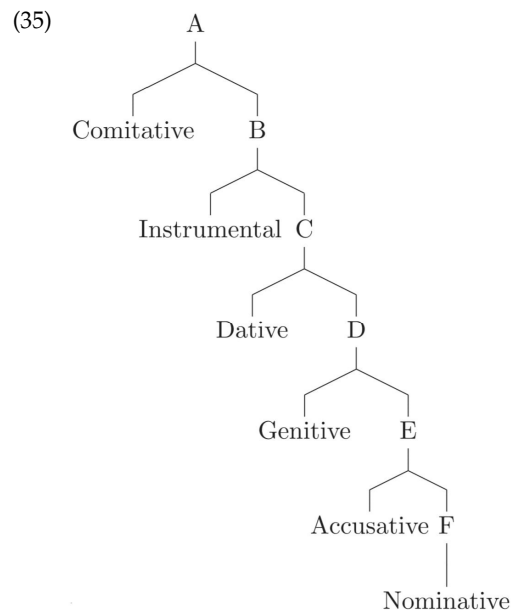
The differences in licensing indirect arguments in BP and EP led us to analyze the context in which the objects were traditionally labeled prepositional accusatives in Historical Portuguese. Recall that the starting point of this paper was the differences between Modern BP and EP, regarding *a*-marked objects in psych predicates (cf. 3). On the surface, *a*-marking has remained in psych predicates in EP (cf. 3a) and has disappeared in the same context in BP (cf. 3b).

As briefly discussed in the introduction, in Modern EP, the preposition *a* is a dummy dative marker that enters the derivation to assign dative Case to this element (for more details, cf. Torres Morais 2007; Figueiredo Silva 2007; Calindro 2015, 2016, 2020), similar to Spanish (Cuervo 2003) and other Romance languages. The dative case is confirmed by the possibility of the argument being displaced, as exemplified in (14), and for the alternation between the marked object (*a*-DP) and the dative clitic *lhe*. This alternation is no longer part of Modern BP, as the clitic *lhe* has been substituted by other strategies, as shown in examples (1b), (2b) and (3b) in the introduction (Calindro 2015); thus, the arguments introduced by *para* receive the structural oblique Case, while the complements of psych verbs are unmarked accusative, i.e., typical direct objects.

The examples in BP, EP, and Italian show crosslinguistic evidence that dative and accusative complements entail a different distribution. According to Manzini and Franco (2016), datives arise in the syntax to reflect a distinct mapping of Participant internal arguments in the event structure when necessary, as the presence of the dative may be due to the activation of a split eventive structure (Svenonius 2002). Manzini and Franco (2016) make an important distinction between Goal datives in ditransitive predicates, as the datives we saw in (1a), and DOM datives. The former are constituents required by the event, while the latter are constituents required by referential properties of the internal argument (similar to Bossong’s and López’s claims addressed previously in Section 2). Thus, based on Svenonius, Manzini, and Franco, DOM datives reflect neither a morphological regularity nor an accident. DOM may be a way to overtly assign case to internal arguments that show referential properties typical of subjects, for instance (cf. Bossong 1991; Gerards 2020).

The data analyzed here showed this may be precisely the case with psych verbs, as the Experiencer argument in these structures remained *a*-marked over the years when it does not enter the derivation with its inherent case alongside its thematic role. Additionally, its

case seems to have shifted from Accusative to Dative. This shift can be explained away using nanosyntax. In this approach, each syntactic–semantic feature is an independent head that projects (Baunaz and Lander 2018, p. 5); thus, cases can be decomposed into more primitive features and hierarchically organized (Caha 2009; Hardarson 2016), as we can see in (35):



(Caha 2009, p. 23)

In short, Caha (2009) observes that syncretic morphology in case assignment is not incidental across languages, as exemplified by the syncretism in Russian colored in gray in Table 3.

Table 3. Syncretism in Russian.

	Table, pl.	Student, sg.	We	Bridge, sg.	Good, f.sg
NOM	STOL-Y	student-∅	my	most-∅	xoroš-aja
ACC	STOL-Y	STUDENT-A	NA-S	most-∅	xoroš-uju
GEN	stol-ov	STUDENT-A	NA-S	most-a	XOROŠ-EJ
PREP	stol-ax	student-e	NA-S	MOST-U	XOROŠ-EJ
DAT	stol-am	student-u	na-m	MOST-U	XOROŠ-EJ
INS	stol-ami	student-om	na-mi	most-om	XOROŠ-EJ

Source: Caha (2009, p. 13).

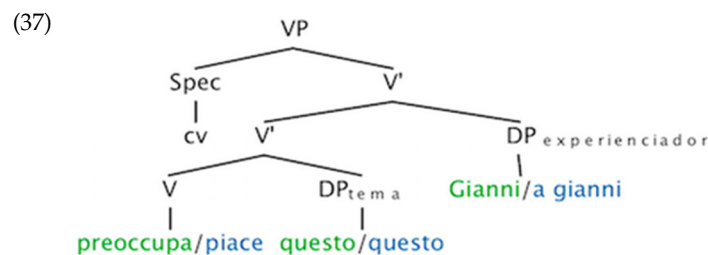
The crosslinguistic data analyzed by Caha (2009) shows a contiguity relation between cases. Nominative pronouns and nominals can be syncretic with accusatives, for instance, but they will not be syncretic with genitive pronouns if the accusative has a different form. In this sense, Hardarson (2016) proposes a slightly different hierarchy of case features from Caha’s (cf. 36), which shows a direct contiguity between accusative and dative (without the interference of the genitive):

- (36) NOM > ACC > DAT > GEN > LOC > ABL/INS>...

According to Hardarson, the dative shares features directly with the accusative, but it is more specific than the accusative, as it is higher in the hierarchy and therefore composed of more features (cf. 36), including, for instance, accusative features.

In the context of psych verbs in EP, it seems that the historical PP-ACC underwent a specialization, which resulted in its accusative features being incorporated into the dative. Consequently, the Experiencer argument, which shows referential properties typical of subjects, displays an *a*-marked dative case in Modern EP to be mapped as an internal participant in the structure.²¹

In BP, on the other hand, the PP-ACC argument with psych verbs is an unmarked accusative (cf. 3b). In order to account for this difference between EP and BP, let us return to the comparison of psych verbs in Italian and BP. In the argument structure of psych verbs, when the Experiencer is the complement of V', the external argument position is available, as we can see in the representation of Italian proposed by Belletti and Rizzi (1988) and adapted by Figueiredo Silva (2007, p. 93) in (37):



For convenience, we repeat the examples with psych verbs in Italian, EP, and BP below:

Italian

(38) Gianni teme questo.
Gianni fears this

(39) Questo preoccupa Gianni.
This worries Gianni

(40) a. Questo piace a Gianni.
This pleases P_{a(to)} Gianni.DAT
b. A Gianni piace questo.
P_{a(to)} Gianni.DAT pleases this

(Belletti and Rizzi 1988, pp. 291–92)

In (39), the Theme *questo* is generated in the complement position of V, as well as the Experiencer *Gianni*, so the external argument position remains available. In (39), the Theme *questo* moves to receive Nominative Case, the unmarked Experiencer *Gianni* remains as an internal argument. Example (40b) shows that the marked Experiencer *a-Gianni* may move to a higher position, probably in the left periphery of the sentence (in Rizzi's (1997) terms), but not to SpecIP, which is occupied by the Theme *questo*.²² In EP, the dislocation of the dative to the left periphery is possible as well, since the Experiencer carries the dative Case by being overtly marked (see 42), as we have just discussed.

(41) O filme agradou ao João. (EP)
The movie pleased P_{a(to)} João.DAT

(42) Ao João agradou o filme. (EP)
P_{a(to)} João.DAT pleased the movie

(43) O filme agradou -lhe. (EP)
The movie pleased -CL.DAT.3SG

For BP, Figueiredo Silva (2007, pp. 93–94) argues that internal arguments licensed with psych verbs have inherent accusative case associated with the Experiencer theta role—*João*

in (44) and (45). Subsequently, the Theme argument, *o filme*, ‘the movie’, in (44) and (45), may move to check its nominative case in SpecIP. Therefore, (45) is ungrammatical because the DP Theme *O João* already bears inherent ACC and cannot receive NOM in SpecIP.

(44)	O	filme	agradou	o	João.	(BP)
	The	movie	pleased	the	John.ACC	
(45)	*O	João	agradou	o	filme.	(BP)
	The	João.ACC	pleased	the	movie	
(46)	O	filme	agradou	-o/ele.		(BP)
	The	movie	pleased	-CL.ACC.3SG./ACC		

It is remarkable that the only possible way to displace the Experiencer argument in BP to the left periphery would be to double it by a personal pronoun that will fill the accusative position, as in (47)²³. It is interesting, though, that this position cannot be occupied by accusative clitics such as *o/a*, ‘him/her’; it may only be occupied by pronouns such as *ele*, ‘he’, which used to be only nominative but has undergone a change, and now it may also display accusative case, as we have seen before in (3b).²⁴

(47)	O	João,	o	filme	agradou	ele.
	The	João	the	movie	pleased	he.ACC

Furthermore, historically, BP has become quite resistant to the inversion of constituents (Galves 2020; Martins 2011), which also explains the ungrammaticality of (45). Therefore, we assume that the fixed SVO order is also a key factor on the loss of marked accusative in BP, as the internal argument receives inherent accusative case when entering the derivation in the complement position of V.

6. Conclusions

In this paper, we explored the variation and change of *a*-marked arguments in Historical Portuguese, with a specific focus on its occurrence in historical texts from the 16th to the 19th centuries. The results of the data analysis demonstrated that the cases of PP-ACCs increased in the 17th century, followed by a decrease in the 18th century. The number of PP-ACC occurrences is much less frequent than of NP-ACCs, though, which has led other authors to treat *a*-marked arguments as a marginal phenomenon in Portuguese. Some authors deal with this phenomenon using the label PP-ACC, while others adopt the concept of DOM, as this is a widespread phenomenon in other Romance languages, such as Spanish. However, in order to accurately address the differences and/or similarities of PP-ACCs and DOM in Portuguese, it would have been necessary to address features of the objects, such as animacy, definiteness, and referentiality of the *a*-marked arguments. These important characteristics will be taken into account in future work, as they were not part of the scope of this paper.

Finally, when the verb classes found in the data were examined, it became evident that the set of psych verbs showed a different path from other contexts, as the quantities even increased along the centuries (see Table 2). In EP, there are still *a*-marked arguments in psych predicates, as opposed to the zero morpheme in BP. The analysis led to the conclusion that, in these cases in EP, the accusative became the structural marked dative, which guarantees the thematic reading of the Experiencer to the internal argument of psych verbs. Meanwhile, in BP, to avoid an ambiguous interpretation, the argument displays inherent accusative case; hence, the Experiencer remains in the accusative position and the Theme rises to receive nominative, so overtly marking the Experiencer is no longer necessary.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used for the research can be found at the Tycho Brahe project website.

Conflicts of Interest: The author declares no conflict of interest.

