

Evidências do Português Médio no *Corpus de Textos Antigos*

Evidences of Middle Portuguese in the *Corpus of Ancient Texts*

Maria Inês Bico^{1,2,a} , Esperança Cardeira^{1,2} , Jorge Baptista^{3,4} , Fernando Batista^{3,5} 

¹ Centro de Linguística da Universidade de Lisboa (Portugal), Portugal

² Faculdade de Letras da Universidade de Lisboa (Portugal), Portugal

³ INESC-ID Lisboa (Portugal), Portugal

⁴ Universidade do Algarve (Portugal), Portugal

⁵ ISCTE-IUL (Portugal), Portugal

✉ ^amariainesb1@edu.ulisboa.pt

Recebido: 08/04/2024; Aceito: 09/04/2025

Resumo

A partir de um conjunto de dados semi-automaticamente anotados do *Corpus de Textos Antigos* (CTA), este artigo propõe-se a analisar os resultados obtidos sobre a síncope de *-d-* intervocálico no morfema da 2.^a pessoa plural, e a consequente resolução do hiato, e as terminações de Particípio Passado *-udo/-ido* nos verbos com origem etimológica nas 2.^a e 3.^a conjugações latinas. A novidade deste artigo está no recurso a métodos de Processamento de Linguagem Natural (PLN) para a otimização da obtenção e extração sistemática dos dados relevantes para análise, contribuindo para um estudo que engloba um maior conjunto de textos. É apresentada a metodologia adotada para a anotação dos dados, e consequente extração dos dados relevantes à análise, afirmando-se a importância do recurso a métodos e ferramentas de PLN para o estudo linguístico e para a descrição dos estados anteriores da língua portuguesa.

Palabras-chave: Português Médio; Corpus de Textos Antigos; História do Português.

Abstract

Based on a set of semi-automatically annotated data from the *Corpus of Ancient Texts* (CTA), this paper aims at analysing the results obtained on the syncopation of intervocalic *-d-* in the second-person plural morpheme, resulting in a hiatus resolution, and the past participle ending forms *-udo/-ido* in verbs with an etymological origin in the 2nd and 3rd Latin conjugations. The novelty of this article lies in the use of Natural Language Processing (NLP) methods to optimise the systematic collection and extraction of relevant data for analysis, contributing to a study that encompasses a larger set of texts. The methodology used for annotating the data and, consequently, extracting the relevant data for analysis is presented, stating the importance of resorting to NLP methods and tools for the purpose of linguistic study and for describing previous stages of the Portuguese language.

Keywords: Middle Portuguese; Corpus of Ancient Texts; History of Portuguese.

1. INTRODUÇÃO

A periodização de uma língua é uma “ilusão útil” (Castro 2006:73). Dada a própria natureza da língua, não é possível determinar o momento concreto da mudança, que só é percebida *a posteriori*. No entanto o estabelecimento de balizas temporais que se podem estender por largos anos ou séculos auxilia a situar o investigador, sem que seja necessário precisar datas. As datas sugeridas para a delimitação do Português Médio são entre 1385 e 1420 para o seu início e entre 1536 e 1550 para o seu término (Castro 2006: 73). Não sendo datas consensuais entre a comunidade académica (Cardeira 2005:28), são datas que se baseiam em acontecimentos extralinguísticos: a vitória a favor dos portugueses na Batalha de Aljubarrota (1385) e a concretização da mudança de paradigma na sociedade (1420); a publicação da primeira gramática da Língua Portuguesa (1536) e o efeito desta reflexão metalinguística nos anos subsequentes (1550).

Independentemente das datas aqui apresentadas, este período que se estende por cerca de século e meio é entendido como um “período crítico, crucial na história da língua portuguesa” (Cardeira 2005: 292). É um período que simultaneamente se afasta das formas e construções do galego-português e seleciona e elabora aquelas que constituem a base do português clássico e do português moderno. As razões para a identificação e caracterização deste período foram amplamente apresentadas por Cardeira (2005), mas é de destacar o processo de elaboração a que a língua foi sujeita. Com a vitória em Aljubarrota, a dinastia de Avis consolida o seu poder a Sul, na cidade de Lisboa, que se desenvolve como centro urbano e de ação política, e ponto de partida para a expansão marítima. A classe burguesa, que apoiara o Mestre de Avis na defesa da independência, floresce, enquanto a antiga nobreza que apoiara Castela cai em desfavorecimento. O novo modelo é a corte de Avis e o dialeto centro-meridional do eixo Lisboa-Coimbra. É neste contexto que se verificam mudanças que se caracterizam pela “rejeição de marcas setentrionais, dialetais e antigas” (Castro 2006:154) e pela demarcação de diferenças com as línguas vizinhas, em especial o galego.

No presente artigo, iremos explorar algumas das mudanças observadas neste período da história da língua. A partir dos textos presentes no Corpus de Textos Antigos (CTA)¹ (CLUL 2015) e do trabalho já realizado no sentido de enriquecer os dados do corpus (Bico et al. 2022), foram selecionados dois dos fenómenos linguísticos –a síncope de *-d-* intervocálico e as terminações *-udo* e *-ido* nas formas de Particípio Passado– que caracterizam o Português Médio e o afastam do Português Antigo. A análise alargada destes fenómenos a um conjunto de sete textos terá dois objetivos. Em primeiro lugar, caracterizar o estado da língua de cada texto e confrontar os resultados obtidos à luz daquilo que se sabe sobre a datação do momento de escrita e do momento de cópia. Em segundo lugar, a análise a um maior conjunto de textos permite comparar os vários resultados entre si, culminando na criação de uma linha temporal.

Este artigo estrutura-se da seguinte forma: na Secção 2 apresenta-se o CTA e todo o trabalho já desenvolvido neste projeto; a Secção 3 faz uma apresentação sumária dos textos selecionados para análise, os contextos de produção e de cópia; a Secção 4 apresenta a metodologia adotada; na Secção 5 apresentam-se os resultados da análise da síncope de *-d-* intervocálico (5.1) e das terminações *-udo* e *-ido* nas formas de Particípio Passado (5.2); na Secção 6 discutem-se os resultados; a Secção 7 apresenta as principais conclusões e trabalho futuro.

¹ <http://teitok.clul.ul.pt/teitok/cta/index.php?action=home> [02-04-2025]

2. CORPUS DE TEXTOS ANTIGOS

O Corpus de Textos Antigos (CTA) é um projeto do Centro de Linguística da Universidade de Lisboa, iniciado em 2015. O corpus tem como objetivo tornar disponível edições de alta-fidelidade aos testemunhos dos textos, procurando ter uma intervenção editorial mínima. Neste sentido, são preservados todos os erros, as lacunas, bem como a grafia original dos testemunhos. Com este objetivo, o CTA procura fornecer dados que ajudem a análise linguística e o estudo da transmissão dos variados textos, além de dar acesso a textos até então inéditos. Os textos publicados no CTA deverão ter sido escritos ou traduzidos para o português até 1525², podendo os testemunhos ser posteriores a esta data. O CTA utiliza a plataforma TEITOK (Janssen 2016), uma ferramenta com uma interface amigável do utilizador, que permite a coexistência de informação linguística e informação textual.

Em fevereiro de 2024, o corpus é composto por 32 edições, existindo três textos que estão representados por mais do que um testemunho: o *Horto do Esposo*, com um manuscrito e três fragmentos; a *Vida de Santa Maria Egípcíaca*, com dois testemunhos; e a *Vida e Milagres de Santa Senhorinha de Basto*, com quatro testemunhos.

