Que ao mestre vai matá-lo? The evolution of prepositional accusatives in Portuguese across time

Helena Rodrigues Menezes de Oliveira Vaz
Universität Wien / Vienna, Austria
rodriguesh25@univie.ac.at

Abstract

This work investigates Differential Object Marking (DOM) in Brazilian Portuguese (BP), specifically *a*-marked objects, or prepositional accusatives (PP-ACCs), across four variables: semantic requirements, constituent order, verb semantics, and textual genre. An optimized parsing model was trained to recognize instances of PP-ACCs and automatically annotate historical documents for these objects for the *Tycho Brahe* and *Colonia* corpora. Contrary to expectations based on the low frequency of these objects and prior diachronic studies on European Portuguese (EP), our results reveal that PP-ACCs remain present in BP from the 18th century onward. Our findings confirm previous patterns for EP and present textual genre (specifically, narrative texts and theater plays) as a possible relevant variable, but this warrants further investigation. Constituent order was proved to be less significant than previously suggested. This work also reveals methodological challenges in using computational models and NLP tools for research in historical Portuguese.

1 Introduction

Differential Object Marking (DOM) is a morphosyntactic phenomenon in which a group of direct objects in a language receive a special overt marker, such as a case morpheme or a preposition, while others remain unmarked (Schwenter, 2014). In Romance languages, *a*-marked objects, also called prepositional accusatives (PP-ACCs), are one of the most investigated structures in DOM.

Spanish is a good example of a language with such clear DOM marking: the preposition *a* marks a subset of direct objects, obligatory when it is an animate one. For instance:

1) Spanish (Cyrino and Irimia, 2019):

(a) He visto **a** tu padre. 'I have seen (to-DOM) your father.'

(b) *He visto **a** tu coche. 'I have seen your car.'

A-marking in Portuguese, on the other hand, is more marginal and less systematic, mainly in optional contexts (Neves, 2011; Cunha and Cintra, 2017; Cyrino, 2017), such as the use of *ao* in (2b). However, it is still found in structures such as in clitic doubling and in topicalized objects, and in comparative and coordinated structures, as long as the object is animated (Cyrino, 2017).

2) Optional object marking (Cyrino and Irimia, 2019):

(a) Pedro viu o menino. 'Pedro saw the boy.'

(b) Pedro viu **ao** menino. 'Pedro saw (to-DOM) the boy.'

Several works have investigated the topic of PP-ACCs in European Portuguese (EP) from a diachronic perspective, particularly between the 16th and 19th centuries (Gibrail, 2003; Döhla, 2014; Pires, 2017, 2020; Calindro, 2024; Cyrino and Irimia, 2019). According to these authors, *a*-marking had its peak during the 17th century but has progressively fallen into disuse from the 18th century onward. However, there is a gap in studies that investigate the behavior of these objects in Brazilian Portuguese (BP), given most of the literature so far has suggested that they are practically inexistent in this variety.

Previous works, notably Gibrail (2003), Pires (2017, 2020), Calindro (2024) have all utilized documents from the *Tycho Brahe Parsed Corpus of Historical Portuguese* (Galves and Faria, 2017) for investigating *a*-marked objects. The corpus has syntactically annotated documents, with a tag for prepositional accusatives (PP-ACC), and as such has been extensively used in diachronic research in *a*-marking in Portuguese. However, not all corpus documents have this annotation available, and there is an imbalance in the number of annotated

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 93–102

documents in EP and BP. There are more annotated documents for EP in the corpus than for BP, making investigations in this variety difficult to carry out due to lack of available data, a methodological gap this study sought to address.

Given the limitations of existing resources, and the fact that there are no existing language models and Natural Language Processing (NLP) applications for historical Portuguese (to the best of our knowledge), data collection for this type of phenomenon becomes a manual and very time consuming task. This study is thus the first attempt to automatize data collection for *a*-marked objects in historical Portuguese, with the goal of addressing the lack of data for BP and providing initial results for the phenomenon in this variety.

Hence, here we present initial findings for object marking in BP from a diachronic perspective, and compare these with what is known about these objects in EP. We look at their frequency across time and analyze the roles that semantic requirements, constituent order and verb semantics play in the realization of these objects, as well as textual genre, not previously explored, not even for EP. As the automatization of data collection for PP-ACCs makes it possible for a broader variety of documents to be analyzed, we were interested in determining whether textual genre can influence the frequency and context of realization of these objects, given Döhla (2014) and Pires (2017) have suggested differences in the frequency of these objects across genre, reflecting a possible discursive and stylistic role of PP-ACCs.

In order to conduct this investigation, we trained a parsing model to recognize contexts of occurrence of these objects and automatically extract candidate sentences with PP-ACCs from unannotated documents from the *Tycho Brahe Parsed Corpus of Historical Portuguese* (Galves and Faria, 2017) and the *Colonia Corpus of Historical Portuguese* (Zampieri and Becker, 2013). We used automatized methods to extract constituent order and most frequent verbs with PP-ACCs.

