http://dx.doi.org/10.31513/linguistica.2021.v17n3a49621

Reconstruindo a história do Português do Brasil pelo *Corpus* Tycho Brahe Brasil: novos dados, novos olhares[1]

Rebuilding the history of Brazilian Portuguese through Tycho Brahe Brazil Corpus: new data, new perspectives

Paulo Ângelo Araújo-Adriano[2]

Williane Silva Corôa[3]

RESUMO

Este artigo apresenta novos dados para a investigação histórica do português brasileiro, a serem introduzidos no *Corpus* Histórico Anotado do Português Tycho Brahe Brasil (CTBB), que passa a conter cartas e atas de Homens Bons da Câmara Municipal de Salvador, peças de teatro de comédia, grande parte do Rio de Janeiro, e cantiga. Como exemplificação do CTBB, o presente texto apresenta uma investigação sobre a colocação clítica em contexto V1 e sobre a perífrase progressiva no português brasileiro, pautando-se não apenas na data de nascimento dos autores, mas também na data de publicação dos textos. Discutimos a questão, grandemente debatida, da emergência do português brasileiro como já existente em meados do século XVIII, conforme defendem Ribeiro (1998) e Corôa (2021). Além disso, argumentamos em favor da importância de se fazer linguística histórica contrastando não só a data de nascimento e a de publicação, mas, também, tipos textuais diferentes.

PALAVRAS-CHAVE: Linguística de *corpus*. Metodologia em linguística diacrônica. Emergência do português brasileiro. Clíticos. Estrutura progressiva.

ABSTRACT

This paper presents new data for the historical investigation in Brazilian Portuguese, to be available on the Brazilian Tycho Brahe Parsed *Corpus* of Historical Portuguese (CTBB), which will contain letters and minutes from Good Men from Salvador City Council and comedy plays, essentially from Rio de Janeiro. As an exemplification of this new set of data, this paper presents an investigation on clitic placement in V1 context and on the morphology of progressive periphrasis in Brazilian Portuguese, based not only on the authors' date of birth but also on the publication date of the texts. We discuss the issue of the emergence of Brazilian Portuguese, widely debated, as already existing in the mid-18th century, as defended by Ribeiro (1998) and Corôa (2021). Furthermore, we argue in favour of the importance of research in historical linguistics by comparing not only birth and publication date but also different textual types.

KEYWORDS: *Corpus* linguistics. Methodology in diachronic linguistics. The emergence of Brazilian Portuguese. Clitics. Progressive structure.

---

[1] Agradecemos aos dois pareceristas anônimos pelas observações e sugestões a este artigo. Ainda que nem todas puderam ser contempladas neste texto, elas serão consideradas em trabalhos futuros.

[2] O presente trabalho foi realizado com apoio da Fapesp, Processo 2019/17443-9. Doutorando da Unicamp, pauloangeloaa@gmail.com, https://orcid.org/0000-0002-9884-0723.

[3] O presente trabalho foi realizado com apoio da Fapesp, Processo 2017/16581-3. Doutoranda da Unicamp, williscoroa@gmail.com, https://orcid.org/0000-0002-5887-7281.

Recebido em 29/01/2022 - Aceito em 19/05/2022

Introdução

Neste artigo apresentamos novos elementos para reconstruir a história do português do Brasil (PB) pelo *Corpus* Histórico Anotado do Português Tycho Brahe-Brasil (CTB-Brasil). Partimos de dois *corpora* que integram esse novo *corpus*: o primeiro constituído por *Cartas da Câmara Municipal de Salvador*, produzidas por brasileiros que ocupavam o cargo de escrivão ao longo dos séculos XVII e XVIII; e o segundo, constituído de texto teatrais e cantigas, também produzidos por brasileiros a partir do século XVII. Analisamos os padrões de colocação de clíticos e a perífrase progressiva.

Apesar de estarem disponíveis diversos *corpora* do português brasileiro sincrônico, como NILC – Núcleo Interinstitucional de Linguística Computacional – (USP, UFSCAR, UNESP) e NURC – Norma Urbana Culta –, para citar alguns, há poucos *corpora* que concentram textos históricos brasileiros. Dos mais notórios há apenas dois, o *corpus* do PHPB – Para a história do português brasileiro – (UFPE, UFPBA, UFMG, UFRJ, UFSC, UFPB, USP) e o *Corpus* Tycho Brahe – CTB – (Unicamp). Muito do que se sabe da história do PB atualmente certamente é devido ao extenso trabalho que pesquisadores do Brasil desenvolveram tendo em mãos o PHPB. No entanto, as fontes históricas de tal projeto não estão anotadas nem morfológica tampouco sintaticamente, além de não estarem disponíveis ao público, o que impede que o linguista histórico trabalhe de forma mais rápida, prática, dada a ausência de anotação, e, o mais importante, que o linguista replique a pesquisa desenvolvida – padrões intrínsecos das ciências. Para além dessas questões, quando não se trabalha com *corpora* anotados, os dados acabam sendo colhidos à mão, o que pode dar margem para falhas e lapsos, resultados indesejados em qualquer campo da ciência. Por esse motivo, defendemos a ideia de que é necessário trabalhar com *corpora* históricos anotados morfológica e sintaticamente, tal como o *Corpus* Tycho Brahe.

Adicionalmente, faz-se relevante estar atento às idiossincrasias ligadas às tipologias textuais que integram tais *corpora*, pois cada tipo textual pode favorecer uma ou outra estrutura dependendo dos seus objetivos e do seu formato. Apenas com *corpora* de diferentes tipos textuais tais nuances aparecem, daí a necessidade de se ter em mãos *corpora* grandes, não só em termos de número de palavras, mas também em relação à diversidade de tipos textuais, de modo que a língua representada por aquele recorte não seja enviesada por um ou outro tipo textual.

Ainda, faz parte do processo metodológico do linguista histórico decidir se o recorte temporal a ser analisado irá se basear na data de nascimento do autor ou na data de publicação do texto. Entretanto, uma escolha não deveria excluir a outra, na medida em que, a depender do tipo textual investigado, a data de publicação pode representar melhor a gramática de um período, e o contrário sendo igualmente verdade. Assim, também defendemos ser necessário considerar tanto a data de nascimento dos autores quanto a data de publicação dos textos, atentando às especificidades de cada tipo textual.

De modo a explicitar tal posicionamento, este texto está organizado da maneira como se segue. Em §2, discutimos alguns desafios do labor do linguista histórico, destacando alguns pontos metodológicos, como escolha do gênero e da datação das obras, se pelo nascimento do autor, se

pela publicação da obra. Em §3, apresentamos os novos dados que passam a compor o CTB, que compreendem cartas e atas de *Homens Bons*[4] da Câmara municipal de Salvador e representações artísticas. Já em §4, analisamos dois fenômenos para ilustrar como esses novos dados podem trazer novos olhares para compreendermos a história do português do Brasil. Por fim, em §5 fazemos algumas considerações finais.

1. Desafios de se trabalhar com linguística histórica/diacrônica

O labor do historiador da língua assenta-se, essencialmente, em fontes escritas que resistiram às intempéries do tempo. Nesse sentido, as informações sobre *quando*, *como*, *onde* e *por quem* o texto foi escrito são indispensáveis para melhor delinear como a mudança linguística acontece (MATTOS E SILVA, 1996). Porém, nos textos históricos, nem sempre é possível acessar tais informações, o que torna a vida do pesquisador em linguística histórica bastante complicada.

Com vistas a sanar esse problema, nos últimos anos, o número de projetos de investigação que visavam à constituição de *corpora* históricos, tanto no Brasil quanto no exterior, deu um salto (JENSET; MCGILLIVRAY, 2017). Isso se justifica, pois os *corpora* são elementos essenciais da linguística histórica, cuja utilização com fins de investigação permite a coleta e a exploração de um conjunto de dados úteis para descrição e análise linguística de forma mais rápida, prática, confiável e, o mais importante, replicável.

Para um pesquisador de linguística histórica, o estudo de qualquer fenômeno linguístico deve fundamentalmente basear-se em um *corpus*. Um *corpus* histórico pode ser definido como um conjunto de exemplos de linguagem natural – desde frases até um conjunto de textos escritos em estágios pretéritos de uma língua (HUNSTON, 2002). Deve, portanto, ser tão extenso quanto possível, visto que o historiador da língua conta apenas com a documentação histórica remanescente. Além disso, deve, quando possível – cientes das dificuldades inerentes – ser representativo considerando as diferentes variedades da língua e os diferentes tipos textuais, a fim de que se possa fazer generalizações. Uma vez que se é impossível investigar todo material linguístico existente de uma língua, o que seria a única maneira de ter um *corpus* realmente representativo, o trabalho do linguista histórico é, de certa sorte, limitado (cf. GARCÍA GARCÍA, 2002, p. 121). Dada essa limitação, o único método para que o objeto de investigação – o *corpus* – seja o mais representativo o possível, é pela variação de tipos textuais.

No que diz respeito à variabilidade do tipo textual, a eleição do *corpus* está intimamente atrelada à escolha do tipo textual e ao recorte histórico a ser investigado. No bojo dessa discussão, "a inexistência de um mesmo tipo de texto em todas as fases da história do Português" (CAMBRAIA, 1994, p. 11) faz com que o historiador da língua manuseie tipologias textuais diferentes, o que pode impactar nos resultados alcançados, visto que tais resultados se mostram sensíveis à seleção do tipo textual. Um exemplo disso aparece nas monografias organizadas por Cyrino e Torres Morais (2018).

[4] O termo homens bons era utilizado para designar os membros da nobreza local, elegíveis às Câmaras municipais. Os requisitos para ser considerado Homem bom eram: ser maior de 25 anos, casado ou emancipado, católico, e sem nenhuma "impureza de sangue".

Galves (2018, p. 442), no *posfácio* do livro, aponta para a influência do tipo textual sobre os resultados encontrados, pois tipologias textuais diversas apresentaram pesos sócio-históricos distintos. Desse modo, as especificidades de cada tipo devem ser levadas em consideração pelo linguista histórico ao constituir seu *corpus*[5].

Outro ponto a ser levado em consideração refere-se a qual momento histórico adotar, se a data de nascimento do autor, se a data de publicação do texto. Dentro do arcabouço da Gramática Gerativa, assume-se que a gramática de um falante é estabelecida no período de aquisição da linguagem, por volta dos 6 anos de idade; adicionalmente, assume-se que a aquisição e a mudança estão intrinsecamente atreladas (cf. LIGHTFOOT, 1979; entre outros). Sob esse ponto de vista, a mudança linguística seria resultado da interação entre a capacidade inata do indivíduo de adquirir uma língua e a experiência linguística vivenciada pelas sucessivas gerações de falantes. Se aquisição e mudança se entrelaçam, é patente adotar a data de nascimento do autor como critério de datação dos textos, isto é, estaríamos diante da *geração biológica* – um termo adotado de Paixão de Sousa (2004).