Com o intuito de enriquecer os dados já presentes no CTA, recorreu-se a ferramentas e métodos de Processamento de Linguagem Natural. Alguns dos textos do corpus foram selecionados para experiências de lematização e anotação de categorias morfossintáticas (part-of-speech). O presente trabalho utiliza a tokenização feita pelo TEITOK aquando da inserção dos textos na plataforma. Esta decisão teve por motivo a necessidade de não alterar os ficheiros XML de onde foram extraídos os dados e a consequente alteração das edições feitas por vários editores.

Primeiro, definiram-se as linhas orientadoras para a lematização e definiu-se o conjunto de etiquetas de anotação (*tagset*) para a anotação morfossintática e respetivos valores flexionais. Para além do princípio da preservação da tokenização feita pelo TEITOK, adotou-se um princípio de simplicidade na anotação, que tinha como objetivo a otimização do esforço inicial da anotação manual dos dados de treino. As linhas orientadoras para a lematização definem que cada lema deve corresponder, sempre que possível, à tradição lexicográfica, isto é: singular nas palavras variáveis em número; masculino nas palavras variáveis em género; infinitivo nos verbos, etc.. Por razões histórico-linguísticas, algumas formas são anotadas com palavras que já não fazem parte do léxico do Português Contemporâneo, como por exemplo as formas átonas dos determinantes possessivos femininos *ma*, *ta*, *sa* e os determinantes/pronomes *esto*, *esso*, *aquesto*, *aqueste*. Também está definido que teriam lemas distintos palavras que, apesar da mesma raiz etimológica, tivessem sofrido processos de evolução distintos, consoante o momento de entrada na língua, independentemente de terem conservado o mesmo significado (exemplo: *paço*/*palácio*). Foi definida a norma para a anotação dos pronomes pessoais, que seriam registados sob o lema *eu*, estando a informação morfossintática distinguida por meio das respetivas etiquetas (caso, pessoa, número, género).

Quanto à anotação morfossintática, adotou-se um conjunto de etiquetas originalmente desenvolvidas e aplicadas para o Português Contemporâneo (Eleutério et al. 1995; Ranchhod 2001; Ranchhod et al. 1999), adaptando-o às particularidades do Português Medieval. São contempladas nove categorias: nome, adjetivo, verbo, pronome, advérbio, preposição, conjunção, determinante e interjeição. No *tagset* incluem-se também subcategorias que

² O limite temporal (1525) deve-se à indexação presente na base de dados da Bibliografia de Textos Antigos Galegos e Portugueses (BITAGAP), que tem catalogados todos os textos e testemunhos que foram inicialmente escritos ou traduzidos para português, galego ou galego-português durante o período medieval.

descrevem com maior granularidade as palavras. Por exemplo, neste conjunto inclui-se a etiqueta para os nomes próprios, para os pronomes/determinantes possessivos, demonstrativos, indefinidos, etc. São também definidas as etiquetas para os valores flexionais de género (masculino/feminino), número (singular/plural), pessoa (1.^a/2.^a/3.^a). No caso dos verbos, um conjunto de 14 etiquetas identifica simultaneamente o Tempo e o Modo verbal de cada forma verbal; também nos pronomes pessoais se faz a identificação do caso (nominativo, acusativo, dativo, oblíquo, reflexivo) através de um conjunto de etiquetas próprias. Todas as marcas gráficas de pontuação têm como lema ‘_’ e como etiqueta ‘SENT’. Para as palavras que expressam a negação, como por exemplo *não*, *nada*, *nenhum*, é aplicada a etiqueta ‘neg’ que é assinalada como traço (*feature*). A atribuição de uma categoria morfossintática e dos respetivos valores flexionais de determinada forma está condicionada pelo contexto particular em que esta se insere.

Posteriormente, selecionou-se o texto do ms. A do *Horto do Esposo* (doravante HdE-A) para a tarefa de anotação manual com vista à criação de um conjunto de dados de treino para a anotação automática. Foram anotados cerca de 150 mil tokens, incluindo pontuação.

Conforme apresentado em Bico et. al (2022), com estes dados foi treinado um primeiro modelo de anotação, usando o TreeTagger (Schmid 1994, 1999), que por sua vez foi aplicado a um conjunto de dados proveniente de dois novos textos: os fragmentos D, C e E do *Horto do Esposo* (HdE-DCE) e o manuscrito G1 da *Vida e Milagres de Santa Senhorinha de Basto* (VMSSB-G1). O comando para a anotação automática dos textos é feito através da linha de comandos. Este comando gera um ficheiro XLSX com os dados verticalizados, isto é, um token por linha³. Para além do token, do lema e da etiqueta morfossintática, este ficheiro contém a identificação do token, o contexto à direita e à esquerda do token. A presença do contexto imediato do token permite a revisão manual dos resultados dados pelo sistema e a sua correção, sempre que necessário.

Na primeira experiência de anotação, os resultados de precisão variaram entre 67% e 90%. A precisão de 90% obtida em HdE-DCE é motivada pelo facto de se tratar de um conjunto de fragmentos do mesmo texto cujos dados foram utilizados para treinar o modelo. O resultado obtido em VMSSB-G1 (67%) evidencia a multiplicidade de novas formas gráficas, que, sendo desconhecidas do sistema, isto é, não estando presentes no conjunto de dados de treino, não foram corretamente identificadas. A título de exemplo, no campo da pontuação, VMSSB-G1 utiliza frequentemente a vírgula. No entanto, HdE-A não regista nenhuma ocorrência deste sinal, pelo que os dados de treino não contemplavam esta forma.

Para melhor serem utilizados, os dados anotados automaticamente foram sujeitos a uma fase de revisão manual. A revisão dos dados permitiu que os dois textos anotados na experiência 1 fossem incluídos no conjunto de dados utilizados para treinar um segundo modelo de anotação. O modelo 2 foi treinado com pouco mais de 160 mil tokens, tendo sido posteriormente aplicado a quatro outros textos: o ms. E da *Vida e Milagres de Santa Senhorinha de Basto* (VMSSB-E), o *Memorial da Infanta Santa Joana* (MISJ), o *Livro dos Mártires* (LdM) e a *História de mui nobre Vespasiano* (Vespasiano). Face à primeira experiência, a precisão aumentou, com os resultados a variar entre os 76% e os 82% para cada um dos diferentes textos. O aumento da precisão evidencia a noção elementar de que quantos mais forem os dados de treino, melhores serão os resultados obtidos. Os resultados desta experiência revelam ainda alguma dificuldade em identificar marcas de pontuação, afetando o desempenho do anotador. Também a entrada de novo léxico, em

³ Um dos objetivos futuros do projeto é a atualização dos ficheiros XML presentes no Corpus de Textos Antigos, de modo que estes incluam os dados linguísticos obtidos através da anotação automática.

particular nos textos que não têm nenhum outro testemunho nos dados de treino (isto é, em MISJ, Vespasiano e LdM) diminuiu os valores de precisão. Também foi possível verificar, particularmente em MISJ, que se encontram também palavras indevidamente segmentadas e, por isso, incorretamente tokenizadas. Todas estas instâncias são identificadas e anotadas de forma que os dados sejam inutilizados. A presença destes casos aumenta o número de tokens com lema e etiqueta morfosintática errados, diminuindo a precisão.

O total de dados anotados linguisticamente, no final das duas experiências, é de cerca de meio milhão de tokens (510.366 tokens).