Notes

- 1 In BP, pronouns that used to be only nominative may also occur in accusative contexts (cf. Kato 2012); we will return to this topic further in this text.
- 2 The first part of the research focuses solely on EP authors from the *Tycho Brahe Corpus* (Unicamp). In future work, we intend to analyze texts from Brazilian authors.
- 3 Cyrino and Irimia (2019, p. 187) point out that direct objects may still be overtly marked in Portuguese. Animated quantifiers may be unmarked or overtly marked, especially if dislocated (see iii): i. *Ele visitou todos*, ‘He visited everyone’; ii. *Ele visitou a todos*, ‘He visited Pa(to) everyone’; iii. *A todos, ele visitou*, ‘Pa(to) everyone, he visited’. We intend to address this context of quantifier, and others mentioned by the authors in future work.
- 4 The example was taken from the *Tycho Brahe Corpus* (Unicamp). The code V_004, 002.1819 indicates the following: V_004—text (in this case: *Sermões do Padre Vieira*—Padre Vieira Sermons); 202—page number in the original; and 1819—line number.
- 5 See Historical Corpus of Portuguese Tycho Brahe: <http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html> (accessed on 4 November 2023).
- 6 The dative clitic *lhe* is still active in some areas of Brazil, but it was recategorized as second person (cf. Figueiredo Silva 2007; Martins et al. 2019).
- 7 We are using Chomsky (1981) Government and Binding terms, because the discussion on psych verbs will be based on Belletti and Rizzi (1988), as well as Figueiredo Silva (2007), who still adopt this framework, and not on more recent versions of the generative approach, such as the Minimalism Program (Chomsky 1995).
- 8 We would like to thank an anonymous reviewer for pointing it out that both *buscar* and *querer* may take an unmarked DO, as well as +human DOs without *a*-marking—*Busco una persona paea este puesto*. As our focus is not Spanish, we intend to confirm if Bossong is right that there are verbs that always select marked objects in future work.
- 9 Additionally, during data collection, in order to use the computational tools described in the methods section, it is not possible to search for such categories, because the NPs in the corpus are not annotated according to these specific features. The methodology chosen permitted us to run a search in a corpus of 39,761 words very quickly. In future work, an analysis based on the thematic role of the complements may be indeed an interesting path to pursue.
- 10 Gibrail (2003) analyzed the texts from the *Tycho Brahe Corpus*. Ramos (1992, pp. 68–69) worked with letters and theater plays written between the 16th and the 20th centuries and newspapers from the 20th century.
- 11 See Pires (2017) for another analysis on marked DOs in Historical Portuguese. Besides animacy, the author also divides the corpus according to the specificity and definiteness of the object. Additionally, works by Schwenter (2006) take specificity and definiteness into account when comparing DOM in Spanish and BP. I intend to analyze these features on this corpus in future work. Finally, Cyrino and Irimia (2019) also address marked DOs in Modern BP.
- 12 According to Gonçalves and Raposo (2014, p. 1175), the following verbs present the same behavior as *agradar*—‘please’: *apetecer*—‘feel like’, *desagradar*—‘upset’, *importar*—‘matter’, *interessar*—‘interest’, *repugnar*—‘repel’ etc.
- 13 See note 1.
- 14 Even though Matias Aires was born in Brazil, his parents were Portuguese, and he moved to Portugal at the age of 11, where he was raised and educated. Therefore, as his contact with written Portuguese was mainly in Portugal and it seems his text would have the same characteristics as the other Portuguese writers, his work was maintained for this analysis.
- 15 The Corpus Search computational tool allows us to look for specific syntactic contexts, which is extremely convenient for syntactic diachronic studies. Query (ii), for example, guarantees the outcome of the search to be the contexts in which there is a PP-ACC (IP* idoms PP-ACC) in a structure where the NP subject precedes the verb (NP-SBJ* iprecedes VB* | TR* | HV* | ET*), hence SVO.

- i. This is the example of a query to search for PP-ACC:
 definition file: port.def
 node: IP*
 query: (IP* idoms PP-ACC)
 AND (IP* idoms VB* | TR* | HV* | ET*)
- ii. This is a query to search for PP-ACCs in SVO contexts:
 definition file: port.def
 node: IP*
 query: (IP* idoms PP-ACC)
 AND (IP* idoms NP-SBJ*)
 AND (IP* idoms VB* | TR* | HV* | ET*)
 AND (NP-SBJ* iprecedes VB* | TR* | HV* | ET*)
 AND (NP-SBJ* idoms !**)

16 It has been pointed out to us that the *social interaction* and *transfer of knowledge* groups seem to overlap. However, in this analysis, the intention was to separate verbs that are related to the idea of knowledge itself—such as *conhecer* (to know), *instruir* (to instruct), *reconhecer* (to recognize)—from social interaction verbs that demand a two-participant scenario, such as *abraçar* (to hug), *acompanhar* (to join), *atacar* (to attack), *calar* (to silence), *consultar* (to consult), *convidar* (to invite), *curar* (to cure), and *imitar* (to imitate). However, the name ‘transfer of knowledge’ may indeed not be the best choice as it is also a type of social interaction; we intend to revise all the data in future work and will take this issue into account.

17 In absolute numbers, we have the following: 200 occurrences in the 16th century, 328 in the 17th, 83 in the 18th, and 13 in the 19th.

18 In fact, if we replace ‘God’ with another noun, such as ‘John’, and use it as an *a*-marked argument with the same verb *adorar ao João*, ‘to worship John’, the term ‘John’ would appear to be some sort of deity. Therefore, in modern BP, if the speaker intends to use *adorar* in the same sense as ‘love’/‘like’, the argument will be unmarked: *Eu adoro João*, ‘I love John’. Moreover, when the complement has [-human] or [-animate] features, using the preposition *a* to label it makes the sentence very odd: *#Eu adoro ao meu gato*, ‘I love my cat’; or ungrammatical: **Eu adoro ao sorvete*, ‘I love ice cream’.

19 In this paper, the context of social interaction verbs will not be addressed. Most examples with this verb class in the corpus were a combination of the verb *servir*, ‘to serve’, and title DPs, such as *Vossa Excelência*, ‘Your Honor’, among others. In the 17th century, for instance, there are 18 occurrences with this arrangement. According to Sornicola (1997), the verb ‘to serve’ derives directly from Latin and has always been marked in Italo-Romance. Therefore, in future work, it will be interesting to investigate this particular context and DO *a*-marking. I would like to thank Luigi Andriani for calling my attention to this fact.

20 An anonymous reviewer has pointed out that the difference among the tokens from the 16th to the 19th centuries may be an issue. We agree, but as explained in the methods section, we searched for PP-ACCs to find all the occurrences in the corpus. Then, we searched for the same verbal contexts with NP-ACCs, so we did not restrict the types of occurrences. In future work, we intend to look more carefully at why there was a decrease in both PP-ACCs and NP-ACCs in the 18th and 19th centuries in the corpus and hopefully compare with data from other corpora.

21 The specific reasons for the shift from accusative to dative we leave for future work.

22 It is not in the scope of this paper to discuss where exactly in the left periphery the Experiencer argument moves to.

23 An anonymous reviewer asked if the following dislocation in EP is possible: *Ao João, lhe/a ele agradou o filme*, ‘To João, CL.DAT.3SG/to him pleased the movie’—this sentence would not be possible in EP due to *lhe* in a proclitic position, and also *a ele* is only licensed in EP with clitic doubling—*O filme agradou-lhe a ele*, with a contrastive reading meaning ‘the movie pleased *him* not *her*’, for example. With enclisis and clitic left dislocation, the topicalization is grammatical in EP: *Ao João, agradou-lhe o filme*, ‘To João, pleased- CL.DAT.3SG the movie’.

24 I would like to thank Matheus Alves for calling my attention to this example.

References

- Andriani, Luigi. 2015. Semantic and Syntactic Properties of the Prepositional Accusative in Barese. *Linguistica Atlantica* 34: 61–78.
- Baunaz, Lena, and Eric Lander. 2018. The Basics. In *Exploring Nanosyntax*. Edited by Lena Baunaz, Karen De Clercq, Liliane Haegeman and Eric Lander. Oxford: Oxford University Press.
- Bazenga, Aline, and Lorena Rodrigues. 2019. Usos do clítico *lhe* em variedades do português [Uses of clitic *lhe* in Portuguese varieties]. In *Pelos Mares da Língua Portuguesa 4*. Edited by António Ferreira, Carlos Morais, Maria Fernanda Brasete and Rosa Lúcia Coimbra. Aveiro: Editora Universidade de Aveiro.
- Belletti, Adriana, and Luigi Rizzi. 1988. Psych-verbs and Theta-theory. *Natural Language & Linguistic Theory* 6: 291–352.
- Bossong, Georg. 1991. Differential object marking in Romance and beyond. In *New Analyses in Romance Linguistics: Selected Papers from the XVIII Linguistics Symposium on Romance Linguistics*. Edited by Dieter Wanner and Douglas A. Kibbee. Amsterdam and Philadelphia: Benjamins, pp. 143–70.
- Caha, Pavel. 2009. The Nanosyntax of Case. Unpublished Ph.D. thesis, University of Tromsø, Tromsø, Norway.
- Calindro, Ana. 2015. Introduzindo Argumentos: Uma proposta para as sentenças ditransitivas do português brasileiro. [Introducing arguments: The case of ditransitives in Brazilian Portuguese]. Unpublished Ph.D. thesis, University of São Paulo, São Paulo.

- Calindro, Ana. 2016. Introducing indirect arguments: The locus of a diachronic change. *Rivista di Grammatica Generativa* 38: 35–44.
- Calindro, Ana. 2017. O acusativo preposicionado na história do português: o caso dos verbos psicológicos. *Revista da Academia Brasileira de Filologia XXI*: 33–44.
- Calindro, Ana. 2020. Ditransitive constructions: What sets Brazilian Portuguese apart from other Romance languages? In *Dative Structures in Romance and Beyond*. Edited by Anna Pineda, Jaume Mateu and R. Etxepare. Berlin: Language Science Press, pp. 73–92.
- Cançado, Márcia. 1996. Verbos Psicológicos: Análise Descritiva dos Dados do Português Brasileiro. *Revista de Estudos da Linguagem* 4: 89–114. [CrossRef]
- Cançado, Márcia. 2013. *Catálogo de Verbos do Português Brasileiro: Classificação Verbal Segundo a Decomposição de Predicados* [Brazilian Portuguese verb catalogue: Verbal classification according to predicate decomposition]. Belo Horizonte: Editora UFMG.
- Carneiro, Beatriz, and Rozana Naves. 2010. Verbos Psicológicos: Hipóteses de mapeamento da estrutura argumental e aquisição da linguagem [Psych Verbs: Mapping hypothesis of the argument structure and language acquisition]. In *Proceedings of IX Encontro do CELSUL*. Palhoça: Universidade do Sul de Santa Catarina.
- Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.
- Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge: The MIT Press.
- Corriente, Federico. 1997. *Poesía dialectal árabe y romance en Alandalús*. Madrid: Gredos.
- Cuervo, Cristina. 2003. *Datives at Large*. Unpublished Ph.D. thesis, MIT, Cambridge, MA, USA.
- Cyrino, Sônia, and Monica Irimia. 2019. Differential Object Marking in Brazilian Portuguese. *Revista Letras* 99: 177–201. [CrossRef]
- Delille, Karl Heinz. 1970. *Die geschichtliche Entwicklung des präpositional Akkusativs im Portugiesischen*. Bonn: Romanisches Seminar der Universität Bonn.
- Döhla, Hans-Jörg. 2014. Diachronic convergence and divergence in differential object marking between Spanish and Portuguese. In *Stability and Divergence in Language Contact: Factors and Mechanisms*. Edited by Kurt Braunmüller, Steffen Höder and Karoline Köhl. Amsterdam: John Benjamins, pp. 265–89.
- Fábregas, Antonio. 2013. Differential Object Marking in Spanish: State of the art. *Borealis. An International Journal of Hispanic Linguistics* 2: 1–80. [CrossRef]
- Figueiredo Silva, Maria Cristina. 2007. A perda do marcador dativo e algumas das suas consequências [The loss of the dative marker and some consequences]. In *Descrição, História e Aquisição do Português Brasileiro*. Edited by Ataliba Castilho, Torres Morais, Maria Aparecida, Ruth Lopes and Sônia Cyrino. Campinas: Pontes Editores, pp. 85–110.
- Galves, Charlotte. 2001. *Ensaio Sobre as Gramáticas do Português* [Essays on the grammars of Portuguese]. Campinas: Editora da Unica.
- Galves, Charlotte. 2007. A língua das caravelas: Periodização do português europeu e origem do português brasileiro [The language of the caravels: European Portuguese periodization]. In *Descrição, História e Aquisição do Português Brasileiro*. Edited by Ataliba Castilho, Torres Morais, Maria Aparecida, Ruth Lopes and Sônia Cyrino. Campinas: Pontes Editores, pp. 513–28.
- Galves, Charlotte. 2020. Mudança sintática no português brasileiro [Syntactic change in Brazilian Portuguese]. *Cuadernos de la Alfal* 12: 17–43.
- Galves, Charlotte, Aroldo Andrade, and Pablo Faria. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. Available online: <https://www.tycho.iel.unicamp.br/home> (accessed on 8 August 2020).
- Gerards, David. 2020. *Differential Object Marking in Romance Languages*. Leipzig: University of Leipzig, Ms.
- Gibrail, Alba. 2003. O Acusativo Preposicionado do Português Clássico: Uma Abordagem Diacrônica e Teórica. [The prepositional accusative in Classical Portuguese: A diachronic and theoretical approach]. Master's thesis, University of Campinas, Campinas, Brazil.
- Gonçalves, Anabela, and Eduardo Raposo. 2014. Verbo e sintagma verbal [Verb and verbal phrase]. In *Gramática do Português*. [Portuguese grammar]. Edited by Eduardo Raposo, Maria Fernanda Nascimento, Maria Antónia Mota, Luísa Segura and Amália Mendes. Lisboa: Fundação Calouste Gulbenkian, vol. II, cap. 28. pp. 1155–218.
- Hardarson, Gísli. 2016. A case for a Weak Case contiguity hypothesis: A reply to Caha. *Natural Language and Linguistic Theory* 34: 1329–43. [CrossRef]
- Kato, Mary. 2012. Caso inerente, caso default e ausência de preposições [Inherent case, default case and the absence of prepositions]. In *Por amor à Linguística*, 1st ed. Edited by Adeilson Sedrins. Maceió: Edufal, pp. 83–99.
- Kato, Mary, Sônia Cyrino, and Vilma Corrêa. 2009. Brazilian Portuguese and the recovery of lost clitics through schooling. In *Minimalist Inquiries into Child and Adult Language Acquisition*. Edited by Acrisio Pires and Jason Rothman. Jason: Mouton de Gruyter, pp. 245–72.
- López, Luis. 2012. *Indefinite Objects: Scrambling, Choice Functions, and Differential Marking*. Cambridge: MIT Press.
- López, Luis. 2016. (In)Definiteness, Specificity, and Differential Object Marking. In *Manual of Grammatical Interfaces in Romance*. Edited by Susann Fischer and Christoph Gabriel. Berlin and New York: de Gruyter, pp. 241–65.
- Manzini, Rita, and Ludovico Franco. 2016. Goal and DOM datives. *Natural Language & Linguistic Theory* 34: 197–240.
- Martins, Ana Maria. 2011. Clíticos na história do português à luz do teatro vicentino [Clitics in history of Portuguese in the light of Vicente's plays]. *Estudos de Linguística Galega* 3: 83–109.
- Martins, Marco Antonio, Kássia Moura, and Franklin Costa da Silva. 2019. Análise diatópico-diacrônica dos complementos pronominais de verbos na escrita brasileira dos séculos XIX e XX [Diatopic-diachronic analysis of verbs pronoun complements in Brazilian writing of the 19th and 20th centuries]. *Working Papers em Linguística* 20: 195–216. [CrossRef]