With this approach, we were able to expand the dataset available for EP and obtain data for BP to investigate object marking, allowing for new patterns to be identified and an initial comparison to be established between the two varieties. This study also points out limitations in using NLP resources for diachronic research in Portuguese, opening new avenues for future investigations and development of new applications.

2 Literature review

Research on *a*-marked objects in Portuguese has been predominantly diachronic and focused on EP. Ramos (1992) was the first study to investigate *a*-marked objects in BP, and has approached the fall of *a* in front of direct objects as an extra case marking resource. Gibrail (2003) has used the *Tycho Brahe* corpus (Galves and Faria, 2017) to conduct an analysis of *a*-marking in EP from the 16th–19th centuries, providing a description of the behavior of these objects at syntactic and semantic levels. Döhla (2014) has compared *a*-marked objects in Spanish in Portuguese from the 13th-19th centuries, proposing that these objects are triggered by syntactic factors, mainly parallelism, topicalization and VSO order. Additionally, Döhla (2014) has proposed that PP-ACCs in Portuguese were heavily influenced by the prestige Spanish had in Portugal during the Iberian Union (1580-1640), period both crowns were unified. This hypothesis was confirmed by Pires (2017, 2020), who also investigated the period between the 16th-19th centuries and found that the contexts of occurrence for PP-ACCs differed between the two languages; most notably, in the 17th century, the increase in the frequency of object marking led to an expansion of its usage contexts. During this time, this construction became more structured around certain categories, such as proper names, relative pronouns, titles and forms of address (Pires, 2017). Thus, PP-ACCs in Portuguese are not parallel to *a*-marked objects in Spanish.

Other studies, such as Cyrino (2017), Cyrino and Irimia (2019) and Calindro (2024) have also investigated syntactic and semantic characteristics of PP-ACCs, both in EP and BP. The first two investigated features such as animacy restriction, verb government and the dative-accusative case alternance observed in other Romance languages, such as Catalan (Cyrino, 2017). The latter also analyzed the dative case shift in object marking but focused on verb semantics, investigating the high frequency and steady rise of PP-ACCs with psych verbs in EP between the 16th-19th centuries (Calindro, 2024).

3 Methodology

3.1 Data

As mentioned, two historical Portuguese corpora were used in this study: the *Tycho Brahe Parsed Corpus of Historical Portuguese* (Galves and Faria

(2017); approx. 3.8M words, 95 texts, 14th–20th c.) and the *Colonia Corpus of Historical Portuguese* (Zampieri and Becker (2013); approx. 5M words, 95 texts, 16th–20th c.). Both include BP and EP and diverse genres, such as letters, plays, narratives, dissertations, poetry, etc. The *Tycho Brahe* also has documents with morphological and/or syntactic annotations, whereas the *Colonia* is automatically annotated for Part of Speech (POS).

A mini-corpus was compiled using texts from both these corpora, balancing century and genre across varieties. The main issue regarding balance between varieties was that, in the *Tycho Brahe*, there are distinctly more EP texts than BP ones. To compensate for this imbalance, 11 documents in total were selected from the *Colonia* corpus and distributed accordingly. Therefore, from the *Tycho Brahe* there is a total of 40 documents in the mini-corpus: 23 for EP distributed in 3 genres (Letters, Newspapers and Prose) and 17 for BP distributed in 5 genres (Letters, Newspapers, Theater, Minutes/Records and Prose). For the *Colonia*, 5 documents from EP were selected (3 Letters and 2 Prose), and 6 for BP (all Prose). For the latter there were no documents from the 16th century, and for the 17th there were only Letters available for this variety in both corpora. In the end, the mini-corpus had 51 files (28 for EP in 3 genres, and 23 for BP in 5 genres)[1]. Syntactically annotated *Tycho Brahe* documents were used as training data, and only unannotated ones from the corpus were used for the linguistic analysis.

3.2 Parsing and model training

Automatizing the annotation of PP-ACCs in a corpus is a quite challenging task due to the low frequency of the phenomenon. A first attempt was made at searching for *a* as a preposition by parsing the documents with Stanza (Qi et al., 2020) and converting them to CoNLL-U format, which was not very productive, as *a* is often misclassified as the feminine determiner *a* by Stanza. Since searching for just *a* as preposition in the parsed documents was quite unsatisfactory, we concluded that automatic identification of PP-ACCs required contextual cues beyond POS tags. Our goal with this methodology became then to identify patterns that could be generalized across the remaining corpus documents and support the training of a model to automatically disambiguate *a* as preposition from determiner, and identify when it was followed by a direct object, thus minimizing manual annotation. The trained model was combined with a rule-based filter, which retrieved candidate PP-ACC sentences where *a* functioned as a preposition and introduced a potential direct object.