3. CTA: TEXTOS SELECIONADOS

São sete os textos do CTA que estão lematizados e morfosintaticamente anotados. Nesta secção, apresentamos cada um deles, de modo a melhor contextualizá-los no período em que se inserem.

Oriunda do *scriptorium* do Mosteiro de Alcobaça, *Horto do Esposo* é uma obra originalmente escrita em português, mas cuja autoria é desconhecida. Terá sido escrita entre 1390 e 1407, sendo um dos textos mais antigos presentes no CTA. Conhecem-se dois testemunhos, o Alc. 198 e o Alc. 212, guardados na Biblioteca Nacional de Portugal, estando as digitalizações disponíveis online⁴. Conhecem-se ainda os fragmentos D, C e E provenientes de um códice do Mosteiro do Lorvão que foi desmembrado e reutilizado na encadernação de livros notariais. O testemunho que se encontra editado no CTA é o Alc. 198, datado de finais do século XIV (1390-1437). Recorrendo a *exempla*, a obra aborda os mais diversos assuntos de uma perspectiva moral, didática e ascética, recorrendo às *auctoritas* e à Sagrada Escritura para ilustrar os seus argumentos (Ferrero & Peixeiro 1993: 315-317).

Da *História de mui nobre Vespasiano* conhece-se apenas um incunábulo, impresso na oficina de Valentim Fernandes a 20 de abril de 1496, conforme indica o colofon. Julga-se que este romance terá descendido do romance francês *La Venjace Nostre Seigneur*, escrito entre os séculos XII e XIII. A tradução portuguesa terá sido realizada por intermédio da tradução castelhana. Esta relação estabelece-se pela existência particular aos testemunhos peninsulares (para além do incunábulo português, conhecem-se quatro testemunhos castelhanos). A narrativa centra-se em torno da figura do imperador Vespasiano que, para curar-se da lepra, procura relíquias vindas de Jerusalém. Após a sua cura milagrosa, Vespasiano parte para Jerusalém que conquista e destrói (Castro 1993: 311-312).

Originalmente escrito em castelhano por Bernardo de Briuega, o *Livro dos Mártires* é a terceira parte de um projeto de cinco livros, encomendado por D. Afonso X de Castela. De natureza hagiográfica, relata a história de vida de santos e mártires, com particular destaque para mártires romanos e os Padres do Deserto. No reinado de D. Dinis (1279-1325), o texto foi traduzido para português. Apesar de hoje não se conhecer nenhum manuscrito da época, esta ilação advém da identificação de formas típicas do português do século XIV na edição de 1513, cujo exemplar único se encontra transcrito no CTA (Sobral & Cardeira 2018).

O *Memorial da Infanta Santa Joana* narra a história da Infanta Joana (1453-1590), filha do rei D. Afonso V e da sua entrada e vida no Mosteiro de Jesus de Aveiro. Conhece-se um testemunho deste texto guardado na Biblioteca do Museu de Aveiro com a cota PT/MA/COD 9. Este testemunho conta com duas narrativas: a *Crónica da Fundação do Mosteiro de Aveiro* e o

⁴ As digitalizações dos códices alcobacenses estão disponíveis na página da Biblioteca Nacional Digital de Portugal (<https://bndigital.bnportugal.gov.pt/> [02-04-2025]).

Memorial da Infanta Santa Joana. Embora distintos, ambos os textos mostram a “preexistência de um plano conjunto de ação” (Sobral 2015:215). A autoria será feminina e dominicana aveirense, mas o nome em concreto é discutido: para Madahil (1939), terá sido Margarida Pinheiro, noviça no mosteiro com entrada em 1467; para Sobral (2015:215), terá sido Isabel Luís, irmã responsável pela instrução das noviças e copista no *scriptorium*. Sobre as balizas temporais, delimita-se o período entre 1513 e 1525 para o momento da escrita do texto, sendo o testemunho posterior a 1525.

A *Vida e Milagres de Santa Senhorinha de Basto* relata a história de Santa Senhorinha, abadessa portuguesa do século X. São conhecidos quatro testemunhos da versão portuguesa, todos editados no CTA. Para o estudo do presente artigo, porém, apenas dois são contemplados (os mss. G1 e E), por estarem linguisticamente anotados. A data de redação inicial do texto terá ocorrido entre 1248 e 1284. O testemunho G1 terá sido produzido entre 1620 e 1645; o testemunho E é uma cópia autógrafa de Torcato Peixoto de Azevedo, realizada entre 1692 e 1705. A análise estemática da tradição de VMSSB levada a cabo por Cruz (2018) concluiu que o testemunho G1 se encontra mais próximo do arquétipo, enquanto o testemunho E está dependente de um sub-arquétipo e terá sofrido contaminação de G1. O testemunho G1 é aquele que também está mais próximo do estado da língua do século XIII.

4. METODOLOGIA

O conjunto de textos selecionados perfaz um total de cerca de meio milhão de *tokens*. A identificação e a extração dos dados relevantes para a análise linguística tornam-se pouco exequíveis sem o recurso a ferramentas de pesquisa e filtragem automáticas. Neste sentido, a lematização e a granularidade da anotação morfosintática, que inclui os valores flexionais, facilitam o processo de extração destes dados.

Assim, para a análise da síncope de *-d-* intervocálico, a pesquisa focou-se na identificação de todos os tokens com a etiqueta ‘V’, relativa à categoria Verbo, e os valores flexionais ‘2p’ que indicam a 2.^a pessoa do plural. Esta pesquisa foi aplicada sobre todos os tempos e modos verbais em que o fenómeno acontece: no Presente (‘P’), no Pretérito Perfeito (‘J’), no Pretérito Imperfeito (‘I’), no Pretérito Mais-Que-Perfeito (‘M’) e no Futuro (‘F’) do Indicativo; no Presente (‘S’) e no Pretérito Imperfeito (‘T’) do Conjuntivo; e nos modos Condicional (‘C’) e Imperativo (‘Y’). Após este levantamento, foram identificadas todas as formas que continham *-d-* intervocálico no morfema e separadas daquelas que já não tinham este grafema.

Para recolha de todas as ocorrências das formas de Particípio Passado, todos os textos foram filtrados para a identificação de uma das seguintes etiquetas –‘V:Kms’, ‘V:Kfs’, ‘V:Kmp’, ‘V:Kfp’– nas quais ‘V’ representa a categoria Verbo, ‘K’ indica o tempo e o modo verbal em causa (Particípio Passado), ‘m’ e ‘f’ dão conta da flexão em género (masculino e feminino); e ‘s’ e ‘p’ (singular e plural) dão indicação da flexão em número. Entre todas as instâncias obtidas, foram apenas selecionadas as formas pertencentes a verbos com origem etimológica na 2.^a ou na 3.^a conjugações latinas (–ERE e –IRE), as únicas pertinentes para o estudo deste fenómeno.

5. RESULTADOS

5.1. Síncope de *-d-* intervocálico e resolução de hiato

No Português Antigo, o *-t-* intervocálico na terminação da 2.^a pessoa plural de todos os tempos do latim clássico na voz ativa sonorizou-se para *-d-* em todos os tempos verbais (ex.:

amatis > *amades*), exceto no Pretérito Perfeito. Posteriormente, o *-d-* intervocálico acaba por sofrer a síncope (ex.: *amades* > *ama-es*), daí resultando um encontro vocálico que é resolvido durante o Português Médio. Para Bechara, este é o “fenómeno balizador por excelência” (1991:70). Segundo a cronologia estabelecida por Cardeira (2005:185-186), até ao início do século XV, as *formas plenas*, isto é, aquelas que conservam o *-d-* intervocálico, são superiores às *formas sincopadas*, aquelas que evidenciam a síncope, e que ocorrem esporadicamente na escrita, mas poderiam ser mais frequentes no registo oral. A tendência inverte-se entre 1410 e 1430, e na segunda metade do século XVI, as formas plenas são raras e arcaicas.