- Naves, Rozana. 2005. *Alternâncias Sintáticas: Questões e Perspectivas de Análise* [Syntactic Alternations: Issues and analysis perspectives]. Unpublished Ph.D. thesis, Universidade de Brasília, Brasília, Brazil.
- Pineda, Anna. 2012. What lies behind dative/accusative alternations in Romance. *Romance Languages and Linguistic Theory 2012: Selected Papers from 'Going Romance' 2012*: 123–40.
- Pineda, Anna. 2017. From dative to accusative: An ongoing syntactic change in Romance. Paper presented at Anglia Ruskin—Cambridge Linguistics Seminars for Easter Term, Anglia Ruskin University, University of Cambridge, Cambridge, UK, May 19.
- Pires, Aline. 2017. *A Marcação Diferencial de Objeto no Português: Um Estudo Sintático-Diacrônico*. [Differential Object Marking in Portuguese: A Syntactic Diachronic Study]. Master's thesis, University of Campinas, Campinas.
- Ramos, Jânia. 1992. *Marcação de Caso e Mudança Sintática no Português do Brasil: Uma Abordagem Gerativista e Variacionista*. [Case Assignment and Syntactic Change in Brazilian Portuguese: A Generative and Variationist Approach]. Unpublished Ph.D. thesis, University of Campinas, Campinas, Brazil.
- Rizzi, Luigi. 1997. The Fine Structure of the Left Periphery. In *Elements of Grammar: Handbook of Generative Syntax*. Edited by Liliane Haegeman. Dordrecht: Kluwer, pp. 281–337.
- Schwenter, Scott. 2006. Null objects across South America. In *Selected Proceedings of the 8th Hispanic Linguistics Symposium*. Somerville: Cascadilla Press, pp. 23–36.
- Sornicola, Rosanna. 1997. L'oggetto preposizionale in siciliano antico e in napoletano antico: Considerazioni su un problema di tipologia diacronica. *Italienische Studien* 18: 66–80.
- Svenonius, Peter. 2002. Icelandic Case and the Structure of Events. *Journal of Comparative Germanic Linguistics* 5: 197–225. [CrossRef]
- Torres Morais, Maria Aparecida. 2007. *Dativos*. [Datives]. Professorship thesis, USP, São Paulo, Brazil.
- Torres Morais, Maria Aparecida, and Heloisa Salles. 2010. Parametric change in the grammatical encoding of indirect objects in Brazilian Portuguese. *Probus* 22: 181–209.
- Torres Morais, Maria Aparecida, and Rosane Berlinck. 2018. O objeto indireto do português: Argumentos aplicados e preposicionados [Portuguese indirect object: Applied arguments and prepositional]. In *História do Português Brasileiro: Mudança Sintática do Português Brasileiro: Perspectiva Gerativista*. Coord. by Sonia Cyrino and Maria A. Torres Morais. São Paulo: Contexto.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Article

Word Order in Colonial Brazilian Portuguese: Initial Findings

Aroldo Leal de Andrade ^{1,*} and Lara da Silva Cardoso ²

¹ Faculdade de Letras, Federal University of Minas Gerais (UFMG), Belo Horizonte 31270-901, MG, Brazil

² Departamento de Ciência Humanas e Letras, State University of Southwest Bahia (UESB), Jequié 45205-490, BA, Brazil; lara.silva@uesb.edu.br

* Correspondence: aroldoleal@letras.ufmg.br

Abstract: Some recent studies have posed the hypothesis according to which the grammatical stage that precedes the cultured trend of Brazilian Portuguese is Colonial Brazilian Portuguese, not Classical Portuguese. Therefore, there are still few works systematically comparing these two varieties. This is the goal of the present paper, which focuses on word order. By undertaking a corpus-based study using the same textual genre from the contemporary authors Eusébio de Matos and António Vieira, we have looked for all word order patterns while paying special attention to the X*VS order (with one or more constituents preceding the verb and a postverbal subject), given that it is quite typical of Classical Portuguese as a V2-like grammar, unlike the modern Portuguese grammars. We have observed that, although the colonial text follows the general trends of the classical language, it starts to depart from a V2-like grammar because it shows a higher frequency of non-V2 orders and a preference for informationally marked constructions involving internal positions to the clause. From a parameter hierarchy viewpoint, the main conclusion is that such differences represent frequency divergences which are consistent with nano- or microparametric changes which took place later.

Keywords: Colonial Brazilian Portuguese; Classical Portuguese; word order; parameter hierarchies; informationally marked constructions

1. Introduction

With the development of research on the history of Brazilian Portuguese and the availability of new data sources, not only written by Portuguese citizens, but also by natives, some investigations are testing the possibility that the differentiation between Brazilian Portuguese and European Portuguese (respectively referred from now on as BP and EP to allude to their modern varieties) started earlier than previously expected. This hypothesis considers that even Colonial Brazilian Portuguese (from now on, ColBP)—the term we use to refer to the variety spoken by the descendants of the Portuguese colonizers—showed some spectrum of variation, which should have been even stronger in other population segments, supposedly due to more robust language contact.

The present work therefore puts into effect a research agenda on ColBP and applies it to a specific case study as follows: the description and analysis of word order in a text written by a Brazilian author from the 17th century, with special attention to the X*VS pattern, in which there are one or more preverbal constituents, the verb, and the subject. This choice is significant for the study of grammars displaying V-to-C movement (V2 grammars) because they show more frequent occurrences of X*VS cases, together with a more varied distribution of informational-structural values related to preverbal constituents.¹ Our main research question is whether there are differences regarding the frequency of constructions involving a preverbal constituent between the grammar of ColBP and the grammar of Classical Portuguese (from now on, CIP).

To achieve this end, in this first stage, we intend to compare data with fronted constituents in main clauses from the following two sermons: *Ecce Homo*,² written by Eusébio de Matos, a printed original document of 25,090 words edited by Ilma Alkimim (cf.

Citation: de Andrade, Aroldo Leal, and Lara da Silva Cardoso. 2024.

Word Order in Colonial Brazilian Portuguese: Initial Findings.

Languages 9, 269. <https://doi.org/10.3390/languages9080269>

Academic Editor: Patrícia Amaral

Received: 7 April 2024

Revised: 14 July 2024

Accepted: 26 July 2024

Published: 1 August 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Alkimim 2014) and the *Sermons*³ by Father António Vieira, a printed original document of 53,855 words edited by Father Gonçalo Alves (cf. Vieira 1907).

The reason for these choices is, in summary, due to the high degree of similarity between these two texts, which represent the same textual genre, and their authors, who had a similar background. Our interpretation is that these multiple similarities are welcome because they suggest that any relevant differences between word order in these texts should be related to the inception of new grammars in variation, especially if they are consistent.

The theoretical framework adopted for this research is generative diachronic syntax, according to which syntactic changes are triggered by mismatches in parameter setting in the human mind, arising from linguistic experience (Lightfoot 1999). Two complementary theories are adopted as follows: the concept of parameter hierarchies, which stems from the notion of syntactic microparameters (Roberts 2019)⁴, and the interface between Information Structure and Syntax mapped into projections at the left periphery of the clause (cf. Rizzi 1997).

This text is organized as follows. Section 2 discusses the main features of ColBP and its relevance for a more detailed understanding of the history of BP and beyond. Section 3 presents background information on word order patterns in CIP. Section 4 details the materials and methods followed in our corpus-based investigation. Section 5 presents the comparative results, both at a general level—involving a comparison of the various word order patterns in main, complement and adjunct clauses—and at a more specific level, regarding the possibilities involved in X*VS order. In Section 6, we include some qualitative observations related to the results. To conclude, Section 7 presents the final remarks.

2. Colonial Brazilian Portuguese: Characterization and Relevance

The divergence between the main modern varieties of Portuguese (BP and EP) is widely recognized and raises questions about the moment of its inception, its motivation, and development. The predominant hypothesis indicates that BP emerged during the Portuguese colonial period (1532–1822), specifically in the mid-18th century (Tarallo 1993; Mattos e Silva 2004; Galves 2007). This assumption was based on sociohistorical factors known to potentially accelerate linguistic changes, such as intense multilingualism, a high degree of miscegenation, and low access to formal education.

The intense linguistic contacts that characterized the colonial period represent a primordial factor for the hypothesis that in Colonial Brazil, especially during the 18th century, an authentic grammar was being developed. During Portuguese colonization in Brazilian territory, the settlers had to deal with a complex linguistic situation: numerous indigenous languages, which were present in the colony since the pre-colonial period; various African languages taken to its territory with enslavement; as well as some regional pidginized languages brought about by the contact with Portuguese. The various sociolinguistic scenarios were surely facilitated by the fact that, although the Portuguese immigrants comprised a demographic minority, they held a position of social supremacy, which allowed them to maintain frequent interactions with African (and, in many cases, also with indigenous) peoples, resulting in a mixed population with several social layers.⁵

These circumstances lasted until the beginning of the 18th century and represented a disadvantage for the use of the Portuguese language, which was often supplanted in favor of more ‘accessible’ languages aimed at communicating and catechizing non-European peoples, known as Indigenous General languages. Being widely adopted until the mid-1750s, inclusively by the Portuguese colonizers and their descendants, these creolized varieties of Tupian languages fell into demise because of a series of coordinated measures involved in linguistic imposition, which favored the Portuguese language.