From the *Tycho Brahe*'s syntactically annotated documents selected (10 EP, 6 BP), 492 sentences containing 641 PP-ACCs were extracted. All previous annotations of these documents (in the Penn Treebank format) were removed, as there were several compatibility issues in the conversion of the sentences. The Universal Dependencies (UD) (de Marneffe et al., 2021) annotation scheme was adopted here. Parsing introduced several mistakes in the data, which was already expected due to current models and NLP tools for Portuguese not having been trained on diachronic data. This led to segmentation issues, wrong lemma attribution, wrong POS tagging and consequentially, many mistakes in the dependency relations. For instance, in subject and object attribution in OVS and SOV sentences, as several PP-ACCs were tagged as indirect objects or obliques when they were in fact direct objects. Manual verification was needed to ensure all the sentences extracted contained PP-ACCs and were correctly annotated: 693 PP-ACCs were manually corrected in Arborator-Grew (Guibon et al., 2020), serving as gold standard for the training of the model.

However, 492 sentences is not a lot of data for training a model, which is why more training data was needed. In total, three models were trained (cf. Table 1), with the first attempt being done with synchronic data. 650 sentences were sampled from the *Bosque* (Rademaker et al., 2017) and *Portinari* (Duran et al., 2023) treebanks each, joined with the gold-standard file and trained with SpaCy (Honnibal and Montani, 2020), totaling 1.142 sentences. This first model correctly tagged prepositions as possible PP-ACCs rather than determiners with several example-sentences. However, the scores for precision, recall and F-score, as well as Labeled (LAS) and Unlabeled Attachment Score (UAS) were quite low compared to *Bosque*, for instance, whose scores are presented in Table 2 for comparison.

The next attempt was with diachronic data: 8 randomly sampled documents from the *Colonia* corpus were stripped of all previous annotation and parsed with Stanza (Qi et al., 2020). 2.000 sen-

[1]Theatre and Minutes/Records were not available in EP, but they were kept in the analysis to balance out the number of words for each century bin for both varieties.

Model nº	Training data	LAS	UAS	SENT F	F-score OBJ	F-score OBL
1	1,142 (*Bosque* + *Portinari* + *Tycho Brahe* sentences)	64.42	74.11	80.17	59.45	57.72
2	2,492 (*Colonia* + *Tycho Brahe* sentences)	86.06	79.60	95.55	79.92	76.82
3	2,492 (*Colonia* + *Tycho Brahe* sentences)	83.94	89.47	96.44	81.36	76.03

Table 1: Performance metrics for the models trained.

Sent F	LAS	UAS
93.10	87.75	90.59

Table 2: Bosque (Rademaker et al., 2017) performance scores

tences with potential PP-ACC contexts (with *a* as either preposition or determiner) were randomly selected from these documents and joined with the gold-standard file, totaling 2.492 sentences. It was also trained with SpaCy (Honnibal and Montani, 2020), and evaluation metrics were promising. However, when running this model on the unannotated documents, the CoNLL-U files presented conversion and formatting issues such as 'XPOS columns = None', inconsistent number of columns and segmentation errors, which led to the POS tags not being recognized by the parser. After resolving these issues by reconverting the files, the third model achieved the most stable performance, as presented in Table 1.

3.3 Manual annotation and analysis

The third and final model trained was run on the unannotated *Tycho Brahe* and *Colonia* documents to extract potential PP-ACC contexts, which were defined as: *a* as ADP, headed by either a NOUN, PROPN, DET or PRON with a CASE relation, in turn headed by a VERB with an OBJ or OBL relation. Each PP-ACC candidate was then manually verified. The number of sentences selected for each document of the corpus ranged from to 3 to up to 1.450 sentences. For this study, an initial phase of a broader analysis, all documents containing potential PP-ACC sentences were manually checked; however, the ones that had more than 200 sentences were checked up to the 150th one. [2] The complete verification of the documents will be done in future works.

For the semantic features annotation, each PP-ACC found was annotated for animacy, specificity and definiteness by adding [+/-animate], [+/-spec], [+/-def] tags in the misc column of the CoNLL-U files. Constituent order was estimated automatically with the *conllu* Python library (Stenström, 2016), based on token IDs rather than dependency relations. This was done to compensate for possible errors in the parsing of the documents by looking at the position of the tokens in the sentences to increase accuracy in word order prediction, rather than relying solely on the dependency relations. Five main orders (SVO, VSO, SOV, OVS, Other) were attested, but many misclassifications were observed, so we had to partly rely on the automated parsing and manually annotate a subset of the data to check its reliability. A balanced subset of 180 sentences was manually checked to assess accuracy, and ratios were compared between the total corpus and subset.

For verb semantics analysis, verbs occurring with PP-ACCs were grouped into 12 classes (e.g. psych verbs, contact, change of possession, perception, etc) following the criteria in *VerboWeb* (Amaral et al., 2017), a database for Portuguese verbs and their semantic-syntactic classification. Frequencies were compared across variety and century.

4 Results

A frequency-based analysis was conducted across four variables: century and genre, constituent order, semantic requirements and verb semantics. While the limitations in parsing introduced noise to the automatic detection of PP-ACCs, this first computational approach to data collection of *a*-marked objects in historical Portuguese demonstrates the feasibility of applying NLP to diachronic data and highlights areas for future improvements. Results for this initial attempt leave space for further developments and contributions, and shed light on the limitations of current models.

[2] For the normalization of the data and results, the word count for all documents that were not entirely checked was adjusted, in order to not skew the bins.

