



UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA

Autorizada pelo Decreto Federal nº 77.496 de 27/04/76



Pró-Reitoria de Pesquisa e Pós-Graduação

COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2024

O USO DE CORPORA ELETRÔNICOS ANOTADOS PARA O MAPEAMENTO DA SINTAXE DO PORTUGUÊS BRASILEIRO

Rayana Cecília Machado de Jesus¹; Antonio José Maria Codina Bobia²

1. Bolsista – Modalidade Bolsa/PVIC, Graduando em Licenciatura em Letras – Língua

Portuguesa, Universidade Estadual de Feira de Santana, e-mail:

rayanadjess63@gmail.com

2. Orientador, Departamento de Letras e Artes, Universidade Estadual de Feira de

Santana, e-mail: ajmcbobia@uefs.br

PALAVRAS-CHAVE: Edição eletrônica de textos; Corpora linguísticos; Anotação morfossintática.

INTRODUÇÃO

Para o fomento dos estudos linguísticos em sintaxe diacrônica do português brasileiro, as ferramentas computacionais e de anotação morfossintática foram desenvolvidas de modo a melhor extrair dados de corpora eletrônicos, para otimizar as buscas de dados. Nesse âmbito de pesquisa das humanidades digitais, este trabalho investiga, em 10 cartas do século XVII e XVIII pertencentes à Coleção de Documentos Históricos do Arquivo Histórico Municipal de Salvador (AHMS), fundo Câmara Municipal: *Série Cartas da Casa de Audiência/Conselho da Câmara de Vereança/Senado da Câmara/Câmara Municipal de Salvador* (1625-1700), do Corpus Eletrônico de Documentos Históricos do Sertão (CE-DOHS) (Carneiro; Lacerda, 2023), a ocorrência de orações relativas.

Para isso, foi realizada, com o auxílio do e-Dictor (Faria; Kepler; Paixão de Sousa, 2013), a edição dessas cartas, gerando a edição eletrônica diplomática, semidiplomática e modernizada; a anotação morfossintática das cartas; além de um levantamento, através da ferramenta de busca (<https://www.tycho.iel.unicamp.br/search>). A ferramenta digital utilizada na pesquisa está integrada à Plataforma Tycho Brahe (Veronesi; Galves, 2024). Essa plataforma conta com textos anotados sintaticamente, facilitando o processo de análise linguística e, à época, era parceira do CE-DOHS (Carneiro; Lacerda, 2023).

A edição de documentos aliado ao processo de anotação morfológica e sintática possibilita que se busquem aspectos idiossincráticos importantes do português brasileiro,

proporcionando uma compreensão ainda mais estreita de questões linguísticas e sócio-históricas do Brasil, além de manutenção de seu acervo cultural.

Sendo a replicabilidade um dos princípios do fazer científico, esta pesquisa se justifica por dar continuidade a estudos que se comprometem a mapear a sintaxe do português brasileiro. Além disso, tal princípio permite que cada vez mais dados históricos e linguísticos possam ser proporcionados a quem se interessa por essas informações acessivelmente. Fazendo isso de maneira rápida e replicável, mais investigações podem ser desenvolvidas de modo responder indagações linguísticas e/ou sócio-históricas.

O fenômeno escolhido para ser pesquisado são as orações relativas. Os estudos de seu comportamento vêm sendo feito por pesquisadores de sintaxe diacrônica. Sua análise permitirá, além de uma compreensão mais aproximada da mudança linguística, o fomento de futuras pesquisas comparativas ou exponencialmente específicas. Dessa maneira, pesquisadores em sintaxe diacrônica chegaram à hipótese de que o português clássico passou por séries de modificações de parâmetros (Galves; Britto; Paixão de Sousa, 2005; Galves; Sândalo, 2012; Galves; Paixão de Sousa, 2017) talvez devido ao contato com as línguas indígenas e africanas. Através de estudos comparativos entre as estruturas sintáticas do PB e do PE, linguistas têm buscado explicações para as divergências entre essas línguas. Para que esse trabalho fosse possível, a Linguística de Corpus tornou-se imprescindível.