Due to this process, scholars such as Mattos e Silva (2004) and Galves (2007) consider the 18th century as a crucial period for the emergence of BP grammar; with the adoption of Portuguese as a language of mandatory and universal use in Brazil, its acquisition occurred through an accelerated and imperfect process by indigenous peoples and by Africans living in Brazil and their descendants, as well as through a natural process by white individuals

born in Brazil, culminating in the configuration of a grammatical system quite different from Portuguese grammar, which served as the norm during the colonization of Brazil.

This sociohistorical reality is aligned with grammatical analyses carried out based on texts written during the colonial period in Brazil. Tarallo (1993), who investigated a range of phenomena (the pronominal system in relation to subject and object positions, relativization strategies, and word order in both declarative sentences and direct and indirect questions) highlighted the following: “O português do Brasil existe como língua literária somente a partir dos anos 1700. Qualquer material anterior àquela data revelaria, pois, traços do português europeu e enviesaria os dados” [Brazilian Portuguese has existed as a literary language only since the 1700s. Any material prior to that date would therefore reveal relics of European Portuguese and would produce data bias] (Tarallo 1993, p. 65). Therefore, there are grounds to believe that, during the 18th century, two grammars were already interacting in Brazil: CIP, ‘transferred’ by the mouths of the colonizers since the 1500s, and a preliminary version of BP.⁶ Retracting the identification of Brazilian Portuguese prior to 1700 has not been deeply explored from the grammatical point of view until now but is compatible with previous studies. Barbosa (1999, p. 14), one of the first descriptions of the Portuguese language written in Brazil during the colonial period, already recognizes the necessity of distinguishing what he calls ‘Portuguese in Brazil’ from ‘Portuguese of Brazil’. In the same line, Ramos and Oliveira (2021, p. 19) state that the Portuguese colonizers fostered the creation of a new dialect in Brazil already at their third generation on the new soil.

With regard to word order in declarative sentences, Tarallo (1993) highlights that the decrease in the use of the verb-subject order is more striking between the end of the 19th century and the beginning of the 20th century. His data, based on the original results in Berlinck (1988), are presented in Table 1.

Table 1. Frequency of use of verb-subject (VS) order in the history of Brazilian Portuguese.

Period	Percent	Occurrences: VS/Total
18th century (1750)	42%	203/486
19th century (1850)	31%	144/469
20th century (1987)	21%	63/1262

Source: Tarallo (1993, p. 71), translated with adaptations.

Before the 19th century, according to the author, there was only an “imitation” of the grammatical pattern of the Portuguese language in Portugal, in which the Brazilian data did not show any grammatical innovations. Other studies have addressed the 18th century in their corpora, but there has not yet been enough data to definitively identify the point of emergence of BP (Pagotto 1992; Coróa 2022).

From this perspective, the main approach to understanding the process of grammatical change in the development of BP has been a comparative analysis that contrasts the Portuguese language from Portugal in the 17th and 18th centuries with Brazilian texts written from the 18th century onwards. The colonial period is seen, therefore, as a transition between two grammars, whose linguistic research is limited to a specific time frame (18th century) to identify syntactic patterns that are close to standard BP. According to this traditional view, the period during which CIP and BP were used partially overlaps, as shown in Figure 1, adapted from Mattos e Silva (2004) and Galves (2007).

Figure 1 represents the traditional hypothesis for the emergence of BP, which already points to the relevance of the colonial period in this process. The vertical lines indicate the probable period of emergence of Portuguese grammars in Colonial Brazil. The horizontal lines indicate the probable maintenance and use of two grammars in the Brazilian territory: CIP and BP. Pagotto (1992), Carneiro (2005), Martins (2009), and Carneiro and Galves (2010), who analyzed syntactic aspects related to the order of clitic pronouns in Brazilian 19th-century texts, propose that CIP remained in use in the Brazilian territory until the mid-19th century. Nevertheless, it has mingled with BP since the 18th century according to

Tarallo (1993), Galves (2007) and Mattos e Silva (2004). According to this scenario, there is no need to assume a different grammar in Colonial Brazil.

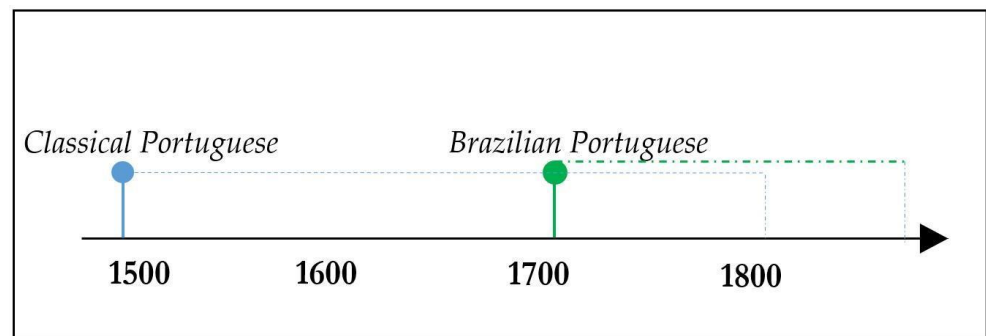


Figure 1. The emergence of Portuguese grammars in Colonial Brazil (first pass).

However, we are taking another stance regarding the study of Portuguese in the colonial period. Following Andrade (2020), Cardoso (2020), and Cardoso et al. (2023), we highlight the relevance of deepening the syntactic investigation into the Portuguese language in Colonial Brazil since the 17th century, based on a comparative analysis between texts written by Portuguese authors and by individuals born in Brazil during this period. This methodological position is based on the hypothesis that a grammatical variety existed between CIP and BP, which we call ColBP. This variety would have been acquired by the descendants of Portuguese citizens born in Brazil in the 17th century. Figure 2 depicts this hypothesis, in accordance with Cardoso et al. (2023).

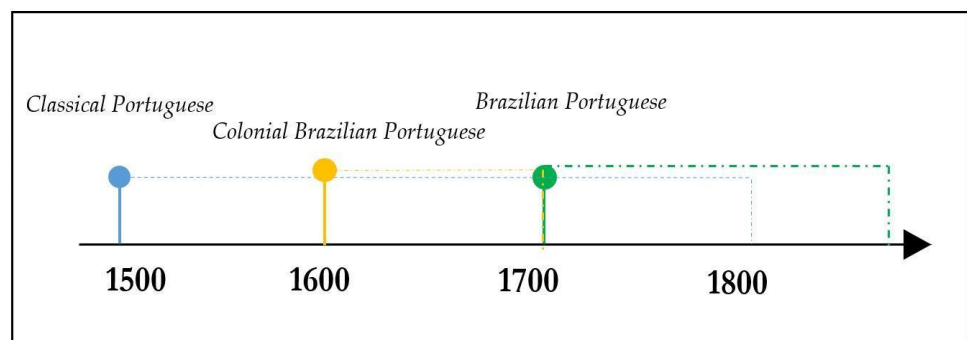


Figure 2. The emergence of Portuguese grammars in Colonial Brazil (final pass).

The sociohistorical and linguistic evidence that supports such a perspective has been presented in Cardoso et al. (2023). A key aspect of the sociohistorical evidence in favor of the existence of ColBP consists in the assumption that linguistic contact had been present in Brazil since earlier periods, although to a lesser degree if compared to the part of the population acquiring Portuguese as a second language. This is the null hypothesis, as there had already been contact between indigenous, European, and African peoples and their descendants since the beginning of the colonization process, also with consequences among the European descendants. Historians who have investigated Brazil’s colonial period have highlighted the noteworthy contact between children of Portuguese and African people during childhood (Mattoso [1979] 2003), a period in which language acquisition occurs (Lightfoot 1999). Therefore, we did not identify a similar linguistic experience to that of Portugal that would support the assumption that CIP was the only variety grammatically acquired and spoken by white individuals during the first 100 years of Portuguese colonization, devoid of any process of change in its grammatical components, as assumed in the hypotheses formulated in Mattos e Silva (2004) and Galves (2007).

Therefore, we recognize the existence of ColBP as a linguistic variant of Portuguese acquired by the descendants of the Portuguese born in Brazil from the 17th century onwards, emerging after the introduction of CIP in Brazil and preceding the formation of ‘cultured’ BP (i.e., the variant of BP spoken by the most prestigious class of the population).⁷ However, although ColBP presents some grammatical differences in relation to the CIP grammar written in the 17th century, we consider that these changes in the grammatical system were less significant if compared to the changes observed in BP. This is mainly due to the general presence of regular linguistic transmission among the likely speakers of this grammatical variety (cf. Lucchesi 1994).

Due to the recent introduction of this proposal, there is yet little linguistic evidence favoring the existence of ColBP. Nevertheless, a comparative study on the placement of clitic pronouns, both in texts written by Portuguese descendants born in Brazil between the 17th and 18th centuries and in texts written by Portuguese citizens living in Brazil during the same time period (Cardoso 2020) has already revealed an important difference: the former data display more frequent proclisis in Variation Context 2, whereas the latter group showed more frequent enclisis, as shown in Figure 3.⁸

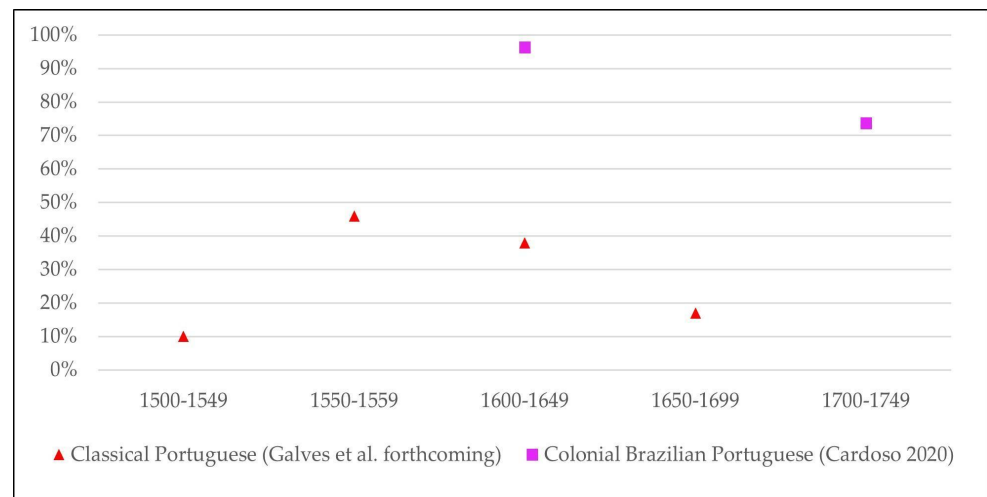


Figure 3. The frequency of proclisis in Variation Context 2 in texts by Portuguese writers compared to Brazilian writers (17th–18th century), considering the authors’ date of birth (adapted from Cardoso 2020, p. 148, Galves et al. Forthcoming).

The difference illustrated in Figure 3 is not taken as innovative because proclisis is also possible in the CIP grammatical system, although it is not the predominant form in the relevant context. Nevertheless, the prevalence of proclisis in ColBP texts stands out in relation to the tendency towards enclisis in CIP texts, highlighting a distinctive feature of these colonial texts that may have led to a grammatical change at a later date. Interestingly, these findings have been confirmed by Corôa (2022), who examined a different set of texts written by Portuguese descendants born in Brazil in the 17th century. A marked preference for a proclitic placement was observed in sentences, including a main verb preceded by a subordinate clause.