O encontro vocálico resultante da síncope de *-d-* intervocálico é resolvido através da ditongação (ex.: *mandades* > *manda-es* > *mandais*) no Presente do Indicativo, no Imperativo da 1.^a e 2.^a conjugação, no Futuro do Indicativo, no Presente do Conjuntivo e no Pretérito-Imperfeito do Conjuntivo de todos os verbos. No Pretérito-Imperfeito, no Pretérito-Mais-Que-Perfeito e no Condicional ocorrem dois fenómenos: assimilação e ditongação (ex.: *mandavades* > *mandava-es* > *mandavees* > *mandaveis*). No Presente do Indicativo e no Imperativo dos verbos da 3.^a conjugação ocorre uma crase (ex.: *dormides* > *dormi-es* > *dormis*).

Atente-se agora nos dados obtidos nos textos anotados do CTA, relativamente à síncope de *-d-* intervocálico no morfema da 2.^a pessoa plural (Tabela 1).

Tabela 1. Total das formas plenas e das formas sincopadas (valores absolutos)

TEXTO	FORMAS PLENAS	FORMAS SINCOPADAS	TOTAL	PROPORÇÃO DE FORMAS SINCOPADAS
VMSSB-G1	21	1	22	4.5%
VMSSB-E	19	1	20	5.0%
LdM	750	81	831	9.7%
HdE-A	141	1	142	0.7%
HdE-DCE	2	0	2	0.0%
Vespasiano	20	126	146	86.3%
MISJ	3	28	31	90.3%
Total	956	238	1,194	19.9%

Em primeiro lugar, tanto HdE-A como os fragmentos HdE-DCE apresentam um resultado expectável, com uma esmagadora presença de formas plenas. Este resultado pode ser explicado sobretudo pelas datas de produção e cópia dos testemunhos: finais do século XIV, início do século XV (1390-1437). Aquando da cópia, as formas plenas ainda seriam aquelas mais frequentes e as forças da mudança ainda não se faziam sentir. Poder-se-á, também, admitir o conservadorismo do contexto de produção. O testemunho HdE-A é proveniente do *scriptorium* de Alcobaça. Como mostrado por Cardeira (2005: 180), nas *Vidas de Santos*, oriundas do mesmo *scriptorium* em data parcialmente coincidente (1431-1446) as formas plenas prevalecem sobre as formas sincopadas.

Os testemunhos G1 e E da VMSSB são cópias altamente conservadoras deste fenómeno. Para o copista do século XVII, as formas plenas seriam meros vestígios de um estado anterior da língua e já não seriam as mais frequentemente usadas. A preservação de formas plenas, em ambos os manuscritos, aponta, como conclui Cruz (2018:256) para o século XIII, época em que estas formas ainda predominavam, e, conseqüentemente, para o arquétipo destes testemunhos.

Os textos Vespasiano e MISJ já se encontram num momento da inversão da tendência, ou até de estabilização da nova variante sincopada, que representa 86,30% e 90,32% das formas, respetivamente.

Inesperadamente, o LdM apresenta a mais elevada presença de formas plenas, que representam 85,96%. Dada a data de produção do testemunho (1513) seria expectável o resultado inverso, ou seja, a prevalência das formas sincopadas e a raridade de formas plenas. Conforme notado por [Sobral e Cardeira \(2018:134\)](#), este resultado coloca o *Livro dos Mártires* “em época anterior a 1438 [...] em época muito próxima do *Orto do Esposo*”. [Cepeda \(1993:508\)](#) comprovou a existência de um testemunho manuscrito português do século XV, do qual se conhece hoje apenas um fragmento, datando a tradução do texto nele copiada do final do reinado de D. Dinis (1279-1325). [Sobral e Cardeira \(2018: 138-139\)](#), através da análise de alguns fenómenos linguísticos, entre eles a síncope de *-d-* intervocálico, concluem que, apesar de já apresentar traços correspondentes ao estado da língua no início do século XVI, o impresso de 1513 conserva traços arcaizantes da língua. A preservação destes traços permite: i) confirmar a datação da tradução do *Livro dos Mártires* no reinado de D. Dinis; ii) admitir a existência de uma cópia do século XV que terá sido usada como original de impressa.

Para a análise das novas regras fonológicas, isto é, as regras fonológicas provocadas pela síncope de *-d-* intervocálico, foi utilizado um subconjunto dos dados anotados. Do total do conjunto dos sete textos, foram selecionados apenas os tempos verbais que registavam mais de 100 ocorrências. Assim, este subconjunto de dados contempla formas do Presente Indicativo, do Presente Conjuntivo e do Imperativo. Feita a distinção entre formas plenas e formas sincopadas, obtiveram-se 55 formas sincopadas do Presente Indicativo, 64 do Presente do Conjuntivo e 92 do Imperativo. Em valores absolutos, estes números poderão não revelar muita informação, mas uma análise mais pormenorizada das várias grafias por tempo e por conjugação verbal poderá ajudar a fazer sentido dos fenómenos observados. A [Tabela 2](#) mostra a distribuição das diferentes terminações gráficas das formas sincopadas por texto, por tempo e conjugação verbal.

Tabela 2. Diferentes terminações por modelo de conjugação e por texto/testemunho

	PRESENTE INDICATIVO			PRESENTE CONJUNTIVO			IMPERATIVO		
	-AR	-ER	-IR	-AR	-ER	-IR	-AR	-ER	-IR
LdM	-ais (1) -aaes (2) -aes (4)	-ees (7) -eys (2)	-is (1) -ijs (1)	-ees (3) -es (1)	-aes (4) -aaes (2)	-	-ay (18) -aj (1) -ai (1) -ae (2)	-ey (10)	-ij (1) -ye (2) -y (1) -i (2)
MISJ	-	-ees (6)	-	-ees (3)	-aes (4)	-	-aae (3) -ay (4) -ae (3)	-é (1)	-ij (1)
Vespasiano	-aaes (1)	-ees (17) -es (2) -eais (1)	-	-ees (16)	-aes (15) -aaes (2) -aes (1) -ays (1)	-aaes (1)	-ay (7) -ai (1) -ae (12) -aae (1)	-ee (20)	-
VMSSB-E	-	-	-is (1)	-	-es (2)	-	-	-	-

	PRESENTE INDICATIVO			PRESENTE CONJUNTIVO			IMPERATIVO		
	-AR	-ER	-IR	-AR	-ER	-IR	-AR	-ER	-IR
VMSSB-G1	–	–	-is (1)	–	-es (2)	–	–	–	–

A duplicação da vogal na terminação *-aaes* no Presente do Indicativo da 1.^a conjugação (ex.: *assaaes*, Vesp.; *firmaaes*, LdM) e no Presente do Conjuntivo da 2.^a conjugação (ex.: *façaaes*, Vesp.; *creaaes*, LdM), encontrada no LdM e no Vespasiano poderá denotar a tonicidade da vogal <aa>, com <e> a representar a semi-vogal. Mais frequentes são as formas *-aes* e *-ees*, enquanto as formas *-ais* e *-eys* no Presente do Indicativo e a forma *-ays* no Presente do Conjuntivo são mais raras. Esta oscilação entre <e>, <y> e <i> para representar a semi-vogal do ditongo é encontrada desde o início do século XV (Cardeira 2005:201). Neste conjunto de dados, é no modo Imperativo que esta oscilação se torna mais evidente e onde é possível estabelecer comparações entre um maior número de textos. Nos textos LdM, MISJ e Vespasiano, todos elaborados no decorrer do século XV ou posteriormente, observa-se a alternância de grafias *-ay*, *-aj*, *-ai*, *-ae* e *-aae*, com estas duas últimas a terem pouca representatividade em LdM e MISJ.