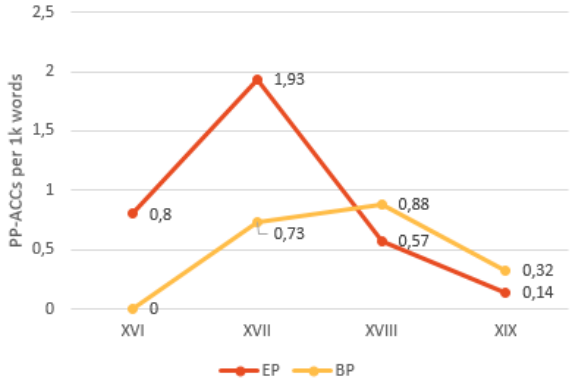

Figure 1: Normalization across century and variety

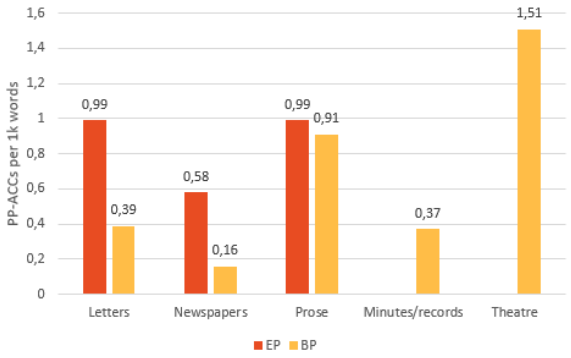

Figure 2: Normalization across genre and variety

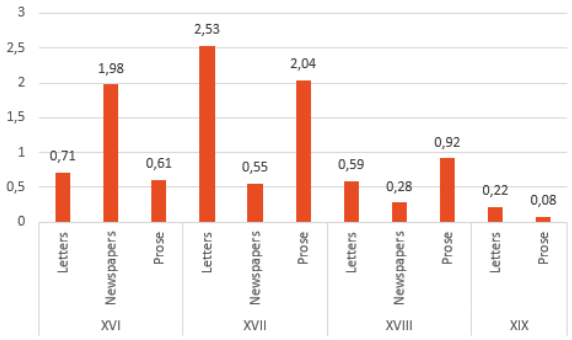

Figure 3: Normalization for century and genre for EP

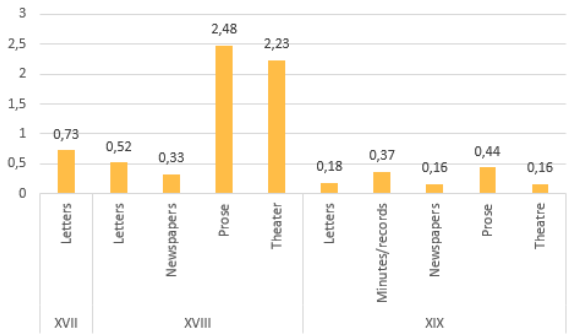

Figure 4: Normalization for century and genre for BP

4.1 Parsing results

Before diving into linguistic results, a few notes on the errors and mistakes in parsing are in order to understand the limitations of this work. Parsing errors mainly involved segmentation, lemma attribution, and dependency and POS labeling, especially in 16th–17th century texts, where long sentences and orthographic variation potentially reduced accuracy. Despite those issues, Stanza provided satisfactory results for later centuries. The final model (as seen in Table 1) trained on gold-standard and diachronic data from the *Colonia* corpus achieved an F-score of 96.4, UAS 89.5, and LAS 83.9. It effectively disambiguated *a* as preposition from determiner in most contexts. The resulting annotations expanded both the *Tycho Brahe* and *Colonia* corpora with annotations for PP-ACCs.

4.2 Results for PP-ACCs

4.2.1 Genre

The normalized frequencies presented in Figure 1 confirm patterns identified by the literature (Gibrail (2003); Pires (2017); Calindro (2024); Döhla (2014)): a peak in the 17th c. for EP (1.93) and in the 18th c. for BP (0.88), followed by a general decline in the 19th. EP shows higher frequencies until the 17th c., whereas BP maintains more stable and proportionally higher rates in the 18th and 19th c., suggesting this construction persisted longer in this variety. These numbers are presented in Figures 1, 2, 3 and 4 in bins of century, genre per variety and genre per century for each variety, respectively.

The 17th-century peak in EP observed in Figure 1 likely reflects Spanish influence during the Iberian Union (1580–1640) (Döhla, 2014; Pires, 2017, 2020), when DOM patterns were more salient. The later decline aligns with the post-Union linguistic distancing, which Döhla, 2014 described as the result of "a strong apathy" towards Spain and desire to separate the two languages during the Age of Enlightenment (ca. 18th century) in Portugal, preserving the linguistic structures of Portuguese and aiming towards its autonomy as a language. On the other hand, BP retained more PP-ACCs than EP into the 18th–19th centuries, perhaps due to broader tolerance of colloquial or regional variants.