The evidence highlighted above, which relate sociohistorical and grammatical aspects, requires the pursuit of the following new research directions that depart from the mainstream practices followed to date:

1. Expanding the scope of investigation into the history of the Portuguese language in Brazil to include 17th-century texts;
2. Carrying out comparative research between CIP and ColBP, without necessarily assimilating the latter to linguistic phenomena associated with BP.

The first measure involves the investigation of the grammar present in texts written by Portuguese descendants born in Brazil in the 17th century. Unlike the previous proposal

on the history of BP (Tarallo 1993), we assume that these materials will not just reveal an “imitation” of the grammar brought by the Portuguese in the previous century. Instead, the study of 17th-century texts shall contribute to the identification of ColBP and the investigation of possible signs of the emergence of a new grammatical variety. The importance of this aspect, however, brings about other challenges, such as the rarity of documents written during this period and the difficulty in identifying their authors. These tasks demand interdisciplinary efforts that involve cooperation among experts in the areas of philology and paleography (Lose 2022).

The second measure involves comparative research between CIP and ColBP, with the aim of identifying the characteristics of the latter and valuing its particularities. Searching for features of BP in the data may limit our ability to discern the grammar of ColBP, considering that we treat these as different varieties. Therefore, we propose a comparative analysis focusing on grammatical phenomena known for their diachronic instability, susceptibility to idiosyncratic patterns, and parametric changes over short periods of time, so that ColBP may be studied as a variety to be described in its own right.

These phenomena tend to correspond to lower-scale parameters (micro- and nanoparameters), as defined by Roberts (2019), which may be illustrated by various grammatical rules, such as clitic placement and the expression of null subjects.⁹ Alternatively, only the frequency of some grammatical options may be affected, a process that clears the way for a parametric change to take place much later. The goal of this comparison is to identify linguistic patterns that, even on a small scale, diverge from CIP, thus providing arguments from internal history that help to identify the nature of ColBP.

Although we privilege a language-internal approach for linguistic change, we do not overlook external motivations that may play a role in the development of ColBP. In fact, it is probable that contact was the main source of innovations in this variety, even though it was much less widespread when compared to General Brazilian Portuguese. Even so, there are elements that suggest ColBP was also related to language shifts, as most of its speakers were not of pure European origin, i.e., they also expressed some degree of miscegenation, which could lead to interference, in which the aspects most susceptible to change are the lexicon and pragmatics. From the side of the Portuguese colonizers, we posit the existence of foreigner talk, typical of situations in which native speakers see learners of the dominant language as socially inferior. Both trends point to a process of language simplification, which may include, in the case of foreigner talk, the avoidance of contractions, shorter and less complex sentences, the retention of optional constituents, and the explicit marking of semantic relations, among other consequences (cf. Matras 2009).

3. Word Order Patterns in Classical Portuguese

The description of word order in CIP has been the origin of much debate regarding the parameter regulating verb movement. Some authors pursue the hypothesis according to which the verb was placed in I since Old Portuguese (cf. Martins 2019). Others have put forward that there was a change in verb position in the clause, from a larger movement (until C) to a narrower movement (until I) (cf. Galves and Kroch 2016). Here, we follow the position that, to some degree, the two proposals can describe most problems, but the latter more clearly identifies the change between older and modern stages of Portuguese. This being so, let us observe the main characteristics of CIP grammar regarding word order.

According to Galves (2020), who adopts Wolfe’s (2015) notion of ‘relaxed V2 grammar’, in CIP, “linear V2 is the dominant order, the pre-verbal field is not specialized for subjects, and subjects very frequently appear in post-verbal position” (p. 384). Unlike in strict V2 languages, V1 is a frequent order and V3 (and even V4) may also occur. According to Galves’ corpus-based research on CIP, there were 59% of V2 cases, among which 27% represent SV occurrences and 73% XV occurrences, in main declarative clauses with transitive verbs. This is consistent with a V2 type of grammar, in which not only is V2 frequent, but there is also a great predominance of non-SV sentences among the cases of V2.

Among the cases of V1, the following two different situations must be considered: absolute V1 and V1 preceded by a conjunction. The first situation may be derived by the presence of a null subject (i.e., rhematic or presentational V1) or by the presence of a postverbal subject, a situation in which there is topic continuity marked by a postverbal subject. Therefore, V1 is seen as totally compatible with a V2 type of grammar. Comparing V2 with V1, the latter represents 32% of the occurrences.

The occurrences of V > 2 are represented by various types of preverbal constituents as follows: fronted or left-dislocated objects, adjuncts represented by simple phrases or by dependent clauses (especially adverbial clauses), subjects and intercalated phrases (parentheticals). To explain their derivation, Galves (2020) follows Benincà's (1995) proposal, according to which only fronted constituents count for V2. Nevertheless, higher positions in the phrase marker may be expressed by base-generated constituents. The proposed representation for the left periphery, considering a cartographic approach for syntax, is shown in (1) (from Galves 2020, p. 391):

$$(1) \quad [_{\text{ForceP}} \quad [_{\text{TopP}^*} (\text{Topic}) \quad || \quad [_{\text{FocP}} (\text{Focus}) \quad [_{\text{KontrP}} (\text{Kontr}) \quad [_{\text{FinP}} \quad [_{\text{TP}}]]]]]]$$

In this structure, the doubled vertical line separates the external domain, in which (left-dislocated or strongly contrastive) topics are base-generated from the internal domain, in which quantified foci and contrastive constituents occur.

Other questions relate to the differences between CIP and the other stages of the history of Portuguese in Europe. Galves and Kroch (2016) present some of the most important parameters distinguishing these stages, which may be summarized as follows:

- OldP > CIP: Multiple AgrS specifiers (yes > no)
- CIP > EP: C attracts V (yes > no)

Accompanying the latter shift, they also assume a change in the domain governing the restriction against clitic-first (changing from a prosodic to a morphosyntactic restriction). Results from recent works suggest, however, that word order changes between OldP and CIP are still to be further explored and may be better explained as microparameters operating at the left periphery (cf. Andrade and Galves 2019, p. 22, fn. 26), more specifically, those related to stylistic fronting (understood as *v*P-movement to Spec, FinP) and to clitic movement to Force (where the complementizer is positioned).

4. Materials and Methods

4.1. Materials

To carry out the research agenda described in Section 2, rigorous observational control and a detailed analysis are necessary to identify peculiar grammatical patterns between the two grammars under analysis. However, there are some challenges to face, primarily the scarcity of texts that meet our minimal prerequisites. Considering the social aspects of the time, such as limited access to schooling, few individuals born in the 17th century had the opportunity to learn to read and write. Furthermore, the time factor presents itself as a problem for studies of past synchronies as not every text was considered relevant to the point of being preserved for centuries, a fact that restricts the material available for analysis. This is reflected by the fact that the archives containing documents from the colonial period, for the most part, contain a set of manuscripts of administrative and/or literary content, preserved according to the relevant criteria of the time in which these materials were produced.¹⁰

The difficulty involved in the selection of the sources lies not only in the localization of the manuscripts, but is also related to the restricted types of the textual genres preserved.¹¹ According to Kabatek (2008), issues such as the purpose of communication and the linguistic-discursive formulas present in certain genres can influence the grammar displayed in the written text, making it confusing to identify the grammar of the individual who wrote it and what corresponds to the writing traditions of the period. To alleviate such problems, we selected for this comparative research two manuscripts of the same textual

genre, produced at the same time by contemporary individuals. Relevant information regarding the selected materials is summarized in Table 2.

Table 2. Corpus selected for the linguistic analysis.

	Colonial Brazilian Portuguese	Classical Portuguese
Author	Eusébio de Matos (Brazil, 1629–1692)	António Vieira (Portugal, 1608–1697)
Textual Genre	Sermon (<i>Ecce Homo</i>)	Sermon (<i>Sermões</i>)
Edition	Conservative edition by I. Alkimim	Conservative edition by G. Alves
Number of words	25,090 words	53,855 words
Source	CE-DOHS	TBC

For the analysis of ColBP word order, we chose to analyze *Ecce Homo*, a collection of sermons delivered at the Jesuit College of Salvador da Bahia during the 17th century, whose first printed version was published in 1677 in Lisbon. Its digital version was made available on the website of *Biblioteca Brasileira Guita and José Mindlin*.¹² However, for this study, we chose the philological edition produced by Alkimim (2014), who used the first edition of the text as the basis to develop a critical edition. Although the author added critical and clarifying comments to the text, considering its context of production and reception, the edition was considered conservative as there were no orthographic, syntactic, or punctuation updates involving distinctive linguistic value (Alkimim 2014, p. 44).¹³ This material is part of a set of manuscripts selected to compose the colonial corpus of the *Electronic Corpus of Historical Documents of the Hinterland* (CE-DOHS, cf. Carneiro and Lacerda 2024).¹⁴

As a representative of 17th-century CIP, we have selected the sermons written by Father António Vieira. Although the first edition of this text was published in Lisbon between 1679 and 1695, for this work, we used the modernized edition published in the year 1907. This edition preserves aspects related to the punctuation and morphosyntax of the original text, and is made available in a parsed version at the *Tycho Brahe Corpus of Historical Portuguese* (TBC, cf. Galves et al. 2017).¹⁵

We observe that the materials chosen for our research corpus present notable similarities. The two texts, published in Lisbon in a similar period, i.e., between the 1770s and 1790s, have significant literary content and belong to the same textual genre. The selection of this corpus was not random; given the number of similarities, it becomes more difficult to attribute grammatical divergences to extralinguistic factors, such as the place and year of publication, or to the influence of discourse traditions used in that period.

Additionally, there are similarities between the authors, contributing to the need for rigorous control for our investigation. As highlighted by Alkimim (2014), both authors maintained a social relationship and were aware of each other’s work. Eusébio de Matos, born in Bahia in 1629, was the brother of the renowned poet Gregório de Matos and studied at the Jesuit College (Colégio da Bahia), where he later served as Professor of Theology and Philosophy. Father Antonio Vieira was born in Lisbon and arrived in Bahia at the age of six, but returned to Portugal when Eusébio de Matos was 11 years old, and did not return to Brazil until the 1680s. Although they did not live together, their work was mutually recognized, with Vieira making some public comments about the Brazilian preacher. According to Machado (1741), Vieira said the following about Matos: “Deus se apostara em o fazer em tudo grande, e não fora mais por não querer” [God had committed to making him great in everything, and he did not become even greater because he did not want to] (p. 766).

A third aspect regarding the similarities between the two texts consists in the authors’ literary style. According to Alkimim (2014), literary critics proposed that Eusébio de Matos was an “imitator” of Vieira, since his sermons carry a theme and style that mirror those of

the famous Portuguese sermonist. A relevant example consists in the mention of the speed of the sun, which is found in both texts, as demonstrated below (examples from Alkimim 2014, p. 294 and Vieira 1907, p. 25, respectively):

- (2) a. Ora notem: o curso do Sol material
 now notice the course of the Sun material
 [...] corre em cada hora
 runs in each hour
 trezentas e oitentamil légoas (ColBP)
 three.hundred and eighty thousand leagues
 ‘Now notice: the course of the material Sun [...] runs three hundred and eighty thousand leagues every hour.’
- b. O Sol [...] corre a cada hora
 the Sun [...] runs in each hour
 trezentas e oitenta mil légoas (CIP)
 three.hundred and eighty thousand leagues
 ‘The Sun [...] runs three hundred and eighty thousand leagues every hour.’

Not only were the themes similar, but also certain stylistic structures present in Eusébio’s texts resemble Vieira’s style. Among these, Alkimim (2014, p. 100) highlights the use of the expression *que muito*, which means “why is it a wonder” or “how strange it is that”. This expression appears in both Eusébio’s and Vieira’s texts, with a greater frequency than with other literary authors of the same period (examples from Alkimim 2014, p. 100 and Vieira 1907, p. 27, respectively):

- (3) a. que muito que diga eu, que
 how much that say I that
 com aquela Capa está Cristo cubrindo
 with that cloak is Christ covering
 nossas culpas. (ColBP)
 our faults
 ‘why should it be a wonder that I say Christ is covering our faults with that cloak?’
- b. que muito é que digam e informem
 how much is that say and inform.3PL
 [...] que lhes sobejam merecimentos
 that to.them remain merits (CIP)
 ‘why should it be any wonder that they say and testify [...] that their merits reach to the rooftops?’