Por outro lado, não é tão claro assim que esse deva ser o caso para textos históricos que foram escritos com o objetivo de interpretar uma narrativa para um público específico, como é o caso de peças teatrais. Uma vez que autores de teatro tentam representar na fala de seus personagens a linguagem corriqueira do seu público-alvo, parece-nos mais adequada adotar a data de publicação do texto, não a data de nascimento do seu escrevente, a fim de investigarmos uma *geração histórica* – outro termo adotado de Paixão de Sousa (2004).

Assim, embora a gramática do autor da peça tenha sido fixada na sua infância, no período de aquisição, a gramática que ele utiliza, idealmente, nas suas obras teatrais é aquela representativa do contexto de sua criação, que depende do seu público-alvo. Isso pode ser evidenciado no comentário/advertência de Artur Azevedo (1900), autor de peças brasileiras de teatro do século XIX, sobre a tentativa de suas peças agradarem e serem destinadas a todo tipo de público-alvo, mesmo com certa resistência da Companhia Lucinda Simões, responsável à época por promover espetáculos destinados apenas à elite econômica (cf. NEVES, 2006, p. 21):

> Esses espetáculos podem ser freqüentados, necessariamente, por todo aquele ou aquela que comprar o seu bilhete e esteja trajado, ou trajada, com certa decência; mas a empresa destina-os especialmente "às mais distintas famílias da elite da nossa sociedade", e conta que o seu teatro seja, às quartas-feiras, um ponto de reunião para as damas e os cavalheiros do monde, como dizem os franceses, ou do high life, como dizem os ingleses. A tentativa é inteligente e simpática, mesmo porque talvez consiga fazer as pazes entre a boa sociedade e o teatro, que há muito se desavieram.

[5] Cambraia (1994) chama a atenção para essa questão tanto no que tange à oposição (i) literário versus não literário, quanto no que tange à oposição (ii) prosa versus verso. Com relação à (i), Cambraia aponta a dificuldade em diferenciar a prosa epistolar literária e as correspondências comuns, dado que não há uma clara fronteira entre o que seria ou não literário. Com relação à (ii), Cambraia (1994) exemplifica a questão com os textos de teatro e pontua que se um pesquisador desejar utilizá-los como fonte de investigação deve considerar que, durante um tempo, tais textos eram escritos em verso. Assim, uma mesma tipologia teria o inconveniente de misturar textos de diferente natureza.

Fica patente, portanto, que o autor, a despeito da pressão econômica, representava nas suas peças todo tipo de público, provavelmente adequando e aproximando a linguagem do seu teatro à linguagem da sociedade.

Nota-se que autores teatrais não ignoram solenemente a fala do seu público; ao contrário, precisam estar atentos à linguagem da sua sincronia para que suas obras não soem artificiais. Um exemplo dessa preocupação é o Primeiro Congresso Brasileiro de Língua Falada no Teatro na Universidade da Bahia em 1956, para discutir exatamente questões relativas à língua na dramaturgia teatral. É claro que, muito embora a escrita teatral não seja totalmente verossímil à fala da época representada, o trabalho do linguista histórico "garimpeiro" – para usar um termo de Barbosa (1998) – é sempre desnudar a língua oral do passado, através da escrita, "como uma *tentativa* de aproximação da realidade" (LOBO, 1998, p. 179, grifo nosso). Dessa maneira, parece razoável afirmar que manusear a data de publicação de um texto também é exitoso, quando se quer investigar o conhecimento linguístico de falantes antepassados.

Olhar para a *geração biológica* ou para a *geração histórica* impacta, portanto, a escolha do *corpus* a ser adotado. A impossibilidade de identificar *por quem* o texto foi escrito tem consequências óbvias sobre a eleição da data de nascimento como critério de datação dos textos. Optar por um ou por outro critério também repercute sobre a periodização, pois olhar para a *geração biológica* implica observar a mudança precocemente se comparado com estudos que consideram apenas a data de produção dos textos, isto é, a *geração histórica*. Não é nosso objetivo com essa discussão bater o martelo e definir qual a melhor escolha, muito menos condenar outra; ao contrário o objetivo é problematizar e salientar que tal escolha não deve ser ingênua, mas consciente, dadas as implicações.

Além dessas questões, é indispensável que quaisquer *corpora* sejam armazenados eletronicamente e possam ser pesquisados por programas de computador. Também urge que os *corpora* sejam anotados, dado que a análise feita manualmente abre grande margem para a possibilidade de erros e até de vieses tendenciosos. Assim, os *corpora* anotados são vantajosos, pois o processo de coleta e classificação dos dados, além de contar com uma fonte de evidências muito mais ampla do que se o levantamento for feito manualmente, é mais preciso, mais transparente do ponto de vista metodológico e facilita a reprodutibilidade dos dados investigados (JENSET; MCGILLIVRAY, 2017).

Dessa maneira, o processo de anotação promove a generalização dos dados para um conjunto de dados maior, possibilitando a replicabilidade – princípio intrínseco ao fazer científico –, o que é impossível caso o levantamento dos dados seja feito manualmente. Portanto, a anotação de *corpus* é uma etapa essencial no processo de investigação linguística e que contribui para testar empiricamente hipóteses a partir dos dados.

No Brasil, apesar da grande tradição de constituição de *corpora* históricos, contamos com poucos *corpora* anotados. Um deles é *Corpus* Tycho Brahe, primeiro *corpus* eletrônico anotado em língua portuguesa. Na próxima seção, apresentamos o *Corpus* Tycho Brahe, com dados majoritariamente do português europeu (PE), e a nova vertente na qual se está repaginando: o *Corpus* Tycho Brahe Brasil.

2. O *Corpus* Tycho Brahe: velhos e novos dados

O *Corpus* Histórico do Português Tycho Brahe é um *corpus* eletrônico anotado (cf. GALVES; FARIA, 2010; GALVES *et al.*, 2017), composto de textos majoritariamente em português europeu, escritos por autores nascidos entre o fim do século XIV e o fim do século XIX. Atualmente, 88 textos (3.544.628 palavras) estão disponíveis para pesquisa, com um sistema de anotação morfológica (aplicada em 58 textos, com um total de 2.280.819 palavras) e sintática (aplicada em 27 textos, com um total de 1.234.323 palavras).

A comunidade científica muito ganhou com esse *corpus* considerado médio-grande[6], que serviu de base para diversos estudos sobre a ordem do sujeito, sistema V2, interpolação, ordem relativa clítico-verbo, o fenômeno do sujeito nulo[7] etc., dando novos subsídios para o entendimento da emergência do português europeu moderno.

Como já mencionado na introdução, *corpora* anotados seja morfologicamente seja sintaticamente possibilitam o manuseio de uma quantidade grande de dados em pouco tempo. Até o momento, no *Corpus* Tycho Brahe há apenas 11 textos do português brasileiro. Entretanto, é imprescindível ampliar o conjunto de dados para se investigar rigorosamente o passado de uma língua: apenas com uma diversidade de gêneros e de dados é que é possível delinear acuradamente a emergência do português falado no Brasil, e até mesmo conjugando aspectos linguísticos com aspectos sócio-históricos. Galves (2018, p. 456), discorrendo sobre o que vai nos permitir fazer descrições mais precisas sobre o português brasileiro, aponta que "serão necessários, portanto, dados, novos dados, muito mais dados". Na próxima seção, apresentamos esses novos dados que passam a fazer parte da nova roupagem do *Corpus* Tycho Brahe: o *Corpus* Tycho Brahe Brasil.

2.1. O *Corpus* Tycho Brahe (Brasil): novos dados.

Mais recentemente o projeto, liderado por Charlotte Galves (Unicamp), propõe-se a ampliar a quantidade de textos brasileiros com a criação do *Corpus* Tycho Brahe-Brasil (CTB-Brasil). O CTB-Brasil será constituído de textos escritos no Brasil e por brasileiros entre o século XVI e o século XX. O objetivo é retratar a dinâmica da emergência do português brasileiro e melhor compreender a separação entre a variedade clássica, europeia e brasileira do português. Dentre os textos que constituirão o CTB-Brasil estão o *corpus* de textos de teatro e cantiga e o *corpus* de Cartas e Atas da Câmara Municipal de Salvador. A seguir, ambos são detalhados.

---

[6] Berber Sardinha (2004, p. 26) considera a seguinte classificação em relação ao tamanho de um *corpus*, relativo ao número de palavras: pequeno (menos de 80 mil), pequeno-médio (80 a 250 mil), médio (250 mil a 1 milhão), médio-grande (1 milhão a 10 milhões) e grande (10 milhões ou mais).

[7] Para ter acesso à lista de trabalhos realizados a partir do *Corpus* Tycho Brahe acesse o link: http://www.tycho.iel.unicamp.br/~tycho/pesquisa/.

2.2. *Corpus* de textos de teatro

A fim de possibilitar novos olhares para o português brasileiro, 11 peças de teatro publicadas entre o século XVIII e o século XXI e 1 cantiga do século XVIII foram incluídas no *corpus*. Os textos anotados morfologicamente (233.183 palavras), fruto do projeto de tese FAPESP (Processo 2019/17443-9), estão todos disponíveis no site do *Corpus*[8], com as informações básicas expostas no quadro 1, a seguir.

Quadro 1: Informações das peças que compuseram o *corpus*

Nascimento	Autor	Nome da Peça	Publicação do texto	Número de palavras	Tipo Textual
1705/séc. 18	Antônio José da Silva	Guerra do Alecrim e da Manjerona	1737/séc. 18	27.224	Comédia teatral
1740/séc. 18	Domingos Caldas Barbosa	Viola de Lereno	1798/séc. 18	18.445	Cantiga
1815/séc. 19	Martins Pena	O juiz de paz na roça	1833/séc. 19	6.897	Comédia teatral
1815/séc. 19	Martins Pena	O noviço	1845/séc. 19	17.563	Comédia teatral
1829/séc. 19	José de Alencar	O demônio familiar	1857/séc. 19	25.319	Comédia teatral
1838/séc. 19	França Junior	Caiu o ministério	1883/séc. 19	14.629	Comédia teatral
1855/séc. 19	Artur de Azevedo	O tribofe	1891/séc. 19	22.434	Comédia teatral
1880/séc. 19	Gastão Tojeiro	Onde canta o sabiá	1920/séc. 20	27.495	Comédia teatral
1934/séc. 20	Gianfrancesco Guarnieri	Eles não usam black-tie	1957/séc. 20	22.571	Comédia teatral
1956/séc. 20	Miguel Falabella	A partilha	1990/séc. 20	17.531	Comédia teatral
1978/séc. 20	Paulo Gustavo	Minha mãe é uma peça	2006/séc. 21	33.075	Comédia teatral
1966/séc. 20	Paulo Sacaldassy	Fulana, Sicrana e Beltrana	2007/séc. 21	8.548	Comédia teatral

A escolha pela tipologia peça de teatro não foi ingênua. Dada a abordagem teórica assumida, a da Gramática Gerativa, a pergunta de pesquisa norteadora leva em consideração o conhecimento interno do falante. Entretanto, quando se faz linguística histórica, na ausência de uma gramática interna viva, é impossível acessar o conhecimento de falantes de outras épocas, cuja gramática já não pode fornecer julgamento de gramaticalidade/aceitabilidade.

