laborhistórico

Revista LaborHistórico
ISSN 2359-6910
v.9, n.1, e52234, 2023
DOI: 10.24206/lh.v9i1e52234

Artigo

Planear a normalização automática: tipologia de variação gráfica do *corpus* das *Memórias Paroquiais* (1758)

Planning spelling normalization: typology of graphic variation in the *Parish Memoirs* (1758) *corpus*

Helena Freire Cameron

Instituto Politécnico de Portalegre. Portalegre, Portugal
Universidade de Évora, Centro Interdisciplinar de História, Culturas e Sociedades (CIDEHUS), Portugal
helenac@ipportalegre.pt

Fernanda Olival

Universidade de Évora, Centro Interdisciplinar de História, Culturas e Sociedades (CIDEHUS). Évora, Portugal
mfo@uevora.pt

Renata Vieira

Universidade de Évora, Centro Interdisciplinar de História, Culturas e Sociedades (CIDEHUS). Évora, Portugal
renatav@uevora.pt

Recebido: 30/04/2022
Aceito: 27/03/2023

Como citar:
CAMERON, Helena Freire; OLIVAL, Fernanda; VIEIRA, Renata. Planear a normalização automática: tipologia de variação gráfica do *corpus* das *Memórias Paroquiais* (1758). *Revista LaborHistórico*, v.9, n.1, e52234, 2023. doi: https://doi.org/10.24206/lh.v9i1e52234

Resumo

No que respeita a fenómenos linguísticos, as Humanidades Digitais são hoje imprescindíveis para estudos sobre *corpora* textuais de grandes dimensões, em que a transformação de textos em dados processáveis requer um tratamento multidisciplinar. Neste artigo iremos apresentar uma abordagem em Humanidades Digitais, aplicada a um *corpus* textual português do século XVIII, reunido a partir de um conjunto documental de elevado valor histórico-patrimonial conhecido como as *Memórias Paroquiais* (1758). Dar-se-á conta de algumas características

da constituição do *corpus*, de questões relativas à variação gráfica reconhecida nos textos, propondo-se uma tipologia da variação com vista ao estabelecimento de uma futura automatização da normalização deste conjunto textual.

Palavras-chave

Humanidades Digitais, Fronteiras disciplinares, Português. século XVIII, Variação linguística, Memórias Paroquiais.

Abstract

Digital Humanities are now essential for studies on large-scale textual *corpora*, where the transformation of text into processable data regarding linguistic phenomena requires a multidisciplinary treatment. In this article we will present an approach in Digital Humanities, which was applied to a Portuguese textual *corpus* from the 18th-century, gathered from a set of documents known as *Memórias Paroquiais* ["*The Parish Memoirs*"], with high historical and heritage value. We will highlight some *corpus* constitution characteristics, questions concerning the expressive spelling variation perceived in the texts. We propose a typology towards a future automatic normalization of this textual *corpus*.

Keywords

Digital Humanities. Disciplinary frontiers. Portuguese. 18th-century. Linguistic variation. Memórias Paroquiais.

Introdução

As Humanidades Digitais (HD) impõem-se, cada vez mais, como um novo campo científico, combinando abordagens tradicionais em História, Paleografia, Linguística ou Filologia com processamento computacional (Edmond, 2020), (Schreibman, 2004). O património textual de épocas pretéritas consegue ser transportado para a sociedade digitalizada do século XXI, tornando-o globalmente acessível não só a estudiosos de história e de linguística, como também disponibilizando estes textos a outras áreas do saber, que dele podem fazer uso e fruí-lo (EU, 2021).

As Humanidades Digitais constituem-se, de modo inequívoco, como uma nova área do conhecimento fundada de raiz numa abordagem multidisciplinar, facilitando o acesso, estudo, processamento e interligação de documentos textuais oriundos de vários períodos históricos e linguísticos.