This attested higher frequency of PP-ACCs in

BP from the 18th century onward was not previously pointed out in the literature. A possible reason for that could be the addition of new data, but also to an influence of genre (cf. Tables 2, 3 and 4), not previously investigated with PP-ACCs. Figure 2 shows that in EP, Letters and Prose exhibit the highest PP-ACC frequencies (0.99), particularly formal letters of the 17th century (2.53, cf. Figure 3), often containing formulaic expressions such as *Guarde Deus a...* or *Fico para servir a...*, likely influenced by Spanish and diffused as a formal epistolary style.

In Prose, EP frequencies drop steadily from the 17th (2.04) to the 19th century (0.08), as observed in Figure 3, while BP shows the opposite trend: peaking in the 18th (2.48) and remaining higher than EP in the 19th (0.44), observed in Figure 4. This divergence supports the idea that BP has retained PP-ACCs for longer, despite being a marginal construction, perhaps due to an ongoing process of syntactic simplification that did not eliminate *a*-marking entirely.

In BP, on the other hand, Theater (1.51) and Prose (0.91) show the highest frequencies of PP-ACCs. Theatrical language likely favors *a*-marked objects due to its proximity to colloquial speech and inherent syntactic flexibility, which may manifest through stylistic parallelism and topicalization (cf. Döhla, 2014). This aligns with the view of BP as a discourse-oriented language (Negrão and Viotti, 2000) with a strong tendency toward topicalization, which may account for the higher incidence of PP-ACCs in theatrical and narrative texts.

More accessible genres such as Newspapers and Minutes/Records [3] display low frequencies of PP-ACCs (0.16 and 0.37, respectively), which is consistent with a clearer language, with a preference for avoiding marked constructions. It could be the case that these objects were progressively dropped, but did not disappear completely, as the frequencies for these genres in the 18th and 19th c. suggest (cf. Figure 4).

4.2.2 Constituent order

A subset of 90 manually annotated sentences per variety was compared with the automatically parsed dataset (960 sentences in total, 655 EP and 305 BP) to assess reliability. The comparison revealed major discrepancies: while manual annotation showed a clear dominance of SVO order (around 60% in both EP and BP), automatic processing underestimated it (28% BP; 34% EP) and inflated non-canonical patterns such as SOV, VSO, etc. Consequently, only the manually corrected subset was used for the analysis, and the percentages of word orders for both varieties can be observed in Figure 5.

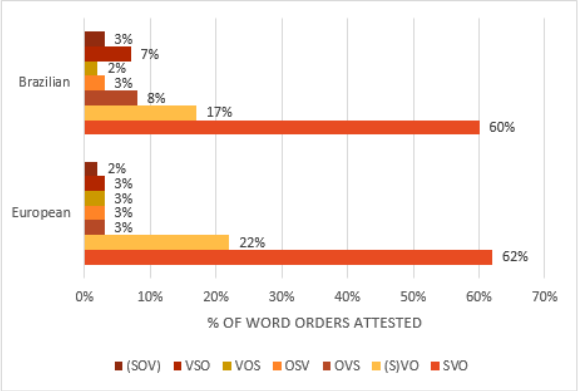

Figure 5: Percentages for total orders in the smaller dataset

The distribution of these orders across centuries, depicted in Figure 6, show that in EP the 16th century displays more variability (OSV 12%, VSO 8%, both SOV and VOS at 4%), which rapidly diminishes in the 17th, when SVO rises to 80%. Later centuries show a stabilization of canonical (S)VO patterns and a general simplification of word order. Since SVO reduces the need for *a*-marking, word order likely played a smaller role in licensing PP-ACCs than previously assumed, a hypothesis also postulated by Calindro (2024). Thus, the 17th c. peak of PP-ACCs may reflect genre-related, discursive or contact factors, such as the influence of Spanish, rather than syntactic ones. The growing preference for SVO order may, in turn, have contributed to the further decline of PP-ACCs in EP after the 17th century.

In BP, however, variation becomes increasingly diverse over time. While only three orders were attested in the 17th c. (SVO 46%, VSO 39%, OVS 15%), the 18th and 19th centuries displayed six different configurations, with SVO increasing steadily (56% in the 18th to 74%in the 19th). Despite this movement towards canonical order, PP-ACCs remain relatively frequent, suggesting further that word order alone does not determine *a*-marking. According to Cyrino (2017), PP-ACCs in BP often function as a disambiguation strategy, especially in coordination structures, preserving argument clarity even in SVO contexts.

[3] Present only in the BP 19th century bin.