Therefore, the only fundamental difference between the texts lies in the fact that they were written by individuals born in different places, who acquired different grammars, with other points of divergence mitigated due to the similarities highlighted above.

4.2. Methods

Since the selected texts are available in different formats, data retrieval involved different steps as follows: for CIP, we could run automated queries using the *CorpusSearch 2* software (Randall et al. 2004), and for ColBP, we carried out manual searches for the desired word order combinations.

We studied word order in two stages. The first was more general and encompassed an investigation into the behavior of subject expression (null subjects, subject-verb order, and verb-subject order) and verb position (V1, V2, and V > 2) in three types of clauses: main clauses, complement clauses, and adjunct clauses. In the second stage, we selected only main clauses with X*VS order (i.e., clauses with a postverbal subject with one or more preverbal constituents) and classified them according to the following syntactic and informational aspects:

1. clause type: main clause, complement clause, or adjunct clause;
2. syntagmatic type of the fronted constituent: nominal phrase; prepositional phrase; adverbial phrase;
3. syntactic function of the fronted constituent: subject, accusative complement, dative complement, oblique complement, adjunct, nominal predicate, dislocated topic, or parenthetical;
4. discursive or informational function: discourse marker, vocative, complementizer, focalizing particle, V2 topic, contrastive topic, frame-setting topic, contrastive focus, or informational focus.¹⁶

The results underwent statistical analysis to verify their significance. Either a chi-square test or a G-test was applied for each table (with partitionings if required).¹⁷ Significant *p*-values are marked in bold. Residual analyses have been applied for significant tables, with the respective results shown in the footnotes.

In the next section, we present the quantitative results organized according to these descriptive levels and, in the subsequent section, we discuss the qualitative aspects observed from the data.

5. Results

5.1. General Word Order Patterns

In this subsection, we present the general results for word order in the three clause types studied. The results for main clauses in the two varieties are shown in Table 3.

Table 3. Comparison of word order patterns in the main clauses of the corpus.

Variety	Verb Position/ Subject Expression	Verb-Subject	Subject-Verb	Null Subject
Colonial Brazilian Portuguese	V1	71 (49%)	0 (0%)	127 (53%)
	V2	58 (40%)	68 (61%)	92 (38%)
	V > 2	17 (11%)	43 (39%)	21 (9%)
	Total	146 (100%)	111 (100%)	240 (100%)
Classical Portuguese	V1	97 (42%)	0 (0%)	74 (49%)
	V2	122 (53%)	86 (61%)	64 (43%)
	V > 2	12 (5%)	56 (39%)	12 (8%)
	Total	231 (100%)	142 (100%)	150 (100%)
<i>p</i> -value		0.0015 *	0.9965 †	0.6960 †

* Chi-square test for independent samples. † G-test for independent samples.

From the results in Table 3, we observe that there are some differences regarding the preferences between V1, V2, and V > 2 positions, which are significant only in the verb-subject order.¹⁸ Unlike in CIP, ColBP shows a comparatively higher percentage of V1 with null subjects and more cases of V1 than of V2 in clauses with verb-subject order:

- (4) a. Temos o exemplo entre mãos
 have.1PL the example between hands (V1; null S)
 ‘We have the example between (our) hands’
- b. Quiseram os antigos pintar a justiça
 wanted the ancient.ones paint the justice (V1; VS)
 mais rigorosa
 more strict
 ‘The ancient ones wanted to represent Justice as (being) more strict’

In CIP, V2 sentences are more frequent, equally with a preverbal or with a postverbal subject:

- (5) a. [NP-S as outras prophecias] cumprem-se
the other prophecies fulfill.3PL=INH
a seu tempo (CIP, SV)
in their time
‘The other prophecies fulfill themselves in their (due) time’
- b. [PP Na parábola das dez Virgens], fallava
in.the parable of.the ten virgins spoke.IMPERF
[NP-S Christo Senhor nosso], propria e
Christ Lord our proper and
literalmente do dia do Juízo (CIP, VS)
literally of.the day of.the Judgment
‘In the Parable of the Ten Virgins, Christ our Lord spoke, properly and literally,
about the day of Judgment’

The results for complement clauses in the two varieties are shown in Table 4.

Table 4. Comparison of word order patterns in the complement clauses of the corpus.

Variety	Verb Position/ Subject Expression	Verb-Subject	Subject-Verb	Null Subject
Colonial Brazilian Portuguese	V1	14 (63%)	0 (0%)	42 (65%)
	V2	7 (32%)	20 (74%)	19 (29%)
	V > 2	1 (5%)	7 (26%)	4 (6%)
	Total	22 (100%)	27 (100%)	65 (100%)
Classical Portuguese	V1	6 (38%)	0 (0%)	31 (80%)
	V2	10 (62%)	9 (75%)	5 (12%)
	V > 2	0 (0%)	3 (25%)	3 (8%)
	Total	16 (100%)	12 (100%)	39 (100%)
<i>p</i> -value		0.2238 †	0.9931 †	0.1532 †

† G-test for independent samples.

Table 4 reveals that the frequency of V1 and V2 are similar between the two varieties, except with postverbal subjects; ColBP shows V1 more frequently than V2 in this context. However, this result, as well as the one found in the other columns, is not significant.

The results for adjunct clauses in the two varieties are shown in Table 5.

Table 5. Comparison of word order patterns in the adjunct clauses of the corpus.

Variety	Verb Position/ Subject Expression	Verb-Subject	Subject-Verb	Null Subject
Colonial Brazilian Portuguese	V1	20 (49%)	0 (0%)	91 (58%)
	V2	17 (41%)	56 (79%)	62 (39%)
	V > 2	4 (10%)	15 (21%)	5 (3%)
	Total	41 (100%)	71 (100%)	158 (100%)
Classical Portuguese	V1	8 (40%)	0 (0%)	8 (42%)
	V2	11 (55%)	1 (20%)	4 (21%)
	V > 2	1 (5%)	4 (80%)	7 (37%)
	Total	20 (100%)	5 (100%)	19 (100%)
<i>p</i> -value		0.5839 †	0.5982 †	0.0002 †

† G-test for independent samples.

Table 5 shows that there are similarities between adjunct clauses and other clause types previously discussed. Nevertheless, the only significant column was that corresponding to null subject clauses.¹⁹ In ColBP, there is a preference for less complex structures, with the

hierarchy V1 > V2 > V > 2, as the examples below suggest (all showing a null subject and a *porque* ‘because’ connector):

- (6) a. *porque* *lhe* *acomodam* *bem* *as* *penas* (ColBP; V1)
 because 3SG.DAT accommodate well the feathers
 ‘because the feathers accommodate well to it [love].’
- b. *porque* [*com* *elas*] *nos* *explica*
 because with them us explains
o quanto *por* *nós* *padeceu.* (ColBP; V2)
 the much for us suffered
 ‘because with them [the wounds] he explains to us how much he suffered for us.’
- c. *porque* [*como* *o* *cair* *é* *pensão*] [...] [*em*
 because as the falling is burden in
nossa própria fraqueza] *temos* *algũa* *desculpa* (ColBP; V3)
 our own weakness have.1PL some excuse
 ‘because, as falling is a burden, we have some excuse for our own weakness.’

In CIP sentences with a null subject, there are more V > 2 structures, comparatively:

- (7) a. *porque* [*aos* *pequenos*] *concede-se-ha* *misericórdia* (CIP; V2)
 because to.the little.ones grant=PASS=FUT mercy
 ‘because mercy will be given to the little ones.’
- b. *porque* [*conforme* *cada* *um* *tem* *o* *coração*],
 because according.to each one has the heart
 [assim]prophetisa. (CIP; V3)
 so prophecies
 ‘because according to (where) each one has their heart, so they prophesy.’

5.2. Preverbal Constituents with Postverbal Subjects in Main Clauses

In this subsection, we present the detailed results for X*VS in main clauses, exploring different aspects of this topic. The most basic aspect is related to general combinations of preverbal elements, as presented in Table 6, in which the type and number of preverbal constituents is considered (O represents any type of object complement and A stands for any type of adjunct).²⁰

Table 6. Comparison of word order combinations in the X*VS clauses of the corpus.

Word Order Combinations	Colonial Brazilian Portuguese	Classical Portuguese
OV	8 (21%)	24 (15%)
AV	22 (60%)	118 (74%)
AAV	5 (14%)	18 (11%)
AOV	2 (5%)	0 (0%)
Total	37 (100%)	160 (100%)
<i>p</i> -value = 0.4904	(G-test for independent samples)	

In ColBP, there are more preverbal objects and preverbal adjuncts in isolation and more combinations of preverbal phrases (i.e., AAV and AOV) in comparison to CIP. Although the results for this table are not significant, it is still noticeable that (i) ColBP shows much fewer cases of X*VS data than CIP (37 versus 160 occurrences) and that (ii) ColBP shows two cases of AOV, whereas CIP displays none, despite the significantly greater amount of relevant data.

Table 7 summarizes the types of preverbal elements, in which a more detailed combination of syntagmatic and functional information on preverbal constituents is offered. The total numbers are different if compared to Table 6 because a constituent may iterate and because some of them were not included before, noticeably those in bold in the left corner.

Table 7. Comparison of types of preverbal constituents in the X*VS clauses of the corpus.

Type of Preverbal Constituent	Colonial Brazilian Portuguese	Classical Portuguese
DP accusative complement	10 (10%)	5 (3.6%)
DP dislocated topic	0 (0%)	13 (9.3%)
DP subject	0 (0%)	1 (0.7%)
PP dative complement	2 (2%)	5 (3.6%)
PP adjunct	36 (39%)	64 (45.7%)
AdvP or Adverbial Clause	43 (46%)	50 (35.7%)
Vocative	1 (1%)	0 (0%)
Parenthetical	0 (0%)	2 (1.4%)
Nominal predicate	2 (2%)	0 (0%)
Total	94 (100%)	140 (100%)
<i>p</i> -value = 0.0069	(G-test for independent samples)	

This table shows significant results, unlike the previous one. This suggests that the relevant factor groups are those capturing either more precise syntactic or informational-structural values.²¹ In relation to CIP, ColBP has more fronted accusative complements, and it shows no dislocated topics (either accusative or dative complements with a resumptive pronoun). This leads us to the next and last table.

Table 8 includes the types of informationally marked constructions, considering the classification in Götze et al. (2007) combined with that proposed in the literature on CIP (cf. Andrade 2015 for a review).²²

Table 8. Comparison of types of marked constructions found in the X*VS clauses of the corpus.

Informationally Marked Construction	Colonial Brazilian Portuguese	Classical Portuguese
Clitic Left Dislocation	0 (0%)	6 (5%)
Hanging-Topic Left Dislocation	0 (0%)	1 (0.5%)
Contrastive Focus	23 (29%)	11 (8.5%)
Aboutness Topic (V2 Topicalization)	7 (9%)	20 (16%)
Frame-Setting Topic	47 (59%)	74 (58%)
Contrastive Topic	2 (3%)	15 (12%)
Total	79 (100%)	127 (100%)
<i>p</i> -value = 0.0124	(G-test for independent samples)	

The results for this table were significant.²³ The differences indicate that ColBP

- does not show any cases of dislocation, unlike in CIP;
- significantly shows more cases of contrastive focus;
- shows fewer sentences with an aboutness topic or a contrastive topic.