Dentre os textos de teatro, o gênero comédia teatral foi o escolhido visto que a comédia é um gênero sobre o cotidiano de pessoas e, em geral, são retratados indivíduos de distintas classes sociais, de modo que possibilita que vejamos retratados nesses textos diferentes variedades da língua da época, inclusive aquelas que não estariam sob pressões normativas. Dessa maneira, concordamos com a afirmação de que "as peças de teatro constituem importante material de pesquisa quando se quer tentar uma aproximação com a fala de sincronia passadas" (DUARTE, 2012, p. 19) e, portanto, defendemos que peças de teatro também são fontes confiáveis e plausíveis de se acessar essa gramática interna de outras sincronias.

[8] http://www.tycho.iel.unicamp.br/~tycho/corpus/en/catalogo.html

No que se refere às obras ineditamente inseridas no *Corpus* Tycho Brahe, a quantidade de textos para cada século foi determinada levando em consideração o início, o meio e o fim do século em questão. Por exemplo, para representar o século XX, uma peça de 1920, uma em torno do meio do século, de 1957, e uma de 1990, na virada do século. Para o século XIX, foram incluídas mais peças, pois é o século cujo número de peças disponíveis é maior. Todos os textos foram coletados levando em consideração a nacionalidade do autor. Na medida do possível, foram selecionadas peças de autores de vários locais do Brasil, porém a maioria é oriunda do Rio de Janeiro, dada a concentração artística e populacional do país nesse estado (cf. SILVA NETO, 1977).

Na próxima seção, vai ser apresentado o conjunto de cartas e atas da Câmara Municipal, escritas nos séculos XVII e XVIII por brasileiros, que também passam a compor o *Corpus* Tycho Brahe Brasil.

2.3. Cartas e atas de Homens Bons da Câmara Municipal de Salvador

As cartas e atas que compõem o *corpus* elaborado ao longo do projeto de tese (Processo Fapesp: 2017/16581-3) fazem parte do Fundo da Câmara Municipal de Salvador, composto por documentos provenientes da administração municipal. A vasta documentação da Câmara de Salvador encontra-se sob a guarda do Arquivo Histórico Municipal, inaugurado em 10 de abril de 1932, administrado atualmente pela Fundação Gregório de Matos.

As cartas e atas, ao lado de alvarás, posturas régias, requerimentos da população, eram os principais meios de comunicação entre a colônia e a metrópole. A responsabilidade de escrever os documentos era do escrivão. Na Câmara da Bahia, no século XVI, as principais funções eram as de escrivão, almotacé de execuções e tesoureiro, mas apenas o cargo de escrivão era exercido por "homens das letras": para exercê-lo, os indivíduos deveriam ser nomeados por provisão real, ter domínio da escrita e conhecimento da legislação portuguesa.

Quadro 2: Informações das cartas e atas que compuseram o *corpus*

Nascimento	Autor	Publicação do texto	Tipo Textual
1602/séc. 17	Rui de Carvalho Pinheiro	1650/séc. 17	Carta
1630/séc. 18	Domingos Garcia de Aragão	1653/séc. 17	Carta
1650/séc. 17	Pedro Dias Pereira	1700/séc. 18	Carta
1670/séc. 17	João de Couros Carneiro Filho	1695/séc. 17	Carta
1670/séc. 17	Manuel Silveira de Magalhães	1715/séc. 18	Carta
1672/séc. 17	Manuel Pessoa de Vasconcelos	1699/séc. 17	Carta
1672/séc. 17	Manuel Pessoa de Vasconcelos	1715/séc. 18	Carta
1700/séc. 18	João de Couros Carneiro Neto	1725/séc. 18	Carta
1710/séc. 18	Jerônimo Sodré Pereira	1765/séc. 18	Ata
1720/séc. 18	Joaquim Rodrigues da Silveira	1770/séc. 18	Ata
1720/séc. 18	Manuel José de Azevedo	1768/séc. 18	Ata
1721/séc. 18	José Álvaro Pereira Sodré	1768/séc. 18	Ata
1725/séc. 18	João Duarte Silva	1775/séc. 18	Ata

O conjunto de cartas e atas faz parte da Série *Documentos Históricos do Arquivo Histórico Municipal* que foi publicada em 1949, na ocasião da comemoração dos 400 anos da Cidade de Salvador. As *Cartas do Senado a Sua Magestade* reúnem, em seis volumes, as cartas produzidas pela câmara municipal entre 1638 e 1730. Já o *Livro de Atas da Câmara* reúne, em dez volumes, as atas produzidas entre 1625 e 1765. Para a tese em elaboração foram anotados todos os volumes de cartas e os volumes 10 e 11 de atas, o que totaliza 269.777 palavras.

Na próxima seção, exemplificamos a praticidade de se fazer linguística histórica com o morfológica e sintaticamente anotado *Corpus* Tycho Brahe Brasil, a partir dos novos dados supramencionados.

3. O *Corpus* Tycho Brahe: novos olhares.

A partir do conjunto de dados organizados e elaborados, **já é possível delinear algumas questões** relativas à emergência do PB e à dinâmica da mudança. Para demonstrá-las, escolhemos dois fenômenos característicos do PB a fim de investigar quando esses fenômenos aparecem nos dados históricos: (i) a estrutura progressiva e a (ii) colocação de clíticos.

Uma característica bastante peculiar na história do PB é a colocação de clíticos, que, apesar de ser um fenômeno bastante estudado (cf. PAGOTTO, 1992; CARNEIRO, 2005; MARTINS, 2009), ainda guarda algumas questões sem resposta. Uma delas é saber se a colocação de clíticos no PB mudou em relação ao PE ou em relação a fases mais pretéritas do português, visto que o PB e o PE são claramente distintos atualmente: enquanto o primeiro apresenta próclise generalizada, o segundo apresenta variação quer em frases finitas quer em frases infinitas.

(1) Colocação clítica do PB
   a. João **te** olhou
   b. **Me** abrace

(2) Colocação clítica do PE
   a. João olhou-**te**
   b. Abrace-**me**

Neste artigo, para o fenômeno da colocação clítica, discutimos os casos em (1b) e (2b), em cujo contexto o PB difere muito fortemente do PE, visto que o clítico não aparece em primeira posição na sentença, isto é, em um contexto V1[9].

Dentre as estruturas típicas que diferem o PE do PB aparece a estrutura progressiva. Enquanto na primeira variedade a perífrase progressiva é veiculada pelo auxiliar *estar* junto de um domínio não finito infinitivo encabeçado pela preposição *a* (*estar* + *a* + *infinitivo*), na segunda, a leitura progressiva também é estruturada pelo verbo *estar*, mas esse auxiliar é seguido por um domínio gerundivo (*estar*

[9] São consideradas sentenças V1 aquelas em que o verbo é o primeiro elemento da sentença e também as que são iniciadas por interjeições, conjunções coordenativas, vocativos e clíticos. Neste trabalho analisamos apenas as sentenças iniciadas por verbo e por um clítico, deixando os demais contextos de lado.

+ gerúndio), conforme os contrastes a seguir ilustram[10].

(3) a. Olha! A Maria **está a correr**. (PE, *PB)

b. Olha! A Maria **está correndo**. (*PE, PB)

Nota-se que, muito embora algumas regiões do sul de Portugal, como Alentejo e Algarve, também façam uso de (3b) com marcas de gerúndio flexionado (cf. *estão me chamandem* retirado de LOBO, 2003, p. 381), o uso de (3a) é bastante disseminado no PE (cf. HRICSINA, 2014; LOBO, 2003; MIRA MATEUS *et al.*, 2003). Observar tal variação se faz, portanto, relevante pois pode fornecer evidências para o período em que a gramática do PB emergiu.

3.1. A colocação de clítico em contexto V1 e a estrutura progressiva no *corpus* de representações artísticas

Quando o verbo aparece em primeira posição na sentença, tanto no Português Clássico (PCl) (século XVI-XVII) quanto no PE a ênclise é obrigatória. Dessa maneira, a gramática do PB mostra-se inovadora permitindo a ocorrência de próclise em sentenças com verbo em primeira posição absoluta. Inclusive, na sincronia, a próclise é categórica. No sistema de busca do *Corpus* Tycho Brahe, a anotação eletrônica do *corpus* possibilita rápida e prontamente a busca pela colocação clítica. A fim de observar sua distribuição ao longo do tempo, a busca pelos dados foi construída considerando a posição em que o clítico aparece. Nos dados de representações artísticas analisados, encontramos 37 casos (7%) de próclise e 496 casos (93%) de ênclise em sentenças com o verbo em posição inicial.

Para os casos de próclise, buscamos[11] pela etiqueta CL*|SE* (para clítico e clítico reflexivo *se*) considerando que o clítico deveria aparecer em primeira posição, por isso a etiqueta *isFirst*. Também foram desconsideradas as vírgulas em relações de precedência imediata. Alguns resultados obtidos para cada século estão apresentados abaixo.

(4) Algoritmo de busca fornecido para encontrar próclise em contexto V1
CL|SE|CL+*|SE+* isFirst

(5) Próclise em contexto V1

a. **Me** dá coceira ver tanto bagunça e tanta coisa fora do lugar! (1705)

b. **Te** arrependeste a tempo. (1829)

c. **Te** peguei, seu capitalista! (1934)

d. **Me** explica melhor o que você está sabendo ... (1978)

---

[10] O asterisco usado no exemplo representa a agramaticalidade da sentença seja na variedade brasileira seja na europeia.

[11] É possível realizar buscas usando o *Corpus* Tycho Brahe de dois modos: (i) a partir da página "Consulta aos textos etiquetados" (http://www.tycho.iel.unicamp.br/~tycho/corpus/texts/csquery/csquery.html), em que se pode construir buscas graficamente – com o auxílio de uma interface – ou manualmente – seguindo a sintaxe de busca usada na ferramenta *CorpusSearch* (disponível em http://corpussearch.sourceforge.net/CS-manual/Contents.html) ou (ii) a partir da instalação da ferramenta *CorpusSearch*. A grande diferença entre os dois modos é que a busca a nível sintático apenas pode ser realizada com a ferramenta CorpusSearch instalada na máquina.