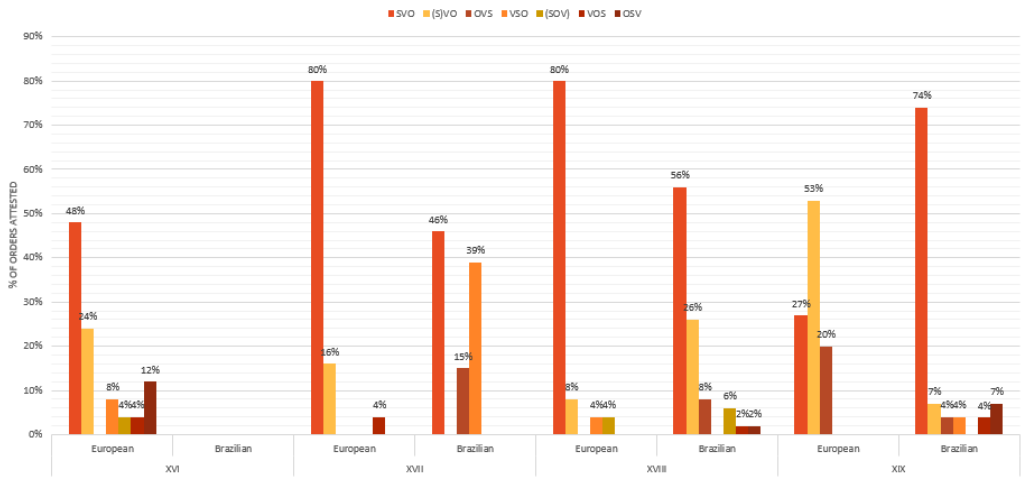

Figure 6: Distribution of word orders across time in the smaller dataset

This discourse-oriented pattern aligns with the well-known topic-prominence of BP (Negrão and Viotti, 2000): topicalization and left dislocation heighten object salience, favoring *a*-marking (Döhla, 2014). These tendencies are most visible in narrative and dramatic, theatrical genres, which employ more spoken-style syntax and topicalized structures as a stylistic device. The persistence of OVS and VSO orders in these contexts supports the idea that PP-ACCs in BP function primarily as discourse-driven markers rather than syntactic ones, maintaining some level of usability despite the general reduction of *a*-marking.

4.2.3 Semantic requirements

Results for animacy, specificity, and definiteness confirm previous evidence that PP-ACCs are favored with [+animate], [+specific], and [+definite] objects (Aissen, 2003; Gibrail, 2003; Cyrino, 2017; Pires, 2020). The results depicted in Figure 7 and here described highlight the patterns found for BP.

In BP, the distribution of [+animate] objects closely mirrors that of EP, with a peak in the 18th century (86%), when PP-ACCs are most frequent (Pires, 2017, 2020). Yet, as observed in Figure 7, BP also shows a comparatively higher share of [–animate] objects, in the 17th century (29%) and rising again in the 19th (20%). These cases often correspond to abstract entities or institutional referents, and others are quantifiers or indefinite pronouns with underlying [+animate] features, as also noted by Gibrail (2003), and exemplified in examples 1 and 2 below. Their gradual increase may suggest a possible connection between PP-ACCs and other DOM phenomena in BP, such as null objects, which unlike in EP or Spanish, occur canonically with [-animate] or [-specific] referents (Schwenter, 2014).

1. Você chama **a isso** brincar? [-animate, +spec, +def]
   'You call to-DOM this play?' (*t_001*)

2. E pedimos agora muito a Vossa Mercê consulte **a este ponto** com os maiores letrados dessa corte (...) [-animate, +spec, +def]
   'And we now strongly urge Your Grace to consult to-DOM this point with the most learned men of this court. (...)' (*va_017*)

   A similar trend emerges for specificity. While most PP-ACCs involve [+specific] objects, typically proper names, pronouns, and forms of address (Gibrail, 2003), the proportion of [–specific] objects also rise steadily, reaching 20% in BP and 30% in EP by the 19th century. These usually correspond to quantifiers or indefinite pronouns, indicating a gradual broadening of PP-ACC contexts beyond referentially stable entities, as shown in the examples below:

3. porque eu só com esta espada hei de vencer **a quantos poetas** há no mundo. [+animate, -spec, -def]
   'Because only with this sword will I be able to conquer to-DOM all the poets in the world.' (*s_004*)

4. Quando uma moça ama **a um homem**, ele pode apanhá-la com amor; (...) [+animate, -spec, -def]
   'When a young lady loves to-DOM a man, he can win her over with love, (...)' (*s_003*)

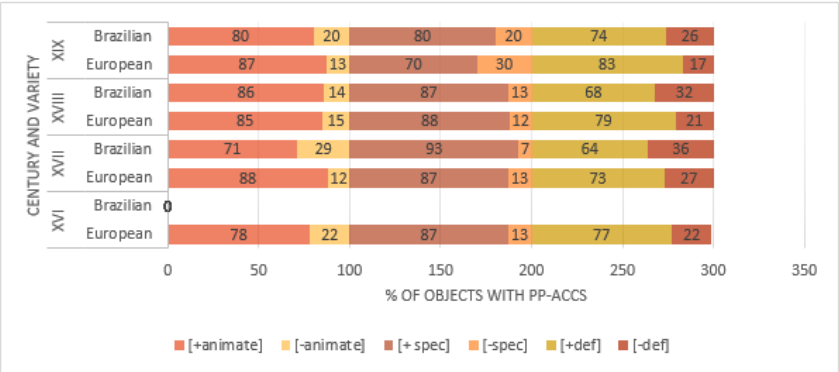

Figure 7: Results for animacy, specificity and definiteness

Definiteness displays lower overall variation but the same directional trend. In EP, [+definite] objects increase from 77% (16th c.) to 83% (19th c.), while BP rises from 64% (17th c.) to 74% (19th c.), though BP overall displayed higher proportions of [-definite] objects, as also observed for animacy and specificity.