METODOLOGIA

Em primeiro lugar, selecionaram-se dez cartas que integram a *Coleção de Documentos Históricos do Arquivo Municipal de Salvador (AHMS), fundo Câmara Municipal: Série Cartas do Senado* do CE-DOHS. Logo, foi iniciada a edição eletrônica das cartas com o auxílio do e-Dictor. Esse processo contou com o auxílio de algumas produções bibliográficas para realizar a expansão de abreviaturas e a modernização das palavras: o dicionário de abreviaturas de manuscritos dos séculos XVI ao XIX (Flexor; 2008) e a dissertação de mestrado “Abreviaturas: uso e função nos manuscritos” (Sobral, 2018).

Figura 01: Tela do e-Dictor com manuscrito e edição semidiplomática



The screenshot shows the e-Dictor software interface. At the top, there is a blue header with the text 'Plataforma Tycho Brahe: Viewer' and 'Corpus CE-DOHS'. Below the header, the document identifier 'documento: 02-CDHAMS-17-05-1753.xml' is displayed. The main area is divided into two sections: a large window on the left showing a handwritten manuscript in cursive Portuguese, and a smaller window on the right showing the digital transcription of the text. The transcription is in a modern, printed-style font. The right window also includes some administrative information and a preview of the document.

1 de 4

Carta 02

Documento contendo 04 fólios .

AHU-Bahia , caixa 120 documento 27 3 .
AHU_ACL CU_005 , Caixa 111 , Documento 8691 .

REQUERIMENTO do Eretor da Fábrica dos Atanados da cidade da Bahia , Tomás Veloso Rebelo ao rei [D. José] a pedir provisão para que os administradores e feitores da fábrica prefiram para tanto a arrematação dos talhos por tempo de nove anos ou o que durar durante

Esta camara como administradora do bem commun , está em posse de mandar rematar os talhos , e curraes donde Je corta a carne pa , o sustento do povo desta cide , os quaes se poem empraca e se remataõ nella a quem por elles lança , deq ma is deq resulta augmento , eutildade de , a o Conco , e a Real Fazenda de Vofsa Maestade . E assentando nesta Cide

Depois da edição, foi realizada a anotação morfossintática automática. Em seguida, a anotação foi corrigida, visto que, ainda que “a automatização dos processos vis[e] diminuir os erros decorrentes dos processos de anotação” (Corôa, 2022, p. 166), por conta de ajustes a serem feitos em sua formatação, a ferramenta pode não desempenhar uma performance sem erros.

Figura 02: Tela do e-Dictor com a etiquetagem morfossintática

The screenshot shows the e-Dictor application interface. On the left, there is a handwritten historical document in Portuguese. On the right, the text is displayed with morphosyntactic annotations. Below the text, a table shows the original text and its corresponding POS tags. The POS tags are as follows:

original	Esta	câmara	como	administradora	do	bem	comum	está	em	posse	de	mandar	rematar	os	talhos	e	cun		
POS tag	D-F	N	C	VB-P-3S	P+D	N	ADJ-G	PUNC	ET-P-3S	P	N	P	VB	VB	D-P	N-P	PUNC	CONJ	N

Por fim, foi realizado um levantamento das orações relativas presentes no corpus anotado. Para recuperar essa informação, buscou-se pela etiqueta WPRO, que retorna todas as ocorrências de pronomes relativos.

RESULTADOS E/OU DISCUSSÃO

O primeiro resultado foi a ampliação dos corpora anotados do português brasileiro. Com efeito, ambos os corpora linguísticos, o CE-DOHS e o Tycho Brahe, foram expandidos com as 10 cartas do século XVII e XVIII (7612 palavras), devidamente anotadas morfossintaticamente para análise. Em relação às orações relativas, encontraram-se 105 casos. Uma análise preliminar foi efetuada, visto que para a análise definitiva precisa-se fazer a anotação sintática do corpus (Fase II do projeto). Nessa primeira análise, não se verificaram estratégias de relativização divergentes da forma canônica, o que confirmaria os achados de Tarallo (1993).