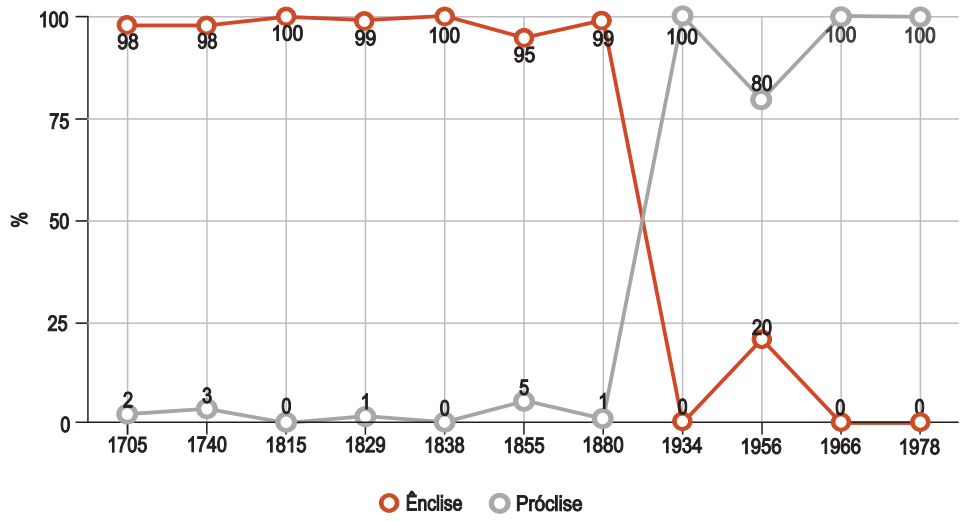
Para os casos de ênclise, buscamos pela etiqueta VB*+CL em primeira posição (*isFirst*) considerando que o clítico deveria aparecer após qualquer tipo de flexão verbal (representado pelo asterisco na etiqueta VB para verbo). Também foram desconsideradas as vírgulas em relações de precedência imediata. Os exemplos abaixo ilustram alguns dos resultados de ênclise para os séculos analisados.

(6) Algoritmo de busca fornecido para encontrar ênclise em contexto V1
VB*+CL isFirst

(7) Ênclise em contexto V1
a. Favoreça-**me** vossa mercê, Senhora Fagundes, com o seu voto, que eu terei bom despacho no tribunal de Cupido. (1705)
b. Vi-**as** em Paris. (1829)
c. Orgulhem-**se** de mim. (1956)

O gráfico 1 mostra em porcentagem o contraste entre próclise e ênclise no contexto de verbo em primeira posição ao longo do tempo. Nela é possível perceber que, muito embora a próclise já permeava o sistema da língua desde 1705, a ênclise era a colocação categórica até o fim do século XIX. A partir de 1880, o cenário inverte e a próclise passa a ser generalizada.

Gráfico 1: Contraste próclise *vs.* ênclise em contexto V1 no *corpus* de representações artísticas



Fonte: elaboração dos autores

No que se refere à busca da estrutura progressiva, para a estrutura prototípica do PE, *estar* + *a* + *infinitivo*, buscamos pela etiqueta ET*, que inclui nos resultados apenas o verbo *estar*; sendo o asterisco uma instrução à busca para se obter resultados com qualquer tipo de flexão no verbo. Para a busca da preposição *a*, foi especificado que se desejava uma preposição, pela etiqueta P, mas não qualquer uma, apenas *a*, com a subespecificação *a*. Por fim, para verbos no infinitivo, a etiqueta que compôs a busca foi VB, de verbo no infinitivo. Alguns dos resultados gerados ao longo dos séculos estão exemplificados abaixo.

(8) *estar* + *a* + *infinitivo*

a. Vá, senhora, não **esteja a choramingar**. (1815)

b. O que **está** o senhor **a dizer**? (1815)

c. Não, abaixei-a para evitar um cascudo que o patrão pretendia dar-me em um belo dia em que **estava a olhar** para a rua, em vez de servir as freguesas, e não voltei mais à loja. (1838)

d. **Estava a olhar** para a luz da lua. (1880)

Para a estrutura típica brasileira, a busca foi praticamente a mesma, com a diferença de que não havia nela a preposição, e, no lugar de um verbo no infinitivo, havia um verbo no gerúndio, representado por VB-G (não um infinitivo), compondo a estrutura *estar* + *gerúndio*. Exemplos dos resultados aparecem a seguir.

(9) *estar* + *gerúndio*

a. Homem, acabemos com isso, venha Dona Clóris, por quem **estou suspirando**. (1705)

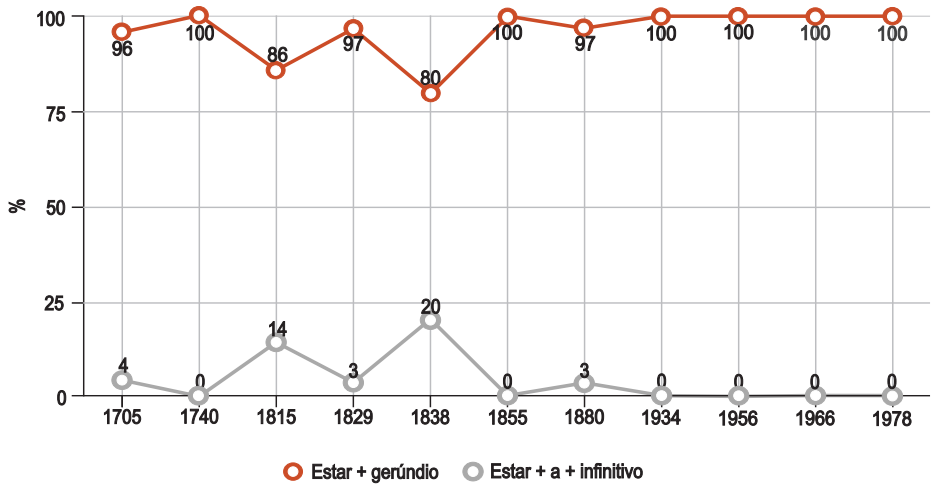
b. Venha, mamãe **está** nos **esperando**. (1829)

c. Ei, Tião, **está** me **ouvindo**. (1934)

d. Carlos Alberto, **estou** te **ligando** pelo seguinte. (1978)

A variação ao longo do tempo está ilustrada no gráfico 2, em que é possível observar que em todos os textos a frequência da estrutura típica brasileira *estar* + *gerúndio* é maior que a variante europeia, padrão semelhante ao do PCl (cf. discussão em §4). Há certa disparidade nos autores nascidos em 1815 e 1838, que usam a variante europeia; porém ainda assim o progressivo é veiculado majoritariamente pela perífrase inconfundível do Brasil, que é a estrutura usada categoricamente a partir do século XIX. Nota-se que a estrutura progressiva está presente desde o início do século XVIII, o que é diferente do comportamento encontrado para a próclise, que só emerge a partir do fim do século XIX.

Gráfico 2: Contraste *estar* + *gerúndio vs. estar* + *a* + *infinitivo* ao longo do tempo no *corpus* de representações artísticas



Fonte: elaboração dos autores

3.2. A colocação de clítico em contexto V1 e a estrutura progressiva no *corpus* de cartas e atas da Câmara Municipal de Salvador

Por mérito da anotação morfossintática, foi possível recuperar facilmente os dados sobre a estrutura progressiva e a colocação de clíticos. Diferente do *corpus* de representações artísticas, nossos dados contam também com anotação sintática. Por isso, a busca por dados é feita com o auxílio da ferramenta *CorpusSearch*[12], que deve ser instalada no computador. A sintaxe das buscas, portanto, difere um pouco do que foi apresentado anteriormente.

Para os dados de ênclise, seguimos as etapas de busca elaboradas por Andrade e Namiuti (no prelo), com a seguinte sintaxe:

(10) Algoritmo fornecido para encontrar ênclise em contexto V1
node: IP-MAT
query: (IP-MAT iDomsMod NP* CL)
AND (IP-MAT iDomsFirst flex_vb)

A sintaxe da busca informa o seguinte: busque todas as sentenças matrizes que dominam imediatamente um clítico (CL|SE), e os únicos nós que intervêm no caminho do IP matriz a um clítico são membros de um sintagma nominal (NP*) e que o IP matriz domina imediatamente como primeiro filho um verbo flexionado (flex_vb).

Para os dados de próclise, a sintaxe foi a seguinte:

(11) Algoritmo fornecido para encontrar próclise em contexto V1
node: IP-MAT
query: (IP-MAT iDomsFirst NP*)
AND (NP* idoms CL|SE)
AND (NP* iprecedes flex_vb)

Busque todas as sentenças matrizes que dominam imediatamente como primeiro filho um sintagma nominal (NP*) e o sintagma nominal domine imediatamente um clítico (CL|SE) e preceda imediatamente um verbo flexionado (flex_vb).

O resultado das buscas soma 57 ocorrências, 49% de ênclise (28 casos) e 51% de próclise (29 casos). Os percentuais e contexto de algumas ocorrências são apresentados a seguir:

(12) Ênclise em contexto V1

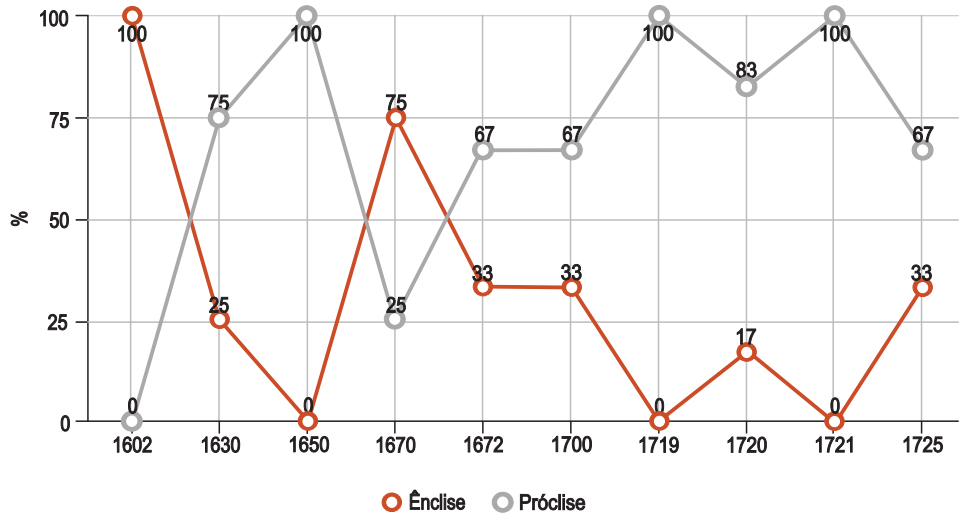
Concedemo-**lo** enquanto a avaria feita a requerimento das partes aos nossos donativos (va_PINHEIRO, 1602).