Together, these tendencies suggest a functional reanalysis of *a*-marking in distinct cases. The steady increase of [+definite] objects, contrasting with the rise of [-animate] and [-specific] ones, suggests a distancing of some of these objects from their canonical semantic requirements. For instance, some PP-ACCs appear with abstract or generic definite referents, typically [-animate] and [-specific], indicating a broader functional range of *a*-marking in later centuries. These cases, exemplified in 5, are more exception than rule, but they point to a gradual reanalysis of PP-ACCs as being more discourse-driven rather than semantically constrained. However, they require more investigation and are beyond the scope of this work.

5. E não só ama **a estes três objetos**, só não acha neles um certo encanto (...) [-animate, +spec, +def] 'And not only does he love to-DOM these three objects, he just doesn't find a certain charm in them.' (*va_010*)

4.2.4 Verb semantics

In this section, we investigated the verbs that occur most frequently with PP-ACCs and their behavior in terms of thematic structure. The classes and verbs with the highest concentrations of PP-ACCs in both varieties were: *Locatum* transfer (*mandar* 'to send', *ajudar* 'to help'), Elucidation (*chamar* 'to call'), Perception (*ver* 'to see', *ouvir* 'to hear'), Permanent change (*matar* 'to kill', *vencer* 'to win') and Psych verbs (*amar* 'to love'). For thematic structure, we confirmed previous findings (Cyrino, 2017) that these objects usually occur with two possible structures: agentive verbs selecting Target/Patient/Beneficiary complements, as is the case with *Locatum*, Elucidation and Permanent change verbs; and verbs selecting Experiencer/Stative Object complements, as in Perception and Psych verbs.

In this section, we compared our results with Calindro (2024), whose work also explored PP-ACCs from the perspective of verb semantics. Her main finding was that Psych verbs in EP show a steady increase in PP-ACCs, which is linked to the persistence of *a*-marking introducing Experiencer arguments (e.g., O vinho agradou **ao** João; 'The wine pleased John'). In BP, however, *a* has disappeared from these contexts. Thus, where in Modern EP the Experiencer receives structural dative case via *a*, in Modern BP it receives inherent accusative case without overt marking.

Calindro (2024) and the present study, however, could not be directly compared due to a mismatch in both scope and methodology. While her work focused on the accusative-to-dative case shift and its connection to object marking (also approached by Cyrino, 2017), this study only maps verbal distribution across classes. Moreover, since Calindro's analysis did not include BP, comparisons were limited to EP.

The main source of divergence concerns the definition of the verb classes, as verb classification based on semantic-syntactic properties is inherently subjective. For instance, there were classes that were not present in (Calindro, 2024) but were here and vice versa, and classes which overlapped

but had the same verbs being classified in different ones. For instance, where Calindro classified *ouvir* 'to hear' as a Contact verb, we considered it as Perception. Calindro's second most frequent verb group (social interaction verbs, such as *abraçar* 'to hug', *convidar* 'to invite' and *atacar* 'to atack') was distributed across multiple classes in this study (e.g., *abraçar* as a Contact verb, *convidar* as Elucidation). These discrepancies likely contributed to differing results for Psych verbs: whereas Calindro, 2024 reported an increase of PP-ACCs with these verbs over time, the present data shows a decline, with no Psych verbs occurring with PP-ACCs in EP by the 19th century. Because of these classification differences and the experimental nature of this analysis, results were preliminary for this comparison. Future investigations are needed, incorporating BP data, to standardize verb class criteria to yield more consistent insights into the relationship between verb semantics and object marking.

5 Conclusion and future works

This study offered a first computational investigation of object marking in Portuguese, addressing a notable gap in the literature for this phenomenon in BP, where annotated diachronic data remain scarce. Using available language models and NLP tools, we collected new data for BP and compared it with EP across four variables: genre, constituent order, semantic requirements and verb semantics.

Using the *Tycho Brahe* corpus, we extracted instances of *a*-marked objects to be used as training data for an optimized parsing model. This model was was used to automatically recognize PP-ACC contexts and extract candidate sentences, which were then manually verified. The data collected confirmed the well-known 17th-century peak of PP-ACCs in EP; however, it revealed that in BP the peak occurs in the 18th century, with the frequency of this construction surpassing that of EP thereafter. Genre emerged as a potentially relevant factor, especially in BP, where narrative and theatrical texts, genres with a richer discourse structuring, exhibited the highest frequencies of PP-ACCs. For animacy, specificity, and definiteness, these objects behaved largely as expected in both varieties, although BP displayed a wider tolerance for [-animate] and [-specific] objects. Verb semantics yielded promising insights, but requires further investigations with more controlled classification criteria for stronger conclusions. Finally, constituent order was found not to be as relevant for object marking as previously thought, although more studies are needed to confirm this claim.