In sum, ColBP displays a less pronounced variation in the types of marked constructions. These results may indicate a preference for the usage of clause-internal constituents at the left periphery in ColBP if we consider that they tend to express contrastive foci, whereas clause-external constituents indicate various types of topics (cf. contrastive topics), as the following examples from ColBP illustrate²⁴

- (8) a. Primeiramente [com aquelas feridas] representa
 firstly with those wounds represents
 Cristo o quanto nos ama (contr. focus)
 Christ the much us loves
 'Firstly, with those wounds, Christ represents how much he loves us'
 (interpretation: '...it is with those wounds that Christ represents how much he loves us')
- b. [A Thomé] mostrou Cristo cinco Chagas;
 since Thomas showed Christ five sores
 porém a nós cinco mil (contr. topic)
 but to us five thousand
 'To Thomas, Christ showed five sores; but to us, five thousand'

6. Discussions

The results shown above reveal, as expected, a great deal of similarity between the two texts, but there are some differences which we consider crucial for distinguishing between the two grammars.

Similarities are stronger in the domain of complement clauses, which do not reveal significant word order differences between the texts. Furthermore, the two texts display variation regarding the usage of double complementizers and of null complementizers, which are illustrated below only with ColBP sentences:

- (9) Pois estai certos, fiéis, [que] se não
 thus be.2PL certain believers that if NEG
 correspondermos de outra sorte a tão
 correspond.FUT.1PL of another luck to such
 grande amor, [que] este mesmo amor
 great love that this same love
 se há de converter em indignação
 PASS will PART turn.into in indignation
 'Thus be certain, believers, that if we do not respond in another way to such great love, this same love will turn into indignation'
- (10) tantos incêndios sentia no coração,
 so.many fires felt in.the heart
 que parece ∅ encerrava no peito
 that seems (that) held in.the chest
 novo Etna, novo Mongibello...
 new Etna new Mongibello
 'He felt so many fires in his heart, that it seems like he held a new Etna, or a new Mongibello in his chest...'

Differences occur in the main clauses of the corpus in the context involving the verb-subject order, which is exactly the combination detailed in Tables 6–8. Crucially, the V2 position is statistically more significant to distinguish between ColBP and CIP, and the former shows smaller numbers of V2 sentences. This may be theoretically interpreted as indicating a future demise of the V2 parameter in ColBP, although the data are still compatible with a V2 grammar. Another significant difference was found in the word order in adjunct clauses, and this may be interpreted in the following way: ColBP has a less complex left periphery (especially in dependent clauses), with a clear preference for V1 over the other categories.

Regarding the more specific comparisons shown in Tables 7 and 8, they confirm that ColBP prefers to fill the lower (internal) positions in the left periphery, which are occupied by various constituents because of movement operations. Higher (external) positions tend not to be accessed in this grammar, and this is even stronger because there are no cases of dislocation constructions. Notice, crucially, that the use of internal positions to the clause triggers proclisis instead of enclisis.

The results do not allow us to affirm that a parametric change is at stake yet. However, they reveal frequency differences that are compatible with changes that will take place later:

- in the order between subject and verb, because BP shows a subject-verb grammar;
- in the order between clitic and verb, because BP shows a proclitic grammar.

EP, on the other hand, still shows more space for VS occurrences and a predominantly enclitic grammar.

7. Conclusions

The selected texts have many similarities, but also several points of divergence which cannot be interpreted as a mere reflection of personal styles. The main aspects of their differences, considering the adopted theory of syntactic change, are as follows:

- unlike in CIP, in which V2 is the preferred option, the ColBP text shows a greater number of V1 order (if compared to V2 and V3) in all clause types;
- unlike in CIP, there is a smaller use of structures with an external topic, especially those embodied in left-dislocation and contrastive-topic constructions.

With changes stemming from these differences, we expect to observe, either in other ColBP texts or in later texts from native Brazilian authors, the following characteristics:

- a higher frequency of subject-verb clauses;
- a higher frequency of proclisis.

Our hypothesis is that ColBP already showed higher frequencies of subject-verb and proclisis than CIP. The first characteristic may be observed from Table 3 if the totals of subject-verb are compared to those of verb-subject in the two texts (43% in ColBP versus 38% in CIP). The second point was already observed in a previous work (cf. Figure 3 and Cardoso 2020). These observations, which are merely made in terms of frequency at this point, may have catastrophic consequences for the later grammar of BP, following the observations of Lightfoot (1999), as empirically tested by Meisezahl (2024) for the acquisition of V2 grammars as depending on a high variability of constituents in first sentence position, with a high frequency of adjunct-verb orders in the primary linguistic data.

The relevance of frequency in the specific constructions of a language may impose a grammar change, to quote a paper by Hinterhölzl (2004). In other words, frequency changes may probably trigger micro- or nanoparameter changes in later linguistic stages. However, we would like to highlight that the results presented here should be considered not only from their quantitative significance, but also (and crucially) from their qualitative meaning as the pressure of language contact should also be considered among the speakers of ‘cultured’ Portuguese in Brazil. Further work on ColBP is needed to verify whether these results are confirmed in other texts and in other word order patterns.

Author Contributions: Conceptualization, A.L.d.A. and L.d.S.C.; methodology, A.L.d.A.; software: A.L.d.A. and L.d.S.C.; validation, A.L.d.A. and L.d.S.C.; formal analysis, A.L.d.A.; investigation, A.L.d.A. and L.d.S.C.; resources, L.d.S.C.; data curation, L.d.S.C.; writing—original draft preparation, A.L.d.A. and L.d.S.C.; writing—review and editing, A.L.d.A.; visualization, L.d.S.C.; supervision, A.L.d.A.; project administration, A.L.d.A.; funding acquisition, L.d.S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Council for Scientific and Technological Development (CNPq, Brazil), L.C.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available at the Center for Open Science (OSF) at https://osf.io/ugdh4/?view_only=bfd778ad066f4ef392949b0bede552d7, accessed on 15 July 2024, reference number 10.17605/OSF.IO/UGDH4. These data were derived from the following resources in the public domain: *Repositório Institucional da UFMG*: <https://repositorio.>

ufmg.br/handle/1843/ECAP-9JXPSX (Alkimim 2014) and the *Tycho Brahe Parsed Corpus of Historical Portuguese*: <https://www.tycho.iel.unicamp.br/corpus/> (Vieira 1907).

Acknowledgments: We thank three anonymous reviewers for their feedback and questions as well as the editorial assistance of Patricia Amaral and the audience at the XX International Conference of the Latin American Association for Linguistics and Philology (ALFAL, Project 3: History of Brazilian Portuguese, from Europe until America) where a portion of this material was presented. We thank the collaboration of Zenaide Carneiro and Mariana Lacerda, the directors of the CE-DOHS corpus, and of Charlotte Galves, who administrated the TBC corpus.

Conflicts of Interest: The authors declare no conflicts of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Notes

- 1 Regarding this issue, an anonymous reviewer has also mentioned that the acceptability of X*VS patterns is still a major point of variation among the Romance languages today.
- 2 This text shall be integrated into the *Electronic Corpus of Historical Documents of the Hinterland* (CE-DOHS; Carneiro and Lacerda 2024).
- 3 This text is available with syntactic annotation in the *Tycho Brahe Parsed Corpus of Historical Portuguese* (TBC; Galves et al. 2017).
- 4 Roberts’ (2019) proposal acknowledges that larger-scale parameters emerge due to the clustering of smaller-level parametric settings, unlike in the original rendering of microparameters by Kayne (2008). We thank an anonymous reviewer for pointing this out.
- 5 Lucchesi (2019) mentions these continuous and frequent interactions as reasons that led to the non-emergence of a Portuguese-based creole in Brazil. The debate about whether Brazilian Portuguese was a creole language was frequent at the end of the 20th century. However, nowadays no one assumes that this has been a characteristic of Brazilian Portuguese in any of its developmental stages (Tarallo 1993).
- 6 It is beyond the scope of this work to address the creation of Brazilian Portuguese varieties, which are divided, according to Lucchesi (1994), into a cultured variety, spoken by Brazilians with a higher level of education and greater socioeconomic condition, generally born in State capitals, and the popular variety, spoken by Brazilians with a lower level of education, having a lower socioeconomic status, generally born in the country’s hinterland. We recognize, however, that such varieties show considerable linguistic divergences, probably associated with the process of formation and the acquisition of Portuguese by European descendants on one hand and Africans and indigenous peoples and their descendants on the other hand, during the colonial period (cf. Mattos e Silva 2004 and Lucchesi 2015 for a more in-depth discussion regarding the origins of Brazilian Portuguese along these two sociolinguistic trends).
- 7 We consider that, although more fundamental grammatical properties of a linguistic system could hypothetically take place, the sociolinguistic context represented by ColBP, together with the lack of time to implement some changes, would not promote the advent of meso- or macroparameters.
- 8 This variety is commonly referred to as *português brasileiro culto falado* (‘cultured’ spoken Brazilian Portuguese).
- 9 Clitic placement in pre- or postverbal position (enclisis or proclisis) was guided by several syntactic contexts in CIP, some requiring proclisis, others requiring enclisis, and yet others allowing for variation between proclisis and enclisis. Among these, Galves et al. (2005) suggest the existence of a Variation Context 2, which includes sentences containing a verb in initial position in a second-coordinate clause or preceded by one or more dependent clauses (examples from André de Barros’ *The Life of the Apostolic Father Antônio Vieira*, 17th century, quoted in Galves et al. 2005, p. 49):

- | | | | | | | | | |
|-----|----|--|--------------|----------------|---------------|-------------|--------------|----------|
| (i) | a. | Achou-os | ditosamente, | falou-lhes, | e | rendeu-os | a | largarem |
| | | found-them | happily | talked-to.them | and | caused-them | to | leave |
| | | aquela vida | | brutal... | | | | |
| | | that life | | brutal | | | | |
| | | ‘[Vieira] found them happily, talked to them and caused them to leave that brutal life...’ | | | | | | |
| | b. | Para os | começar a | render, | amimou-os | com | donativos... | |
| | | to them | start to | surrender | pampered-them | with | donations | |
| | | ‘To start to surrender them, [Vieira] pampered them with donations...’ | | | | | | |

This context showed a tendency for enclisis combined with a more pronounced variation between the studied authors.

- 10 Examples of archives containing documents related to the colonization of Brazil are the Arquivo Histórico Ultramarino (AHU) and the Arquivo Nacional da Torre do Tombo (IAN-TT), both located in Portugal. The task of retrieving these texts may be costly and laborious because the collections are organized according to a specific classification that does not always meet the relevant criteria for linguistic research.

- 11 The sermons written by Father António Vieira have already been the subject of linguistic studies which, when compared with the letters written by the same author, pointed out the strong presence of stylistic marks interfering in the behavior of certain linguistic phenomena, such as the order of clitic pronouns in the sentence, cf. Galves (2002).
- 12 Available online: <https://digital.bbm.usp.br/handle/bbm/1>. accessed on 3 January 2024.
- 13 The lack of any relevant morphosyntactic changes in the new edition provides support for the choice of this material for linguistic research.
- 14 The CE-DOHS platform brings together a significant set of texts written in Brazil between the 16th and 20th centuries, the vast majority of which were produced by Brazilians of different ethnicities (Portuguese, Indigenous, and African descendants), in addition to containing oral data produced in the late 20th century. These documents have significant socio-historical control related to the authorship and social context of writing the manuscripts, in addition to providing an interface that allows an exploration of the material based on philological and/or grammatical goals.
- 15 G-tests were used if a table displayed at least one cell with less than 5 occurrences, and this option is indicated in the respective captions.
- 16 From an analysis of chi-square residuals, the difference between ColBP and CIP resides in the V2 factor (with 40% and 53%, respectively).
- 17 From an analysis of chi-square residuals, the difference between ColBP and CIP resides in the V > 2 factor (with 3% and 37%, respectively).
- 18 The TBC platform is composed of texts written in Portuguese between 1380 and 1978, with more than one million words of parsed text. Most of its texts are representative of CIP and modern European Portuguese, although some representative texts of Old Portuguese and modern Brazilian Portuguese have been recently included as well.
- 19 From an analysis of chi-square residuals, the difference between ColBP and CIP resides in the ‘DP accusative complement’ factor (with 10% and 3%, respectively).
- 20 There were no clear cases of Informational Focus at the left periphery, so this category was ignored in the table.
- 21 From an analysis of chi-square residuals, the difference between ColBP and CIP resides in the ‘contrastive focus’ factor (with 29% and 8.5%, respectively).
- 22 The two first discursive/informational functions (discourse marker and vocative) are placed at the external clausal domain; the last one (informational focus) is believed not to exist at the left periphery of Portuguese in any stage but was included there for the sake of completeness.
- 23 The total is not equivalent to the total of XVS sentences in Table 3 because there are other requirements regarding the category, removing other possible combinations.
- 24 As already mentioned, the notions of internal and external left-peripheral positions are indirectly motivated by the difference between enclisis and proclisis and directly motivated by their position in the syntactic marker, higher than FocP (cf. Benincà 1995). Interestingly, frame-setters are also external to the clause, but their percentages are similar in the two grammars. According to Benincà and Poletto (2004), they occupy the Spec, FrameP position at the very top of the left periphery; however, the results in Galves et al. (2005) suggest, that in CIP, frame-setters may occupy either a clause-external or clause-internal position, possibly Spec, KontrP in Galves and Kroch’s (2016) terms, depending on its length in terms of phonological words. Cf. Prévost (2003) for an explanation for why some locative phrases may be considered as focus-like.