[12] http://corpussearch.sourceforge.net/

(13) Próclise em contexto V1
A nossaobrigação ao Serviço de Vossa Alteza, eade procurar obem Comum | desta Republiqua da B[*ahi*]a *[s:syn-clause]* **nos** leva por esta aos Reaes Pés de V[*ossa*] A[*lteza*] a Repre- | zentar o prejuizo q[*ue*] se seguem de sehirem fundando, e fazendo pella | terra dentro muitos Eng[*enh*]os deaSuc[*are*]S junto hus dos outros sem fun- | damento deterra bastante aoq[*ue*] demandão delenhas para oseu gasto que he muito grande (va_Aragão, 1630).

O gráfico 3 mostra em porcentagem o contraste entre próclise e ênclise em sentenças V1 entre nascidos entre os séculos XVI e XVII. Observa-se que já na primeira metade do século XVII, o padrão de colocação de clítico encontrado em sentenças V1 é proclítico, diferindo do PCl (GALVES, BRITO; PAIXÃO DE SOUSA, 2005), cujo padrão é enclítico. O padrão se mantém entre os nascidos ao longo do século XVIII.

Gráfico 3: Contraste ênclise *vs.* próclise em contexto V1 ao longo do tempo no *corpus* de cartas e atas da Câmara Municipal de Salvador



Fonte: elaboração dos autores

Nos dados, já há evidências de que o padrão de colocação de clíticos começa a se afastar do padrão encontrado na gramática clássica, apontando para a emergência de uma gramática brasileira.

Os dados relacionados à estrutura progressiva também foram coletados com o auxílio da ferramenta *CorpusSearch*. A sintaxe da busca foi a seguinte para a estrutura *estar* + a + infinitivo:

(14) Algoritmo fornecido para encontrar *estar + a + infinitivo*
node: IP*
query: (flex_estar precedes PP)
AND (PP iDoms P)
AND (P idoms a)
AND (PP iDoms IP-INF*)

Busque todas as sentenças em um nó IP qualquer em que o verbo *estar* flexionado (flex_estar) preceda um PP (sintagma preposicionado) e o PP domine imediatamente uma P(reposição) *a*, que domine imediatamente uma sentença infinitiva. Já a sintaxe para a estrutura *estar* + gerúndio informa o seguinte:

(15) Algoritmo fornecido para encontrar *estar* + *gerúndio*
node: IP*
query: (flex_estar HasSister IP-GER)

Busque em um nó IP qualquer em que o verbo *estar* flexionado (flex_estar) tenha como irmão uma sentença no gerúndio.

Os dados de progressivo somam 49 ocorrências, sendo a estrutura *estar* + a + infinitivo a menos encontrada (4%), diferente da estrutura *estar* + gerúndio com 96% dos casos. A seguir reproduzimos algumas ocorrências no *corpus* analisado:

(16) *estar* + *a* + *infinitivo*

a. E porque Sua Excelência várias vezes tem recomendado o seu conserto por ser aquela ladeira e por onde se há de conduzir o Marquês e Arcebispo que **estão a chegar** da Índia requereu se fizesse na dita ruína inspeção ocular para à vista dela se lhe dar a providência que pede a necessidade o que visto pela vereação assim o mandaram de que mandaram fazer este termo que assinaram (va_Sodré, 1710).

b. A causa que defendo contra os irmãos da Casa da Misericórdia desta cidade com que nos receberam os embargos **está a ressoar** sobre estes (va_Magalhães, 1680).

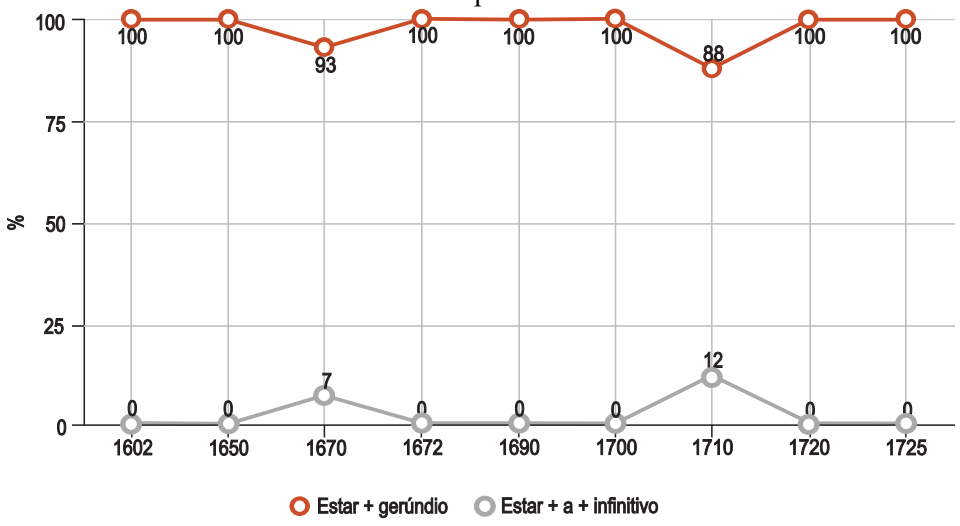
(17) *estar* + *gerúndio*

a. E na dita vereação requereu o procurador atual da Câmara que como capitão André Marques não tinha feito a conta do que era devedor ao donativo real **estava devendo** a este, (va_Sodré, 1710).

b. Além da ruína que **está ameaçando** a terra dos quintais[,] por lhe faltar o disparo dos muros por causa do inverno[,] caíram (va_Silveira, ~1725).

De um modo geral, os resultados encontrados apontam para a manutenção da expressão da estrutura progressiva tal como ocorre no PCl. Há prevalência pela estrutura *estar* + *gerúndio* nos dados brasileiros entre os séculos XVII e XVIII, com alguns poucos casos de *estar* + *a* + *infinitivo*, atestado em indivíduos cultos, que viveram por um período de tempo em Lisboa. Os dados, considerando a data de nascimento, são apresentados no gráfico 4:

Gráfico 4: Contraste *estar + gerúndio vs. estar + infinitivo* ao longo do tempo no *corpus* de cartas e atas da Câmara Municipal de Salvador



Fonte: elaboração dos autores

A seguir, discutimos os resultados.

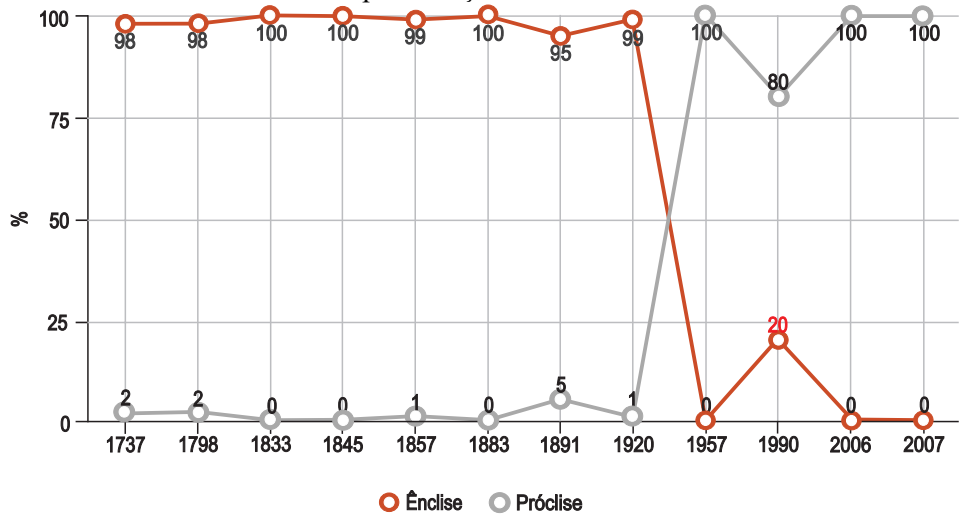
3.3. Discussão dos resultados

Visto que os dados relativos à perífrase progressiva são similares em ambos os *corpora* e o que de fato difere são os dados de colocação de clítico, passamos a discutir o aspecto ligado à tipologia textual e ao contraste *geração biológica versus geração histórica*.

Do ponto de vista da data de publicação dos textos, os dados diacrônicos em ambos os *corpora* mostram que a mudança é detectada mais tardiamente, comparados com os dados a partir da data de nascimento dos autores. Na seção anterior, mostramos que a mudança do padrão ênclise para próclise ocorreu nos textos de representações artísticas no fim do século XIX, com um crescimento abrupto de 1% para 100% de próclise em contexto V1 (cf. gráfico 1). Já para o *corpus* de Cartas e Atas da Câmara Municipal de Salvador, a mudança aparece ainda na segunda metade do século XVII (cf. gráfico 3), de modo não tão abrupto como nos *corpora* de representações artísticas.

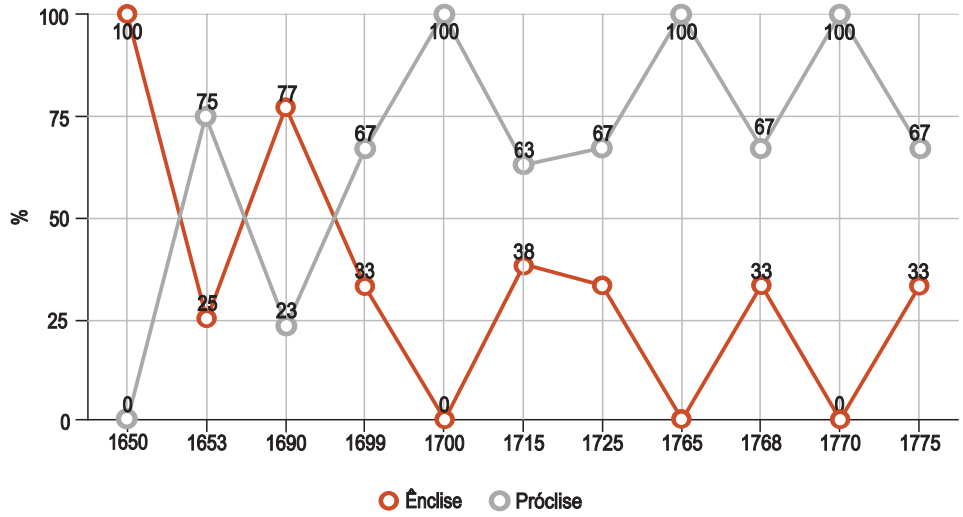
Quando o parâmetro de análise passa a ser a data de publicação dos textos, a mudança ênclise-próclise ocorre não mais no fim do século XIX, mas no começo do século XX nos textos de representação artística, quando a colocação clítica passa a se comportar de maneira oposta ao padrão até então vigente, como fica patente no gráfico 5. O mesmo acontece com relação à percepção da mudança nos dados do *Corpus* de cartas e atas da Câmara Municipal de Salvador, cuja mudança torna-se perceptível no início do século XVIII, como aparece no gráfico 6.