Methodologically, this study demonstrates both the potential and limitations of current computational tools to historical Portuguese. The parsing errors that hindered this study highlight the urgent need for NLP resources tailored to diachronic varieties of Portuguese, comparable to those already available for English and French, for instance. With more robust tools, it would be possible to test the patterns observed here using statistical and regression models, especially by comparing the frequencies of *a*-marked versus unmarked direct objects. Such quantitative modeling was not feasible here due to time constraints and noise in the automatically parsed dataset, but it represents a key direction for future research.

Our study offers an initial yet valuable picture of how *a*-marked objects behave diachronically in both EP and BP, and open new avenues for research at the intersection of historical linguistics, typology, and computational methods. With improved resources and larger, cleaner datasets, future studies will be able to confirm these patterns and deepen our understanding of the evolution of object marking in Portuguese.

Acknowledgments

I would like to thank Jun.-Prof. Dr. Annemarie Verkerk at Universität des Saarlandes for her supervision of this work, and Dr. Diego Fernando Válio Antunes Alves for his valuable suggestions and support in the publication of this paper.

References

Judith Aissen. 2003. Differential Object Marking: Iconicity vs. Economy. *Natural Language & Linguistic Theory*, 21(3):435–483.

Luana Amaral, Márcia Cançado, and Letícia Meirelles. 2017. VerboWeb: syntactic-semantic classification of Brazilian Portuguese verbs. Lexical Database. UFMG. Available online at: http://www.letras.ufmg.br/verboweb.

Ana Regina Calindro. 2024. A Diachronic Overview of the Prepositional Accusative in Portuguese. *Languages*, 9(6):194.

Celso Cunha and Luís Filipe Lindley Cintra. 2017. *Nova gramática do português contemporâneo*. Lexicon. Copyright, Rio De Janeiro, Brasil.

Sônia Cyrino. 2017. Reflexões sobre a marcação morfológica do objeto direto por a em Português Brasileiro. *Estudos Linguísticos e Literários*, Special(58):83.

Sônia Cyrino and Monica-Alexandrina Irimia. 2019. Differential Object Marking in diachrony: the case of Brazilian Portuguese. *Revista Letras*, 99(1).

Marie-Catherine de Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Magali Duran, Lucelene Lopes, Maria das Graças Volpe Nunes, and Tiago Pardo. 2023. The Dawn of the Porttinari Multigenre Treebank: Introducing its Journalistic Portion. In *Proceedings of the 14th Symposium in Information and Human Language Technology (STIL)*, pages 115–124.

Hans Döhla. 2014. Diachronic convergence and divergence in differential object marking between Spanish and Portuguese. In Kurt Braunmüller, Steffen Höder, and Karoline Kühl, editors, *Stability and Divergence in Language Contact: Factors and Mechanisms*, pages 265–289. John Benjamins, Amsterdam.

Charlotte Galves and Pablo Faria. 2017. Tycho Brahe Parsed Corpus of Historical Portuguese. URL: texts/psd.zip.

Alba Gibrail. 2003. *O Acusativo Preposicionado Do Português Clássico: Uma Abordagem Diacrônica E Teórica*. Master thesis, Universidade Estadual de Campinas (Unicamp).

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Matthew Honnibal and Ines Montani. 2020. spaCy: Industrial-strength natural language processing in python. Software Library.

Esmeralda Vailati Negrão and Evani Viotti. 2000. Brazilian Portuguese as a Discourse-Oriented Language. In Mary Kato and Esmeralda Vailati Negrão, editors, *Brazilian Portuguese and the Null Subject Parameter*, page 105–125. Iberoamericana: Madrid; Vervuert: Frankfurt am Main.

Maria Helena Neves. 2011. *Gramática de usos do português*, 2nd ed. edition. Editora Unesp, São Paulo.

Aline Pires. 2017. *A Marcação Diferencial de Objeto no Português: Um Estudo Sintático-Diacrônico*. Master thesis, Universidade Estadual de Campinas (Unicamp).

Aline Pires. 2020. A influência da gramática espanhola na Marcação Diferencial de Objeto no português diacrônico. *Cadernos de Linguística*, 1(2):01–20.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. ArXiv preprint arXiv:2003.07082.

Alexandre Rademaker, Fabricio Chalub, Livy Real, Cláudia Freitas, Eckhard Bick, and Valeria de Paiva. 2017. Universal Dependencies for Portuguese. In *Proceedings of the Fourth International Conference on Dependency Linguistics (Depling 2017)*, pages 197–206, Pisa, Italy. Linköping University Electronic Press.

Jânia Ramos. 1992. *Marcação de caso e mudança sintática no português do Brasil: uma abordagem gerativa e variacionista*. Ph.d. thesis, Universidade Estadual de Campinas (Unicamp), Instituto de Estudos da Linguagem.

Scott Schwenter. 2014. Two kinds of differential object marking in Portuguese and Spanish. *Issues in Hispanic and Lusophone linguistics*, 1:237–260.

Emil Stenström. 2016. conllu.

Marcos Zampieri and Martin Becker. 2013. Colonia: Corpus of Historical Portuguese. In *ZSM Studien, Special Volume on Non-Standard Data Sources in Corpus-Based Research*, volume 5. Shaker.