References

- Alkimim, Ilma. 2014. *Sermões de Eusébio de Matos: Edição Crítica e Estudo*. Ph.D. dissertation, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.
- Andrade, Aroldo. 2015. On the Emergence of Topicalization in Modern European Portuguese: A Study at the Syntax-Information Structure Interface. *Estudos Linguísticos* 11: 13–34.
- Andrade, Aroldo. 2020. A gramática do português colonial brasileiro: Uma agenda de pesquisas. [Apresentação no painel “Sociolinguística Histórica, tratamento de corpora e história do português brasileiro”] Abralín ao Vivo: Linguists Online, December 7. Available online: <https://aovivo.abralin.org/lives/sociolinguistica-historica/> (accessed on 1 July 2024).
- Andrade, Aroldo, and Charlotte Galves. 2019. Contrast and word order: A case study on the history of Portuguese. *Glossa: A Journal of General Linguistics* 4: 107–29. [CrossRef]
- Barbosa, Afrânio. 1999. *Para uma História do Português Colonial: Aspectos Linguísticos em Cartas de Comércio*. Ph.D. dissertation, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil.
- Benincà, Paola. 1995. Complement clitics in medieval Portuguese: The Tobler-Mussafia Law. In *Language Change and Verbal Systems*. Edited by Adrian Battye and Ian Roberts. Oxford: Oxford University Press, pp. 325–44.
- Benincà, Paola, and Cecilia Poletto. 2004. Topic, Focus and V2: Defining the CP sublayers. In *The Structure of CP and IP: The Cartography of Syntactic Structures*. Edited by Luigi Rizzi. Oxford: Oxford University Press, vol. 2, pp. 52–76.
- Berlinck, Rosane. 1988. *A Ordem V SN No Português do Brasil: Sincronia e Diacronia*. Master’s thesis, Universidade Estadual de Campinas, Campinas, Brazil.

- Cardoso, Lara. 2020. A Gramática dos Pronomes Clíticos no Brasil Colônia: O Português Clássico na História do Português Brasileiro. Master's thesis, Universidade Estadual de Feira de Santana, Feira de Santana, Brazil.
- Cardoso, Lara, Aroldo Andrade, and Zenaide Carneiro. 2023. O português colonial brasileiro: Uma nova agenda de pesquisas entre o português clássico e o português brasileiro moderno. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada* 39: 202339253749. [CrossRef]
- Carneiro, Zenaide, and Charlotte Galves. 2010. Variação e Gramática: Colocação de clíticos na história do português brasileiro. *Revista de Estudos da Linguagem* 18: 7–38. [CrossRef]
- Carneiro, Zenaide. 2005. Cartas Brasileiras (1809–1904): Um Estudo Linguístico-Filológico. Ph.D. dissertation, Universidade Estadual de Campinas, Campinas, Brazil.
- Carneiro, Zenaide, and Mariana Lacerda. 2024. CE-DOHS: Corpus Eletrônico de Documentos Históricos do Sertão. Available online: <https://www.uefs.br/cedohs/> (accessed on 10 January 2023).
- Corôa, Williane. 2022. Rastreado as Origens do Português Brasileiro: A Dinâmica da Mudança na Escrita de “Homens Bons” na Bahia Colonial. Ph.D. dissertation, Universidade Estadual de Campinas, Campinas, Brazil.
- Galves, Charlotte, and Anthony Kroch. 2016. Main syntactic changes from a Principles-and-Parameters view. In *The Handbook of Portuguese Linguistics*. Edited by Leo Wetzels, Sérgio Menuzzi and João Costa. New York: John Wiley & Sons, pp. 487–503.
- Galves, Charlotte, Aroldo Andrade, and Pablo Faria. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. Available online: <https://www.tycho.iel.unicamp.br/corpus/> (accessed on 10 January 2023).
- Galves, Charlotte, Aroldo Andrade, Cristiane Namiuti, and Maria Clara Paixão de Sousa. Forthcoming. *Classical Portuguese: Grammar and History*. Oxford: Oxford University Press.
- Galves, Charlotte. 2002. Syntax and Style: Clitic-placement in Padre Antonio Vieira. *Santa Barbara Portuguese Studies* 6: 387–403.
- Galves, Charlotte. 2007. A língua das caravelas: Periodização do português europeu e origem do português brasileiro. In *Descrição, História e Aquisição do Português Brasileiro*. Edited by Ataliba Castilho, Maria Aparecida Torres Moraes, Ruth E. Vasconcellos Lopes and Sonia Maria Lazzarini Cyrino. Campinas and São Paulo: Pontes and Fapesp, pp. 513–28.
- Galves, Charlotte. 2020. Relaxed V-Second in Classical Portuguese. In *Rethinking Verb Second*. Edited by Rebecca Woods and Sam Wolfe. Oxford: Oxford University Press, pp. 368–95.
- Galves, Charlotte, Helena Britto, and Maria Clara Paixão de Souza. 2005. The change in clitic-placement from Classical to Modern European Portuguese: Results from the Tycho Brahe Corpus. *Journal of Portuguese Linguistics* 4: 39–67. [CrossRef]
- Götze, Michael, Thomas Weskott, Cornelia Endriss, Ines Fiedler, Stefan Hinterwimmer, Svetlana Petrova, Anne Schwarz, Stavros Skopeteas, and Ruben Stoel. 2007. Information Structure. In *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics and Information Structure*. Edited by Stefanie Dipper, Michael Götze and Stavros Skopeteas. Potsdam: Universitätsverlag Potsdam, pp. 147–87.
- Hinterhölzl, Roland. 2004. Language Change versus Grammar Change: What diachronic data reveal about the distinction between core grammar and periphery. In *Diachronic Clues for Synchronic Grammar*. Edited by Eric Fuß and Carola Trips. Amsterdam: John Benjamins, pp. 131–60.
- Kabatek, Johannes. 2008. *Sintaxis Histórica del Español y Cambio Lingüístico: Nuevas Perspectivas Desde las Tradiciones Discursivas*. Madrid: Vervuert Iberoamericana.
- Kayne, Richard. 2008. Some Notes on Comparative Syntax, with Special Reference to English and French. In *The Oxford Handbook of Comparative Syntax*. Edited by Guglielmo Cinque and Richard Kayne. Oxford: Oxford University Press, pp. 3–69.
- Lightfoot, David. 1999. *The Development of Language: Acquisition, Change, and Evolution*. Malden: Blackwell.
- Lose, Alicia. 2022. Ver más allá del texto: Análisis material de los Pasquines Sediciosos de la Revolución de los Sastres en Bahía en el siglo XVIII. *Espacio, Tiempo y Forma, Serie IV. Historia Moderna* 35: 71–96. [CrossRef]
- Lucchesi, Dante. 1994. Variação e norma: Elementos para uma caracterização sociolinguística do português do Brasil. *Revista Internacional de Língua Portuguesa* 12: 17–28.
- Lucchesi, Dante. 2015. *Língua e Sociedade Partidas: A Polarização Sociolinguística do Brasil*. São Paulo: Contexto.
- Lucchesi, Dante. 2019. Por que a crioulação aconteceu no Caribe e não no Brasil? Condicionamentos sócio-históricos. *Gragoatá* 24: 227–55. [CrossRef]
- Machado, Diogo Barbosa. 1741. *Bibliotheca Lusitana*. Lisboa Occidental: Oficina de Antonio Isidoro da Fonseca, vol. 1.
- Martins, Ana Maria. 2019. Against V2 as a general property of Old Romance languages. In *Romance Languages and Linguistic Theory 15: Selected papers from Going Romance 30, Frankfurt*. Edited by Ingo Feldhausen, Martin Elsig, Imme Kuchenbrandt and Mareike Neuhaus. Amsterdam and Philadelphia: John Benjamins, pp. 10–33.
- Martins, Marco Antônio. 2009. Competição de Gramáticas do Português na Escrita Catarinense dos Séculos 19 e 20. Ph.D. dissertation, Universidade Federal de Santa Catarina, Florianópolis, Brazil.
- Matras, Yaron. 2009. *Language Contact*. Oxford: Oxford University Press.
- Mattos e Silva, Rosa Virgínia. 2004. *Ensaio para uma Sócio-História do Português Brasileiro*. São Paulo: Parábola.
- Mattoso, Katia. 2003. *Ser Escravo no Brasil*, 3rd ed. São Paulo: Brasiliense. First published 1979.
- Meisezahl, Marc. 2024. Learning to Lose: The Role of Input Variability in the Loss of V2. Ph.D. dissertation, University of Edinburgh, Edinburgh, UK.
- Pagotto, Emílio. 1992. A Posição dos Clíticos em Português: Um Estudo Diacrônico. Master's dissertation, Universidade Estadual de Campinas, Campinas, Brazil.

- Prévost, Sophie. 2003. Les compléments spatiaux: Du topique au focus en passant par les cadres. *Travaux de linguistique* 47: 51–77. [CrossRef]
- Ramos, Jânia M., and Marilza de Oliveira. 2021. Introdução. In *História do Português Brasileiro, vol. 10: Dialetação e Povoamento. Da História Linguística à História Social*. Edited by Jânia M. Ramos and Marilza de Oliveira. São Paulo: Contexto, pp. 11–45.
- Randall, Beth, Antony Kroch, and Beatrice Santorini. 2004. CorpusSearch 2 Users' Guide. Available online: <https://www.ling.edu/~beatrice/corpus-ling/CS-users-guide> (accessed on 10 January 2023).
- Rizzi, Luigi. 1997. The fine structure of the left periphery. In *Elements of Grammar: Handbook of Generative Syntax*. Edited by Liliane Haegeman. Dordrecht: Kluwer, pp. 281–337.
- Roberts, Ian. 2019. *Parameter Hierarchies and Universal Grammar*. Oxford: Oxford University Press.
- Tarallo, Fernando. 1993. Diagnosticando uma gramática brasileira: O português d'aquém-mar e d'além-mar ao final do século XIX. In *Português Brasileiro: Uma Viagem Diacrônica*. Edited by Ian Roberts and Mary Kato. Campinas: Editora da Unicamp, pp. 69–105.
- Vieira, António. 1907. *Sermões, Pe. António Vieira*. Edited by Gonçalo Alves. Porto: Chardron/Lello & Irmão Editores.
- Wolfe, Sam. 2015. Microvariation in Medieval Romance Syntax: A Comparative Study. Ph.D. dissertation, University of Cambridge, Cambridge, UK.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

MDPI AG
Grosspeteranlage 5
4052 Basel
Switzerland
Tel.: +41 61 683 77 34

Languages Editorial Office
E-mail: languages@mdpi.com
www.mdpi.com/journal/languages



Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.



Academic Open
Access Publishing

mdpi.com

ISBN 978-3-7258-2061-0