Gráfico 5: Contraste próclise vs. ênclise em contexto V1 por data de publicação no *corpus* de representações artísticas



Fonte: elaboração dos autores

Gráfico 6: Contraste próclise vs. ênclise em contexto V1 por data de publicação no *corpus* de Cartas e Atas da Câmara Municipal de Salvador



Fonte: elaboração dos autores

Olhando para o *corpus* de representação artística, uma das vantagens de se adotar a *geração histórica* é que a data de publicação possibilita observar o comportamento do português brasileiro no século XXI, visto que não é possível olhar para a *geração biológica* no mesmo século: autores que nasceram no advento do século XXI ainda não publicaram peças de teatro. Futuramente, com peças escritas por autores que nasceram no século XXI, vamos ter dados o suficiente para contrastar e compreender mais acuradamente se de fato a data de publicação de uma peça reflete o conhecimento linguístico do período publicado ou se é a data de nascimento do autor. Como ainda não temos o queijo e a faca em mãos, acaba sendo importante e fulcral levar em consideração os dois parâmetros temporais, quando se usa peças de teatro na diacronia.

Alguns estudos sincrônicos mostram que o português contemporâneo é categoricamente proclítico (cf. CARNEIRO, 2016; VIEIRA, 2016), mas desde quando essa característica aparece? Com os dados do *corpus* de Cartas e Atas da Câmara Municipal de Salvador podemos aventar que essa característica aparece já no final do século XVII, apesar de parecer "desaparecer" nos dados de textos teatrais. Aqui fica evidente que a tipologia textual é um fator importante a ser considerado quando olhamos para os dados diacrônicos.

Ribeiro (1998, p. 114), ao avaliar o clítico em início de sentença, lista uma série de dados assistemáticos sobre esse fenômeno e sugere que "só com uma busca em documentos de diferentes tipos e autores se poderá ter uma ideia do valor dessas construções". É isso que fazemos nos gráficos 7 e 8.

Mas antes, insistimos na ideia de que, para se evitar que façamos linguística histórica de apenas uma tipologia textual, é fundamental "misturar" diferentes tipologias a fim de buscar um *corpus* "equilibrado" de textos, "representativo" o máximo possível da língua de uma época. Desse modo, é essencial fazer

> uma história da língua que estuda as diferentes tradições sem limitar-se a uma só, mantendo a diferenciação – uma história da língua menos monolítica que permitirá saber em quais T[radições] D[iscursivas] uma inovação é criada, como se difunde ao longo das TD, e também onde há TD resistentes às inovações, TD que preservam elementos que em outras TD não se usam mais (KABATEK, 2006, p. 516).

Superficialmente, poderíamos afirmar que os textos teatrais por estarem mais próximos da oralidade refletiriam de modo "mais genuíno" a gramática de falantes de determinado período. Porém, observando essa tradição discursiva, vê-se que os textos teatrais parecem ser mais resistentes às inovações do que as cartas e atas. Portanto, é necessário delinear a função desse tipo textual desde o momento do chamado "descobrimento" do Brasil para compreender o porquê de as inovações não serem tão evidentes.

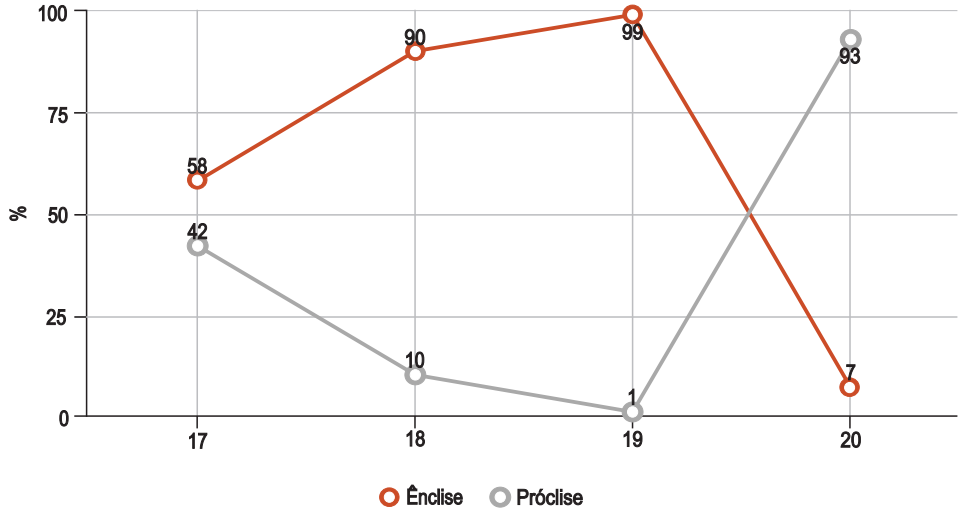
Inicialmente, as manifestações teatrais no Brasil Colônia tinham um caráter pedagógico, pois pretendiam educar religiosamente os indígenas (MAGALDI, 1997). Posteriormente, a partir do século XVIII, segundo Mayor (2015, p. 105), "[...] o teatro na colônia adquiriu mais uma função além de seus significados políticos, pedagógicos e religiosos, pois começou a se tornar também uma mercadoria". Mayor (2015, p. 106) também considera "as produções culturais da colônia não como autônomas, mas sempre relacionadas com as portuguesas" e chama a atenção para o fato de que, nesse período,

> muitas das peças apresentadas, por exemplo, eram de autoria do dramaturgo Antônio José da Silva, o Judeu, que passou toda sua curta vida em Lisboa, apresentando-se no Teatro do Bairro Alto. Por isso, o desafio do estudo dos temas coloniais é justamente investigar quais seriam as especificidades da Colônia, considerando os modelos europeus (MAYOR, 2015, pp. 106-7).

Fica evidente, desse modo, que, apesar de ser uma tipologia mais próxima da oralidade, os textos teatrais cumpriam várias funções na colônia – políticos, pedagógicos e religiosos – de sorte que seguiam muito mais de perto o modelo português do que textos de tradições discursivas mais fortemente fixadas, como as cartas e atas.

Como defende Kabatek (2006), a visão de conjunto da história da língua só poderá ficar mais tangível se olharmos para um *corpus* diacrônico multidimensional. Nesse sentido, o *Corpus* Tycho Brahe cumpre essa "missão", visto que é resultado da reunião de diferentes tipologias textuais. Dispor de tipologias textuais variadas é fundamental para poder tirar conclusões quantitativas e qualitativas sobre o que se quer analisar. Por isso, unimos os dados de ambos os *corpora* para ter uma visão mais global dos fenômenos investigados. Comecemos com os dados apresentados no gráfico 7 de clítico em posição inicial.

Gráfico 7: Contraste próclise vs. ênclise em contexto V1 por século de nascimento em ambos os corpora investigados



Fonte: elaboração dos autores

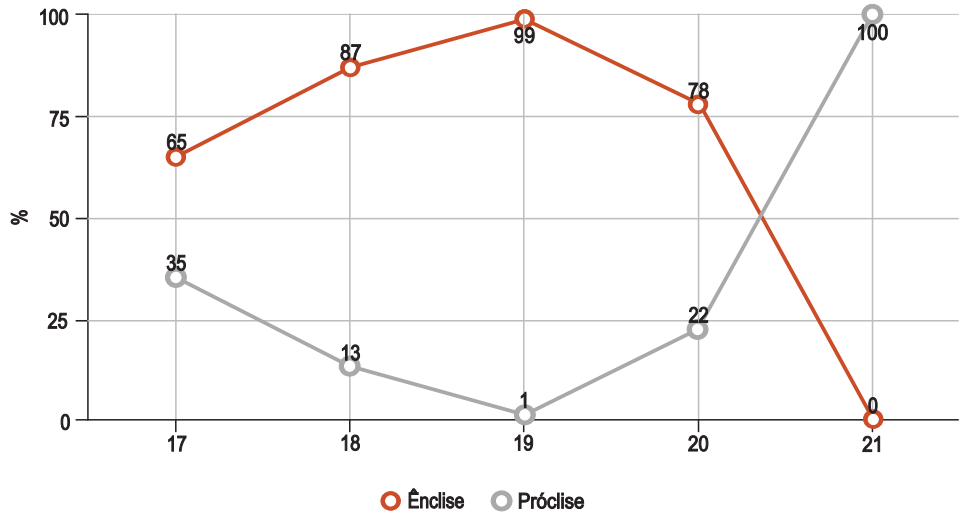
Com base nos resultados apresentados no gráfico 7, as ocorrências de próclise em orações com o verbo em primeira posição absoluta aparecem já no século XVII e somam 10% no século XVIII. Chama a atenção que tanto nos textos teatrais quanto nas cartas e atas da câmara o percentual de próclise é maior do que registrado em trabalhos anteriores (cf. PAGOTTO, 1992; LOBO, 2001; CARNEIRO, 2005; MARTINS, 2010; CARDOSO, 2020).

Entre os nascidos no século XIX, atesta-se um aumento no emprego da ênclise; e a próclise em contextos V1 é bastante marginal, ocorrendo em apenas 1% dos dados. Já entre os nascidos no século XX, o aumento nas taxas de próclise em V1 é abrupto: os valores passam de 1% para 90% dos casos.

Os padrões encontrados, quando olhamos para o século de publicação (cf. gráfico 8), diferem-se. Nos textos publicados no século XX, ainda há um alto percentual de ênclise (78%), padrão que reflete

a gramática do PE. O padrão encontrado nos textos publicados nos séculos 19 e 20 reflete uma pressão da norma culta padrão (PAGOTTO, 1998; CARNEIRO, 2005; MARTINS, 2010), o que não deixa claro quando a gramática brasileira é implementada.