Na 3.^a conjugação, todas as formas encontradas do Presente do Indicativo já apresentam a crase (*fugis*, HdE-A; *dormis*, VMSSB-G1, VMSSB-E, LdM; *sentijs*, LdM), mas um conjunto de quatro ocorrências não constitui uma base suficiente para uma análise mais aprofundada.

No Imperativo, a forma *destruye* com duas ocorrências no LdM poderá ser equiparada à forma *abrij* se se entender a presença dos dois elementos vocálicos como marcação da vogal tónica, tendo a crase já acontecido. A inexistência de outras formas semelhantes e a pouca frequência de verbos da 3.^a conjugação com formas sincopadas não permitem tirar conclusões mais seguras. No MISJ, encontra-se a terminação de ocorrência única *-é*, na forma *fazé*. Pode-se supor a hipótese de esta forma ser um ensaio de uma crase que não vingou.

Apenas com um maior conjunto de dados, obtidos através da inclusão de mais e maiores textos, poderá a análise da representação de novas regras fonológicas obter conclusões mais seguras. Conclusões mais seguras sobre a representação das novas regras fonológicas poderão ser obtidas. Neste sentido, a utilização de um corpus anotado traz a vantagem do fator tempo, uma vez que reduz em muito o tempo que o investigador utiliza na procura por estas formas. Acresce também o facto de a utilização de ferramentas e métodos de PLN permitir uma maior facilidade de pesquisa e extração sistemática e exaustiva de informação de um largo conjunto de dados, possibilitando a comparação entre textos e através do tempo.

5.2. Particípios em *-udo* / *-ido*

No Português Antigo, conservavam-se três terminações de particípios passados: *-ado*, *-udo* e *-ido*. A divergência relativamente ao latim deve-se à fusão da 2.^a e 3.^a conjugações latinas ainda durante o latim vulgar (Cardeira 2005:203). Deste modo, “só aos verbos com origem na 4.^a conjugação latina (*-IRE*, *salīre*) correspondia o Particípio Passado em *-ido* (*-ITU*); a terminação do Particípio dos verbos de vogal temática *e* (da segunda e da terceira conjugações latinas, respetivamente *-ĒRE* e *-ĒRE*) era *-udo*” (Cardeira 2005:204). Assim era no Português Antigo até que a terminação *-udo* começa a desaparecer e a ser substituída por *-ido*, encontrando-se formas esporádicas já no final do século XIII. Durante o século XIV, a alternância entre as formas seria mera variação. No século XV, a variante *-ido* é acolhida pela

corte e adquire prestígio, permitindo a sua rápida expansão (Cardeira 2005:214). Na Tabela 3 mostram-se os dados obtidos no conjunto dos sete textos selecionados.

Em HdE-A, a alternância entre formas é bastante visível: 56,70% das formas de verbos com origem na 2.^a e na 3.^a conjugação latina terminam em *-ido* e 43,30% em *-udo*. A proporção de formas *-udo* obtidas nesta recolha é superior àquela apresentada por Carvalho (1996:92), que indica uma representação de 35% destas formas no HdE-A. Entre as formas terminadas em *-udo*, 30,95% (equivalentes a 26 formas) correspondem ao verbo *ter* ou verbos derivados (*conter*, *manter*, *reter*, *suster*). O verbo *ter* e todos os verbos dele derivados apresentam apenas as terminações em *-udo*. O mesmo sucede para os verbos *ler* e *haver* (5 e 4 ocorrências, respetivamente). São 12 os verbos com *e latino* nos quais se regista variação: *aprimir* (=oprimir), *confundir*, *crescer*, *derreter*, *dever*, *entender*, *esconder*, *meter*, *perceber*, *resolver*, *saber* e *vencer*. Em seis destes verbos, a variante *-ido* é mais frequente, representando entre 60% e 96% das formas dos respetivos verbos. Destaque-se o verbo *esconder*, que regista 24 ocorrências de particípio terminado em *-ido* e apenas uma com o etimológico *-udo*.

Tabela 3. Total de formas de particípio passado em verbos da 2.^a e 3.^a conjugações latinas

TEXTO	FORMAS -UDO	FORMAS -IDO	TOTAL	PROPORÇÃO DE FORMAS -IDO
VMSSB-G1	4	4	8	50%
VMSSB-E	3	6	9	67%
LdM	46	697	743	94%
HdE-A	84	110	194	57%
HdE-DCE	0	0	0	0
Vespasiano	0	40	40	100%
MISJ	0	127	127	100%
Total	137	984	1,121	88%

No LdM, são apenas 6,19% os verbos que têm a forma de Particípio terminada em *-udo*, e que correspondem a 46 formas. Estas formas distribuem-se por 14 verbos. Em dez destes verbos há variação na terminação, sendo *-ido* predominante em sete verbos (*esconder*, *ler*, *meter*, *receber*, *saber* e *vencer*). Os verbos *crer* e *haver* apresentam igual distribuição entre terminações (5) e, nos verbos *descrer*, *conter* e *ter* regista-se um maior número de formas *-udo* (*descrer*: 4 *-udo*, 1 *-ido*; *conter*: 7 *-udo*, 1 *-ido*; *ter*: 18 *-ido*). Em dois verbos (*descer* (1) e *manter* (3)) não ocorre variação, sendo a terminação sempre *-udo*.

Textos publicados no mesmo espaço temporal que LdM, como o *Flos Sanctorum* (1513) e o *Cancioneiro Geral* de Garcia de Resende (1516) contêm alguns vestígios desta terminação, quer por razões rimáticas, quer por força da cristalização das formas (ex.: verbo *ter* - *teudo*). Se se admitir que as formas dos Particípios Passados do verbo *ter* e verbos derivados possam ter cristalizado devido ao frequente uso, as ocorrências da terminação *-udo* poderão ser consideradas meros vestígios de uma alternância que no início do século XVI já estava estabilizada. Ao mesmo tempo, a presença da terminação *-udo* no LdM, ainda que diminuta no contexto geral do texto, aponta novamente para o manuscrito do século XV que serviu de base ao original de imprensa. Por meio de comparação com outros textos, Sobral e Cardeira (2018: 135) colocam o LdM entre 1416 e 1438. Ficará, porém, por responder, por ausência

dos testemunhos que permitem estabelecer a comparação, até que ponto o copista do século XV atualizou linguisticamente a tradução do século XIII.