CONSIDERAÇÕES FINAIS

Entre os estudos linguísticos, o nível sintático destaca-se por apresentar uma maior dificuldade de identificação de mudança linguística devido à sua variação mais lenta e menos perceptível (Corôa, 2022, p. 166). Por esse motivo, a identificação das orações relativas por meio da anotação automática foi analisada pelo estudo comparativo com outras pesquisas voltadas à mudança na sintaxe diacrônica. Aspectos sócio-históricos não deixarão de ser considerados, como bem explicitado por Mattos e Silva (2004), que explana a importância do estudo de documentos datados e localizados para o estudo diacrônico. Os resultados das análises das estratégias de relativização não foram

concluintes. No entanto, espera-se que na segunda fase da pesquisa mais dados possam ser analisados para chegar a uma conclusão mais categórica.

REFERÊNCIAS

- [1] CARNEIRO, Zenaide de Oliveira Novais; LACERDA, Mariana Fagundes de Oliveira (org.). Plataforma CE-DOHS - Corpus Eletrônico de Documentos Históricos do Sertão. Disponível em: <http://www.ufes.br/cedohs>. Acesso em: 20 out. 2023.
- [2] CORÔA, W. A sintaxe da ordem no corpus de cartas e atas produzidas por homens bons da Câmara Municipal de Salvador. Tese de doutorado. Universidade Estadual de Campinas. 2022.
- [3] FLEXOR, Maria Helena Ochi. Abreviaturas: manuscritos dos séculos XVI ao XIX. 3. ed. Rio de Janeiro: Arquivo Nacional, 2008.
- [4] GALVES, Charlotte; BRITTO, Helena; PAIXÃO DE SOUSA, Maria Clara. The change in clitic placement from Classical to Modern European Portuguese: results from the Tycho Brahe Corpus. *Journal of Portuguese Linguistics*, Lisboa, v. 4, n.1, p. 39-68, 2005.
- [5] GALVES, Charlotte; SÂNDALO, Filomena. From Intonational Phrase to Syntactic Phase: the grammaticalization of enclisis in the history of Portuguese, *Lingua*, 122(8), 952-974, 2012.
- [6] GALVES, Charlotte; PAIXÃO DE SOUSA, Maria Clara. The Position of the Verb in the History of Portuguese: Subject Position, Clitic Placement and Prosody, *Language*, 93(3): p. 152–80, 2017.
- [7] MATTOS E SILVA, Rosa Virgínia. *Ensaios para uma sócio-história do português brasileiro*. São Paulo: Parábola Editorial, 2004.
- [8] PAIXÃO DE SOUSA, M. C.; KEPLER, F. N.; FARIA, P. P. F. E-Dictor: Novas perspectivas na codificação e edição de corpora de textos históricos. In: SHEPHERD T.; SARDINHA T. B.; PINTO M. V. (Org.). *Caminhos da linguística de corpus*. Campinas: Mercado de Letras, 2010.
- [9] SOBRAL, M. das G. T. Abreviaturas: uso e função nos manuscritos. Dissertação (Mestrado em Letras e Linguística) – Instituto de Letras, Universidade Federal da Bahia, Salvador, 2007. Disponível em: <https://repositorio.ufba.br/ri/handle/ri/28200>. Acesso em maio. 2024.
- [10] TARALLO, F. Diagnosticando uma gramática brasileira: o português d'aquém e d'além-mar ao final do século XIX. In: ROBERTS, I. & KATO, M. (orgs.) *Português Brasileiro: uma viagem diacrônica*. 2 ed. Campinas: Editora da Unicamp: 1993, p. 69-106.
- [11] VERONESI, L. H. L.; GALVES, Charlotte. The Tycho Brahe Platform. 2024.