Gráfico 8: Contraste próclise vs. ênclise em contexto V1 por século de publicação em ambos os corpora investigados

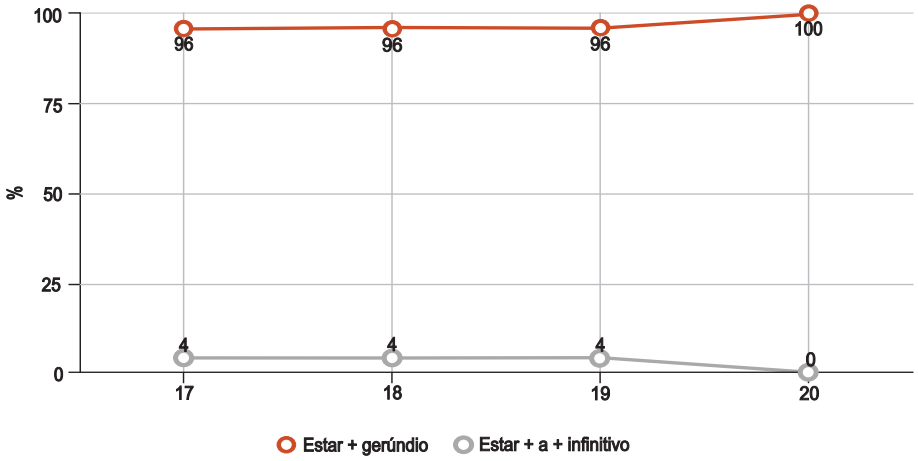


Fonte: elaboração dos autores

De modo a termos uma visão ampla do fenômeno da variação da estrutura progressiva, unimos os dados de ambos os *corpora* no gráfico 9. A partir dela, temos evidência de que, independente do tipo textual, a estrutura *estar* + *gerúndio* para veicular progressividade estaria presente na gramática brasileira desde o século XVII, seguindo o padrão do PCl, sendo a estrutura *estar* + *a* + *infinitivo* a inovação do PE: até o século XIX, o PE fazia uso extensivo de complemento infinitivo encabeçado pela preposição *a* com verbos como *começar*, *tornar* e *dever*, mas não com o verbo *estar*: seu complemento típico também era o gerúndio (cf. HRICSINA, 2014 para dados do PE do século XIV ao XX). É no século XIX ou ligeiramente antes que o complemento infinitivo toma lugar generalizado das perífrases verbais[13] (HRICSINA, 2014). Nota-se pelo gráfico que alguns autores talvez tenham sido influenciados pela metrópole, porém nem mesmo a pressão da norma culta padrão lusitana foi capaz de inverter o cenário em terras tupiniquins. Diante disso, é razoável pensar que se houve alguma tentativa de implementar a estrutura *estar* + *a* + *infinitivo* no PB, essa tentativa foi fracassada.

[13] Hricsina (2014, p. 393-395) nota que o infinitivo substituiu plenamente o complemento das perífrases, sendo usado inclusive com *ir* (e também com gerúndio), *vir* (e também com gerúndio), *continuar*, *andar*, *ficar*. Ainda nota que, muito embora em orações adverbiais reduzidas se use o gerúndio (*Chovendo, não vamos à praia*), em orações adjetivais reduzidas e orações reduzidas também a preferência é pelo infinitivo: *... pessoas a fazer palhaçadas* (equivalente ao *... pessoas fazendo palhaçada* do PB) e *O Pedro ficou sentado a falar com o seu pai* (equivalente ao *O Pedro ficou sentando falando com seu pai do PB*).

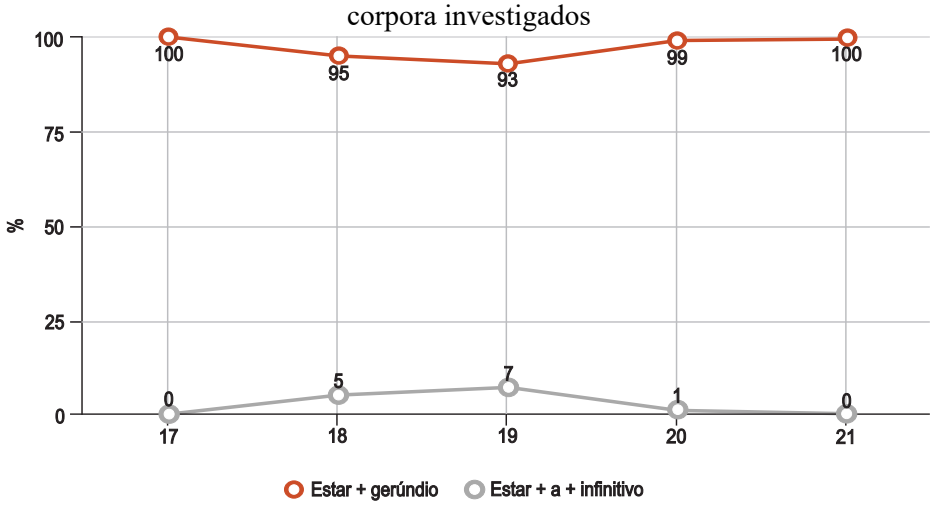
Gráfico 9: Contraste estar + gerúndio vs. estar + a + infinitivo por século de nascimento em ambos os corpora investigados



Fonte: elaboração dos autores

Quando a análise é baseada no século de publicação, os dados do fenômeno em questão também se diferem substancialmente. O gráfico 10 mostra esses resultados, sugerindo que talvez os textos publicados no século XIX tenham sido influenciados pela pressão normativa de Portugal, já que é nesse século que mais aparece a variante europeia nos textos publicados. Considerando os resultados de Hricsina (2014, p. 400), parece que o processo de mudança que "deve estar situado no século XIX ou ligeiramente antes (século anterior)" no PE influenciou a produção de *estar + infinitivo* nos textos publicados aqui, justamente por observamos ausência dessa estrutura nos textos publicados no século XVII, caracterizando textos puramente do PCl, mas um aumento tímido dessa estrutura exatamente quando ela emergiu em Portugal. Dessa maneira, sob o ponto de vista da geração histórica, a hipótese de que a eventual implementação de *estar + a + infinitivo* no PB teria sido uma tentativa fracassada, levantada anteriormente, parece se confirmar.

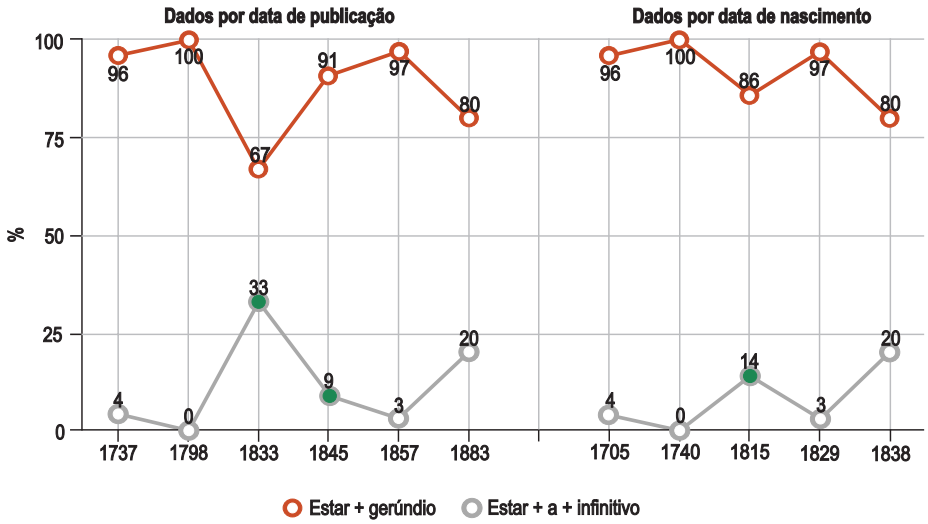
Gráfico 10: Contraste estar + gerúndio vs. estar + a + infinitivo por século de publicação em ambos os corpora investigados



Fonte: elaboração dos autores

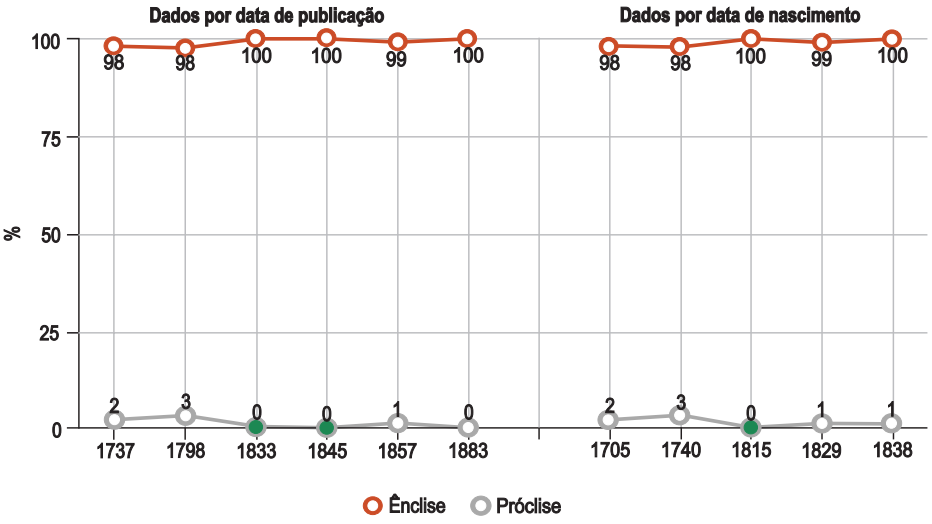
Para dar cabo à questão da *geração biológica* contrastivamente à *geração histórica*, chama atenção os resultados de Martins Pena – autor de duas peças publicadas em épocas diferentes, 1833 e 1845. Caso nos baseemos apenas na data de nascimento, por exemplo, não capturamos o fato de que, muito embora se trate do mesmo autor, há variação na gramática empregada nas peças. Isso é muito claro na comparação feita entre o uso de ênclise e próclise com a estrutura progressiva: muito embora o autor tenha a mesma gramática em relação à colocação clítica em ambas as peças (cf. gráfico 12), houve variação no uso da estrutura progressiva, com diferença de 24 pontos percentuais entre um texto e outro (cf. gráfico 11). Esse fato curioso, por exemplo, só é capturado quando a data de publicação é trazida à análise.

Gráfico 11: Contraste estar + gerúndio vs. estar + a + infinitivo em Martins Pena por ano de publicação e ano de nascimento



Fonte: elaboração dos autores

Gráfico 12: Contraste ênclise vs. próclise em Martins Pena por ano de publicação e ano de nascimento



Fonte: elaboração dos autores

Com relação à tipologia textual, os dados de progressivo mostram-nos um comportamento bastante similar entre os *corpora*. O mais interessante é que os dados de progressivo também nos informam sobre as origens do PB se recolocamos a pergunta de Ribeiro (1998, p. 115): "Pode-se falar em mudança linguística do PB tendo como parâmetro unicamente os dados do PE?".

A resposta à pergunta acima, olhando para os dados de estrutura progressiva, é negativa. Enquanto o padrão prototípico da gramática do PE *estar* + *a* + *infinitivo* aparece apenas a partir do século XIX, segundo Hricsina (2014, p. 399), o padrão encontrado no PB é similar ao padrão do PCl. Nesse sentido, a gramática do PE é mais inovadora do que a gramática do PB e difere do padrão encontrado em outras línguas românicas, com exceção do francês; e diferente do que acontece com a colocação clítica, os dados de progressivo não parecem sofrer pressão da norma culta. Sob esse ponto de vista, ainda, podemos afirmar, consoante Corôa (2021, p. 6757), que "a mudança ocorrida em Portugal a partir do século XVIII não afeta a gramática desenvolvida no Brasil", pelo menos não inteiramente.