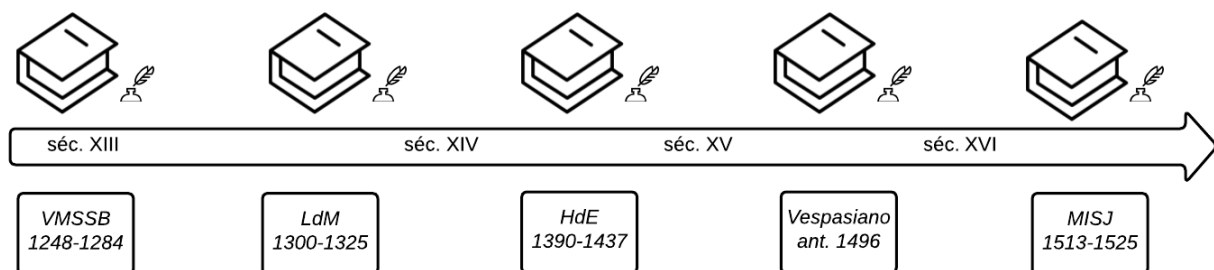
No *Vespasiano* e no *MISJ* não se encontraram formas de Particípio terminadas em *-udo*, somente em *-ido*. O incunábulo do *Vespasiano* data de 1496 e o *Memorial da Infanta Santa Joana* é redigido no primeiro quartel do século XVI, evidenciando que o processo de substituição já estava estabilizado.

Na VMSSB-G1, texto do século XIII em testemunho do século XVII (1620-1645), às formas *perdudo* (2), *estendudo* (1), *estendudos* (1), *metido* (1), *offrecido* (1), já identificadas por Cruz (2018:275), acresce a forma *escondidas* (2). São, assim, as 8 formas de verbos com origem etimológica na 2.^a e na 3.^a conjugações latinas, cujas terminações de particípio passado se distribuem em igual proporção. Com base nos seus resultados, Cruz (2018:275.) obtém uma proporção de 66,7% de formas em *-udo* e 33,3% em *-ido*. As evidências, no entanto, são escassas para se tirar conclusões mais significativas: nem no momento de redação do apógrafo no século XIII, nem no momento da cópia, no século XVII, existia a alternância de formas. Se, no século XIII, já fosse possível encontrar formas em *-ido*, estas eram esporádicas e raras; no século XVII, as formas *-udo* já teriam saído por completo da gramática dos falantes. A presença, portanto, das formas em *-udo* atesta este estado anterior da língua e as formas *-ido* podem ser resultado da interferência da língua do copista seiscentista. O testemunho E de VMSSB contrapõe três formas *-udo* e seis em *-ido*. Destas últimas, uma é exclusiva a este testemunho (*offendido*), a outra é a atualização de uma das ocorrências de *perdudo* em G1 que em E figura como *perdido*. As ilações retiradas são, assim, semelhantes àquelas obtidas na análise de VMSSB-G1.

6. PARA UM MELHOR CONHECIMENTO DOS TEXTOS

Dado aquilo que se sabe sobre a datação da produção do texto, a linha cronológica criada pelo conjunto dos sete textos anotados poderia ser representada através da Figura 1. O primeiro texto escrito terá sido a *Vida e Milagres de Santa Senhorinha de Basto*, ainda no século XIII, seguindo-se a tradução portuguesa do *Livro dos Mártires*, durante o reinado de D. Dinis, e o *Horto do Esposo* durante o século XIV; no final deste século terá sido escrito o *Vespasiano*; e só durante o primeiro quartel do século XVI é que o *Memorial da Infanta Santa Joana* terá sido escrito.

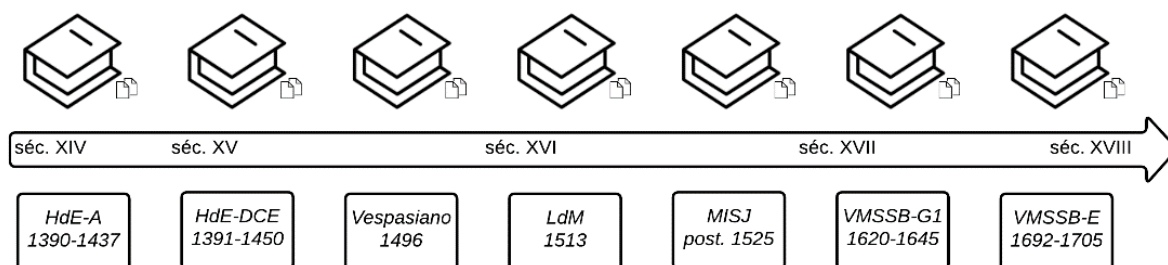
Figura 1. Cronologia das datas de produção dos textos selecionados.



Porém, os testemunhos presentes no *corpus* não são originais, mas antes cópias mais ou menos distantes do original. O alinhamento cronológico por data dos testemunhos é, por isso, diferente (Figura 2). Por se distanciarem do original, estes testemunhos trazem a marca da

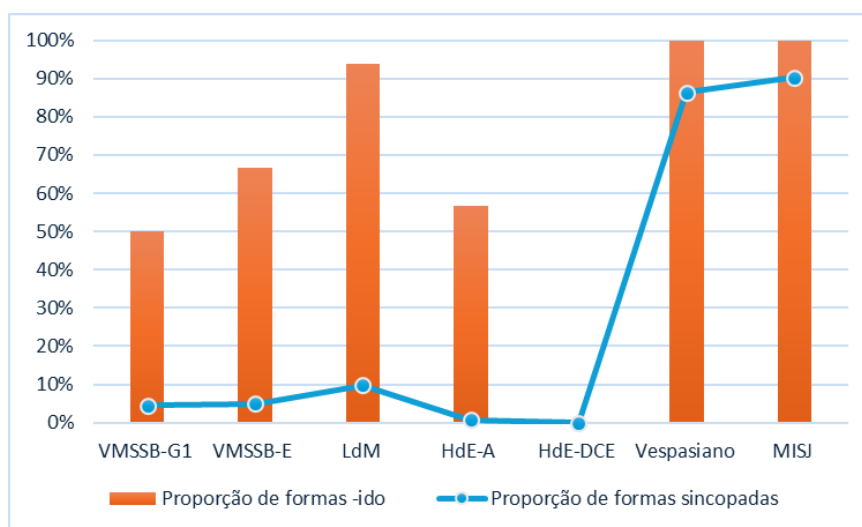
transmissão dos textos, quer através de sucessivas atualizações linguísticas, que atestam a evolução da língua, quer através da inserção, consciente ou não, de erros.

Figura 2. Cronologia das datas de cópias dos testemunhos.



No campo da História da Língua, em particular do Português Médio, observámos dois fenómenos caracterizadores deste período: a síncope de *-d-* intervocálico no morfema de flexão da 2.^a pessoa plural e as formas de Particípio Passado de verbos com origem etimológica na 2.^a e 3.^a conjugações latinas. A conjugação de todos os resultados (Figura 3) permite tecer alguns comentários face ao estado da língua de cada texto relativamente ao momento de produção e ao momento da cópia.

Figura 3. Total percentual dos resultados obtidos nos dois fenómenos.



Em primeiro lugar, HdE-A, Vespasiano e MISJ são os testemunhos cujo estado da língua presente no texto mais se aproxima do estado da língua do momento da produção. No caso do HdE-A, o texto é escrito na viragem do século XIV para o século XV; no Vespasiano e no MISJ, na viragem do século XV para o século XVI. A quase totalidade de formas da 2.^a pessoa plural no HdE-A são formas plenas e há uma alternância relativamente equiparada entre as terminações *-udo* (57,59%) e *-ido* (42,41%) nas formas de Particípio Passado. A combinação destes fatores, coloca o HdE-A mais recuado na linha temporal aqui ensaiada. O mesmo sucede para os fragmentos HdE-DCE. Por sua vez, e em sentido inverso, o estado da língua do Vespasiano e do MISJ, pelo que é observável nos dois fenómenos analisados, coloca estes

testemunhos em pontos mais avançados na linha cronológica. A presença de formas plenas é minoritária em ambos os textos (9,68% em MISJ e 13,70% no Vespasiano) e nenhum deles apresenta formas de Particípio Passado terminadas em *-udo*.