Conclusões

O objetivo central deste artigo foi apresentar os novos *corpora* que passam a fazer parte do *Corpus* Tycho Brahe, quais sejam, cantiga, comédias teatrais cariocas, cartas e atas de Homens Bons da Câmara Municipal de Salvador. Apontamos que com esses novos dados podemos lançar novos olhares para a diacronia do português brasileiro.

Considerando que o objeto da linguística histórica é o texto escrito, mostramos que o linguista diacrônico, mais especificamente o que trabalha sob a perspectiva gerativista, para acessar indiretamente o conhecimento linguístico de falantes de tempos remotos, precisa tomar certas decisões antes de iniciar seu labor. Uma delas é qual tipo/gênero textual vai representar a língua do seu recorte temporal e qual parâmetro temporal vai ser adotado, a data de publicação desses tipos/gêneros textuais ou de nascimento dos seus autores. Neste artigo, demonstramos a significância de tais escolhas a partir de dois fenômenos que distinguem claramente o PE do PB: a colocação do clítico em contexto V1 e a perífrase progressiva.

Quando contrastamos dois tipos textuais, percebemos que certos fenômenos são mais sensíveis a essa escolha que outros. Por exemplo, a leitura de progressividade era categoricamente veiculada pela estrutura *estar* + *gerúndio* tanto nas representações artísticas quanto nas cartas e atas, o que sugere que tanto em tipologias textuais mais orais, como nas peças, quanto em tipologias textuais mais escritas, a estrutura era a mesma. No entanto, analisando a colocação clítica, observa-se que as cartas e as atas apresentam um comportamento majoritariamente proclítico desde o século XVII, enquanto as peças, um comportamento mais conservador, passando a ser majoritariamente proclítico a partir do século XIX. Sem a comparação desses novos dados, duas análises distintas poderiam emergir: a próclise não/é uma inovação do século XIX.

Adicionalmente, demonstramos que, a depender da periodização, considerar a geração biológica ocasiona uma análise da mudança mais precoce. Caso a análise seja feita a partir da geração histórica, a mudança será delineada mais tardiamente. Por exemplo, a clara mudança do padrão enclítico para o proclítico em contexto V1, tendo como base as peças de teatro, ocorreu de fato no fim do século XIX (com a geração biológica) ou no meio do século XX (com a geração histórica)?

Em suma, neste artigo tornamos públicos textos históricos ineditamente anotados (morfologicamente e morfossintaticamente) que passam a compor o *Corpus* Tycho Brahe, agora com 2.783.779 palavras, contribuindo com novos dados a fim de termos novos olhares para a história do português brasileiro. Ficou notório que apenas com novos dados é possível expandir nossas análises e, sobretudo, compreender o que está em jogo quando se escolhe por certo tipo textual ou por certa periodização temporal. Esperamos ter mostrado que tais escolhas não são isentas de consequências e que, no melhor dos mundos, o linguista diacrônico deve ter consciência das suas implicações.

Referências

ANDRADE, Aroldo; NAMIUTI, Cristina. Gone without the verb: clitic interpolation and clitic climbing in the history of European Portuguese. *Caderno de Estudos Linguísticos*, Campinas, v. 58, n. 2, pp. 201-19, 2016.

BARBOSA, Afranio. O português escrito no século XVIII: fontes reunidas na Biblioteca Nacional do Rio de Janeiro. *In*: CASTILHO, Ataliba (ed.). *Para a história do português brasileiro*. São Paulo: Humanitas Publicações, 1998, v. Primeiras Idéias, pp. 229-38.

CAMBRAIA, César Nardelli. A pesquisa diacrônica e o problema do *corpus*. *Anais da Semana de Estudos de Língua Portuguesa*, v. 2, n. 2, pp. 11-9, 1994.

CARDOSO, Lara. *A gramática dos pronomes clíticos no Brasil Colônia*: o português clássico na história do português brasileiro. 2020. Dissertação (Mestrado em Estudos Linguísticos) – Programa de Pós-Graduação em Estudos Linguísticos, Universidade Estadual de Feira de Santana, Feira de Santana, 2020.

CARNEIRO, Zenaide de Oliveira Novais. *Cartas brasileiras (1809-1904)*: um estudo lingüístico-filológico. 2005. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem, Unicamp. Campinas, SP, 2005.

CARNEIRO, Zenaide de Oliveira Novais. Colocação de clíticos em orações finitas em duas vertentes do português oral feirense: um contexto não variável. *In*: ALMEIDA, Norma Lucia Fernandes de; ARAUJO, Silvana Silva de Farias; TEIXEIRA, Eliana Sandra Pitombo; CARNEIRO, Zenaide de Oliveira Novais. (org.). *Variação Linguística em Feira de Santana*. Feira de Santana: UEFS Editora, 2016, v. 1, pp. 141-74.

CORÔA, Williane. Novos elementos para a periodização do português no Brasil. *Fórum Linguístico,* v. 18, n. 3, pp. 6748-59, 2021. https://doi.org/10.5007/1984-8412.2021.e78982.

CYRINO, Sonia; TORRES MORAIS, Maria Aparecida. *Mudança sintática do português brasileiro: perspectiva gerativista*. História do Português Brasileiro (Coordenador geral: Ataliba de Castilho), São Paulo: Contexto, 2018. v. 6.

DUARTE, Maria Eugênia. Apresentação. *In*: DUARTE, Maria Eugênia (ed.). *O sujeito em peças de teatro (1833-1992): estudos diacrônicos*. São Paulo: Parábola, 2012. pp. 11-9.

GALVES, Charlotte. Posfácio: o retrato da emergência de uma nova gramática. *Mudança sintática do português brasileiro: perspectiva gerativista*. História do Português Brasileiro (Coordenador geral: Ataliba de Castilho), São Paulo: Contexto, 2018. v. 6, pp. 441-56.

GALVES, Charlotte; FARIA, Pablo. *Tycho Brahe Parsed Corpus of Historical Portuguese*. 2010. URL: http://www.tycho.iel.unicamp.br/~tycho/corpus/en/index.html.

GALVES, Charlotte; ANDRADE, Aroldo Leal de; FARIA, Pablo. *Tycho Brahe Parsed Corpus of Historical Portuguese*. 2017.

GARCÍA GARCÍA, L. A case study in historical linguistic research. *In*: BRUTON, Anthony; GARCÍA, Luisa García; DOMÍNGUEZ, Joaquín José Fernández (orgs). *Perspectives on the genitive in english: Synchronic, Diachronic, Contrastive and research*. Universidad de Sevilla: 2000, pp. 118-29.

HRICSINA, Jan. Substituição do gerúndio pela construção a + infinitivo no Português Europeu (estudo diacrónico). *Studia Iberystyczne*, v. 13, pp. 383-401, 2014.

HUNSTON, S. *Corpora in appled linguistics*. Cambridge: Cambridge University Press, 2002.

JENSET, Gard B.; MCGILLIVRAY, Barbara. *Quantitative historical linguistic*s: A *corpus* framework. Oxford University Press, 2017.

KABATEK, Johannes. Tradições discursivas e mudança linguística. *In*: LOBO, Tânia; *et al*. (orgs.). *Para a história do português brasileiro*. Salvador: EDUFBA, 2006, v. 6, t. 1-2. pp. 505-27.

LIGHTFOOT, David. *Principles of diachronic syntax*. Cambridge, UK: Cambridge University Press, 1979.

LOBO, Maria. *Aspectos da Sintaxe das Orações Subordinadas Adverbiais do Português*. 2003. Dissertação de Doutoramento em Linguística/Sintaxe – Universidade Nova de Lisboa, Lisboa, 2003.

LOBO, Tânia. Depoimento sobre a constituição de um *corpus* diacrônico do português brasileiro – Bahia. *In*: CASTILHO, Ataliba (ed.). *Para a história do português brasileiro*. São Paulo: Humanitas Publicações, 1998. vol. Primeiras Idéias, pp. 171-95.

LOBO, Tânia. *Para uma sociolinguística histórica do português no Brasil*: edição filológica e análise linguística de cartas particulares do Recôncavo da Bahia, século XIX. 2001. 808f. Tese (Doutorado em Filologia e Língua Portuguesa) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. São Paulo, 2001.

MAGALDI, S. *Panorama do teatro brasileiro*. São Paulo: Global, 1997.

MARTINS, Marco Antônio. *Competição de gramáticas do português na escrita catarinense dos séculos 19 e 20.* 2009. 326 f. Tese (Doutorado em Linguística) – Universidade Federal de Santa Catarina, Florianópolis, 2009.

MATTOS E SILVA, Rosa Virgínia. Notícia sobre o Programa para a História da Língua Portuguesa. *Estudos Lingüísticos e Literários*, Salvador, número especial, pp. 231-7, 1996.

MAYOR, Mariana Soutto. O teatro do século XVIII no Brasil: das festas públicas às casas de ópera. Revista Aspas, v. 5, n. 2, pp. 103-10, 2015.

MIRA MATEUS, Maria Helena et al. *Gramática da Língua Portuguesa*, revista e aumentada 5. ed. Lisboa: Caminho, 2003.

NEVES, Larissa de Oliveira. *As Comédias de Artur Azevedo – Em Busca da História*. 2006. 212 f. Tese (Doutorado em Letras) – Instituto de Estudos da Linguagem, Unicamp, Campinas, 2006.

PAGOTTO, Emílio. *A posição dos clíticos em Português*: um estudo diacrônico. 1992. 168 f. Dissertação (Mestrado em Linguística) – Instituto de Estudos da Linguagem, Unicamp. Campinas, SP, 1992.

PAGOTTO, Emílio. Norma e Condescendência: Ciência e Pureza. *Língua e Instrumentos Linguísticos*, Campinas, v. 2, n. 1, pp.49-68, jul./dez. 1998.

PAIXÃO DE SOUSA, Maria Clara. *Língua barroca*: Sintaxe e história do português nos 1600. 2004. 455 f. Tese (Doutorado em Linguística) – Instituto de Estudos da Linguagem, Unicamp. Campinas, SP, 2004.

RIBEIRO, Ilza. A mudança sintática do Português Brasileiro é uma mudança em relação a que gramática? *In:* CASTILHO, Ataliba T. de. (org.). *Para a história do português brasileiro*: primeiras ideias. V.1. São Paulo: Humanitas, 1998. pp.101-9.

SARDINHA, Tony Berber. *Lingüística de corpus*. Barueri: Manole, 2004.

SILVA NETO, Serafim da. *Introdução ao estudo da língua portuguesa no Brasil.* Rio de Janeiro: Presença, 1977.

VIEIRA, Maria de Fátima. *A ordem dos clíticos pronominais nas variedades urbanas europeia, brasileira e são-tomense:* uma análise sociolinguística do português no início do século XXI. 2016. 238 f. Tese (Doutorado) - Universidade Federal do Rio de Janeiro, Faculdade de Letras, 2016.