Por um lado, os dois testemunhos da *Vida e Milagres de Santa Senhorinha de Basto* atestam a língua falada em Portugal em meados do século XIII, principalmente através da preservação do *-d-* intervocálico no morfema de pessoa-número na flexão verbal. Por outro, como evidenciado aquando da análise das formas de Particípio Passado, aquelas terminadas em *-ido* não são próprias do século XIII, nem aquelas terminadas em *-udo* próprias do século XVII. É observável, então, uma atualização linguística que o copista consciente ou inconscientemente fez. Curiosa é, porém, a diferente atitude perante os dois fenómenos. Perante a síncope de *-d-* o copista seiscentista é altamente conservador, mas quanto ao Particípio Passado deixa mais facilmente o seu estado da língua permear o trabalho de cópia. É possível que o copista tenha conservado as formas com *-d-* para manter o carácter arcaizante do texto, atualizando, por outro lado, talvez de forma inconsciente, as terminações do Particípio Passado, já completamente anacrónicas.

Os resultados obtidos no LdM apontam em direções opostas. Contrariamente ao que seria expectável para um testemunho do início do século XVI, o LdM apresenta ainda uma considerável representação de formas plenas, que conservam o *-d-* intervocálico (85%). Simultaneamente, as formas sincopadas apontam já para a resolução do encontro vocálico provocado pela síncope, quer através da duplicação da vogal tónica, quer pela utilização de vários elementos gráficos para representar o elemento menos vocálico do ditongo. Se este último aspeto aproxima o testemunho do momento da sua produção, o conservadorismo nas formas plenas afasta-o do século XIV, aproximando-o mais dos séculos XIII e XIV e, conseqüentemente, do momento da tradução do texto para o português. Conforme anteriormente dito, terá existido um manuscrito do século XV, feito a partir da tradução trecentista, que terá servido como original de imprensa do incunábulo de 1513 (Cepeda 1993; Sobral & Cardeira 2018). Deste modo, o incunábulo atesta a língua de quatrocentos, apontando para uma fase mais recuada ainda. Na variação das terminações do Particípio Passado, o LdM apresenta apenas 6,85% de formas em *-udo*, sendo muitas delas pertencentes do verbo *ter* ou verbos derivados. A presença relativamente residual destas formas aponta já para uma estabilização da variação. Ficaré por responder, porém, em que medida terá o incunábulo de 1513 atualizado estas formas relativamente ao manuscrito do século XV.

Assim, entre todos os textos, os testemunhos do HdE (A e DCE), o incunábulo do Vespasiano e o testemunho de MISJ serão aqueles cujo estado da língua no momento de cópia mais se aproxima daquele no momento de produção. Estes quatro testemunhos colocar-se-iam em pontas opostas da linha temporal: os testemunhos A e DCE do *Horto do Esposo* revelam a língua falada no final do século XIV e nos inícios do século XV; Vespasiano e MISJ apontam já para a língua de Quinhentos, no advento do Português Clássico. O testemunho do LdM é ambivalente: se, por um lado, apresenta traços que o aproximam do momento de produção do original (séc. XIII-XIV), por outro lado, outros traços o aproximam do momento de produção.

Se a presença residual de formas de Particípio Passado em *-udo* podem evidenciar a estabilização da inversão da tendência no início do século XVI, a elevada frequência de formas plenas não é coerente com a língua de 1513, mas antes com a língua de um ou dois séculos antes. Os testemunhos G1 e E da *Vida e Milagres de Santa Senhorinha de Basto* têm comportamento semelhante, com os traços mais modernos a serem reflexo da influência linguística dos seus copistas.

7. CONCLUSÕES

Neste artigo, procurou-se aprofundar o conhecimento de uma seleção de sete textos presentes no *Corpus de Textos Antigos* (CTA) relativamente a dois fenómenos caracterizadores do Português Médio, período da história do português que decorre desde meados do século XIV até aos inícios do século XVI. O conjunto dos sete textos selecionados perfaz um total de cerca de meio milhão de tokens que se encontram lematizados e anotados morfossintaticamente. Recorrendo a esta anotação foi, assim, possível extrair sistematicamente do corpus todas as instâncias relevantes para o estudo da síncope de *-d-* intervocálica no morfema da 2.^a pessoa do plural e da variação entre *-ido* e *-udo* nas terminações do Particípio Passado nos verbos com origem etimológica na 2.^a e 3.^a conjugações latinas. Desta recolha, obtiveram-se 1,194 formas verbais correspondentes à 2.^a pessoa do plural, repartindo-se entre 956 formas plenas e 238 formas sincopadas; e 972 formas de Particípio Passado.

A análise subsequente procurou caracterizar o estado da língua nos vários testemunhos face à datação e cronologia dos vários fenómenos. Quanto à síncope de *-d-*, HdE-A, MISJ e Vespasiano obtiveram resultados expectáveis face à data de cópia dos testemunhos: HdE-A apresenta uma esmagadora presença de formas plenas; MISJ e Vespasiano evidenciam já a inversão da tendência, se não mesmo a sua estabilização. Apesar de serem os testemunhos mais tardios, VMSSB-G1 e VMSSB-E são conservadores na cópia, o que, por um lado, mostra o zelo dos seus copistas durante o processo da cópia. Por outro lado, estes testemunhos atestam a existência do arquétipo duocentista. Por sua vez, LdM apresenta os resultados mais inesperados, com a prevalência das formas plenas sobre as formas sincopadas. Este resultado parece comprovar a existência de um testemunho anterior, datado do século XV, que teria usado um testemunho anterior do século XIII, onde a alternância entre formas não se faria sentir. Não obstante, as formas sincopadas presentes apontam já para um período da evolução da língua no qual os hiatos resultantes da síncope de *-d-* já estariam resolvidos. No que diz respeito às terminações do Particípio Passado, HdE-A apresenta uma alternância entre as formas, que evidencia o período em que as duas formas ainda coexistiam em variação; em LdM, as formas *-udo* são percentualmente reduzidas, podendo admitir-se o seu uso em formas cristalizadas ou ainda como vestígio do manuscrito do século XV; no Vespasiano e no MISJ não há presença das formas *-udo*, apontando, assim, para o momento em que a mudança já estava estabilizada. Quanto aos dois testemunhos de VMSSB (G1 e E), a presença de formas em *-ido* não se coadunam com o arquétipo duocentista, sendo, por esta razão, marcas da influência da língua do século XVII dos copistas.

A comparação destes dois fenómenos no conjunto dos dados anotados permitiu a comparação do estado da língua representada nos textos face ao estado da língua do momento de produção e ao momento de cópia. Esta análise evidencia a importância da análise linguística para o conhecimento da história da transmissão dos textos. A presença ou a ausência de determinadas particularidades linguísticas permite desvendar o trilho que determinado texto percorreu até chegar aos dias de hoje. Como demonstrado no LdM, cujo testemunho data do início do século XVI, a presença de formas atípicas a este tempo põe a descoberto a existência de um manuscrito contemporâneo de tais formas. Do ponto de vista linguístico, esta análise ajuda também a compreender e a datar melhor a progressão dos fenómenos linguísticos.

O recurso a dados anotados e a sua extração automática possibilita a recolha sistemática e exaustiva de uma maior quantidade de dados num menor período de tempo e uma análise comparativa em maior escala. Se é verdade que o mesmo esforço pode ser feito manualmente,

o recurso a ferramentas e a métodos (semi-)automáticos otimiza o tempo utilizado, libertando o tempo do investigador para a análise, permitindo-lhe colocar perguntas talvez até agora não colocadas. Este trabalho não quis ser exaustivo no estudo dos fenómenos linguísticos caracterizadores do Português Médio. Procurou antes pôr em evidência a utilidade que um corpus anotado e ferramentas de extração automática da informação podem ter em estudos desta envergadura. Como trabalho futuro, estará a extensão da análise a outros textos do CTA e a outros fenómenos do Português Médio. Por exemplo, a análise da convergência das terminações nasais em *-ão*, que abundam no português e ostentam múltiplas formas gráficas, beneficiará de uma abordagem semelhante de extração automática dos dados.

Financiamento

A investigação para este trabalho foi financiada por Fundos Nacionais da Fundação para a Ciência e a Tecnologia: J. Baptista e F. Batista (INESC-ID Lisboa, proj. ref UIDB/50021/2020), E. Cardeira (Faculdade de Letras da Universidade de Lisboa, Centro de Linguística da Universidade de Lisboa, proj. ref. UIDP/00214/2020) e M.I. Bico pela Bolsa de Doutoramento (proj. ref. UI/BD/152806/2022).

Referências bibliográficas

- Almeida, Maria Inês. s.d. *História de mui nobre Vespasiano*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M1008T1008.xml> [02-04-2025].
- Bechara, Evanildo. 1991. As fases da língua portuguesa escrita. Em Dieter Kremer (ed.), *Actes du XVIIIe Congrès International de Linguistique et de Philologie Romanes*, vol. III. 68-76. Tübingen: Max Niemeyer Verlag.
- Bico, Maria Inês, Jorge Baptista, Fernando Batista & Esperança Cardeira. 2022. Early Experiments on Automatic Annotation of Portuguese Medieval Texts. Em Gianmaria Silvello, Oscar Corcho, Paolo Manghi, Giorgio Maria Di Nunzio, Koraljka Golub, Nicola Ferro & Antonella Poggi G. Silvello (eds.), *Linking Theory and Practice of Digital Libraries. TPD L 2022. Lecture Notes in Computer Science*, vol 13541. Cham: Springer, Cham. https://doi.org/10.1007/978-3-031-16802-4_44 [02-04-2025].
- Cardeira, Esperança. 2005. *Entre o Português Antigo e o Português Clássico*. Lisboa: Imprensa Nacional – Casa da Moeda.
- Carvalho, Maria José Simões Pereira de. 1996. *Do Português Arcaico ao Português Moderno. Contributos para uma nova proposta de periodização*. [Dissertação de Mestrado]. Coimbra: Faculdade de Letras da Universidade de Coimbra.
- Castro, Ivo. 1993. História de Vespasiano. Em Giulia Lanciani & Giuseppe Tavani (coords.), *Diccionario da Literatura Medieval Galega Portuguesa*. 311-312. Lisboa: Caminho.
- Cepeda, Isabel Vilares. 1993. Os «Quarenta Mártires de Sebaste». Um testemunho manuscrito do século XV em português. *Theologica* 28(2), 507-514. <https://doi.org/10.34632/theologica.1993.11938> [02-04-2025].
- CLUL. 2015. *Corpus de Textos Antigos até 1525*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=home> [02-04-2025].

- Cruz, Marta. 2018. *A vida de Santa Senhorinha de Basto em português: estudo estemático e linguístico*. [Dissertação de Mestrado]. Lisboa: Faculdade de Letras da Universidade de Lisboa. <https://repositorio.ulisboa.pt/handle/10451/34477>. [02-04-2025].
- Cruz, Marta. s.d. *Vida e Milagres de Santa Senhorinha de Basto – E*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M5602T12967.xml>. [02-04-2025].
- Eleutério, Samuel., Ranchhod, Elisabete Marques, Freire, Helena, & Baptista, Jorge. 1995. A System of Electronic Dictionaries of Portuguese. *Lingvisticae Investigationes* XIX (1), 57–82.
- Ferrero, Ana Diaz & Horácio Peixoto. 1993. Horto do Esposo. Em Giulia Lanciani & Giuseppe Tavani (coords.), *Diccionario da Literatura Medieval Galega Portuguesa*. 315-317. Lisboa: Caminho.
- Firmino, Jessica & Cristina Sobral. s.d. *Memorial da Infanta Santa Joana*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M1093T6287.xml>. [02-04-2025].
- Franco, Patrícia & Cristina Sobral. s.d. *Horto do Esposo – A*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M1114T1082.xml>. [02-04-2025].
- Janssen, Maarten. 2016. TEITOK: Text-Faithful Annotated Corpora. Em Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, & Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. 4037–4043. European Language Resources Association (ELRA). <https://aclanthology.org/L16-1637>. [02-04-2025].
- Madahil, António Gomes da Rocha. 1939. *Crónica da fundação do Mosteiro de Jesus, de Aveiro, e memorial da infanta santa Joana: filha del Rei Dom Afonso V (códice quinhentista)*. Aveiro: Edição de Prof. Francisco Ferreira Neves.
- Ranchhod, Elisabete Marques. 1999. Construções com Nomes Predicativos na Crónica Geral de Espanha de 1344. Em Isabel Hub Faria (ed.), *Lindley Cintra. Homenagem ao Homem, ao Mestre e ao Cidadão*. 667–682. Lisboa: Edições Cosmo / Faculdade de Letras da Universidade de Lisboa.
- Ranchhod, Elisabete Marques. 2001. O uso de dicionários e de autómatos finitos na representação lexical das línguas naturais. Em Elisabete Marque Ranchhod (ed.), *Tratamento das Línguas por Computador: Uma introdução à Linguística Computacional e suas aplicações*. 13–47. Lisboa: Editorial Caminho.
- Schmid, Helmut. 1994. Probabilistic part-of-speech tagging using decision trees. Em *Proceedings of International Conference on New Methods in Language Processing*. 44-49. Stuttgart: Association for Computational Linguistics.
- Schmid, Helmut. 1999. Improvements in Part-of-Speech Tagging with an Application to German. Em Susan Armstrong, S. Kenneth Church, Pierre Isabelle, Sandra Manzi, Evelyne Tzoukermann & David Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora*, volume 11. 13-25. Dordrecht: Kluwer Academic Publishers.
- Sobral, Cristina & Esperança Cardeira. 2018. O Livro dos Mártires de Bernardo de Brihuega: dois séculos de leitura em português. *Estudos de Lingüística Galega* 10, 129-141. <https://doi.org/10.15304/elg.10.4613>.

- Sobral, Cristina. 2015. A Vida da Princesa Santa Joana de Portugal: hipóteses de autoria. *Revista de Literatura Medieval* 27, 213-224. <http://hdl.handle.net/10017/28060>.
- Sobral, Cristina. s.d. *Horto do Esposo – DCE*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M3397T1082.xml> [02-04-2025].
- Sobral, Cristina. s.d. *Vida e Milagres de Santa Senhorinha de Basto – G1*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=M1614T12967.xml> [02-04-2025].
- Sonsino, Ana, Marta Cruz & Cristina Sobral. s.d. *Livro dos Mártires*. <http://teitok.clul.ul.pt/teitok/cta/index.php?action=file&id=2265.xml> [02-04-2025].