Quando falamos de Humanidades digitais é preciso ressaltar que não é suficiente apenas passar toda a documentação para o meio digital para preservar a informação, com uma mera transferência de suporte dos textos, digitalizando documentos microfilmados ou em

formato de imagem e disponibilizando-os digitalmente. Como sabemos, estes documentos assim tornados acessíveis têm um valor muito reduzido para os investigadores: por vezes a qualidade das imagens dificulta a leitura, os formatos de imagem entravam a captura do texto por parte de *software* específicos e o reconhecimento destes caracteres textuais requer treino em *softwares* dedicados, nem sempre com resultados satisfatórios. Alguns profissionais das instituições de Memórias têm, por vezes, uma atuação unicamente focada na vertente de preservar, sem ter em conta que a disponibilização em ambiente digital, se feita apenas como uma mera transferência de suporte, não só se torna redutora, como pode até obstacularizar o pleno acesso aos textos. Em algumas instituições portuguesas, dado que os documentos originais passaram a estar vedados ao público, uma vez que estão disponíveis cópias em suporte digital, o investigador, em caso de dúvida na leitura do texto (muitos textos manuscritos foram digitalizados a partir de microfilme), vê-se confrontado com entraves para cotejar o texto original. Assim, não consegue esclarecer devidamente a sua dúvida. O mesmo se passa quando necessita de dados respeitante à materialidade do texto e dos paratextuais do suporte, e cada vez se perspetiva o texto como um todo, que tem uma história. A sua trajetória não termina quando quem o redigiu colocou o ponto final.

Mesmo resolvendo a recuperação da informação de forma útil, a acessibilidade digital a documentos históricos tem de ser devidamente enquadrada e não pode ser encarada com ligeireza. Sob uma perspetiva de Humanidades Digitais, o estudo de textos pré-contemporâneos requer sólidos conhecimentos de realidades históricas passadas e de anteriores estádios das línguas naturais combinadas com conhecimentos computacionais. Assim, as Humanidades Digitais são uma oportunidade de poder ser construído um novo campo do conhecimento, assente em rigorosos conhecimentos científicos em áreas tradicionais, mas desenvolvido num ambiente multiplicador, numa nova abordagem computacional.

A interligação entre vários campos do saber constitui, estamos convictos, a grande riqueza das Humanidades Digitais. Estas são hoje imprescindíveis para estudos sobre *corpora* textuais de grandes dimensões, cujo processamento manual impede a transformação destes em dados processáveis no que respeita a fenómenos linguísticos. (MacGillivray; Tóth, 2020). A aplicação de ferramentas de Linguística de *corpus*, de métodos e técnicas de Humanidades Digitais, e de abordagens computacionais de Processamento de Linguagem Natural a fontes textuais históricas, quer manuscritas, quer transcritas, permitirá processar automaticamente grandes massas lexicais, revelando relacionamentos lexicais, interconexões e correlações impossíveis de ser realizados manualmente. Através das Humanidades Digitais, conseguimos passar do texto aos dados, gerando novo conhecimento.

Neste artigo iremos descrever uma abordagem em Humanidades Digitais, detalhando a constituição do *corpus* textual português do século XVIII das *Memórias Paroquiais* e algumas das suas características, entre as quais a variação gráfica, que é expressiva neste conjunto textual. Com recurso à ferramenta AntConc, foram feitas, listagens lexicais com as respetivas frequências de ocorrência. Dada a extensão do *corpus*, foi construído um subcorpus de estudo, de menor dimensão, que foi manualmente anotado com recurso à plataforma

INCEpTION, propondo-se, para cada variante, a respetiva correspondente em ortografia PE (português europeu) pós-acordo de 1990, com vista a obter pares lexicais, com a forma original e a respetiva forma normalizada. A partir da observação dos DataSets constituídos, as variantes gráficas foram indexadas numa tipologia da variação gráfica, descrita neste artigo, tendo em conta a frequência de ocorrência das formas originais. Esta tipologia, que advém do estudo de um *corpus* textual plural como é do das *Memórias Paroquiais*, é fundamental, cremos, para o treino de sistemas automáticos que permitam normalizar automaticamente a totalidade deste conjunto textual das *Memórias Paroquiais* e aplicar esse investimento em outros textos da época.

A normalização manual de textos é um processo moroso e muito exigente, requerendo sólidos conhecimentos linguísticos e históricos. Esta tarefa, aplicada a um volume textual tão elevado como é o *corpus* destes textos, exigiria um investimento temporal não praticável. Por outro lado, a automatização do processo de normalização requer um treino de modelos dirigidos às características linguísticas do português do século XVIII, entre as quais a variação gráfica, já que modelos treinados para o estágio contemporâneo da língua produzem resultados menos satisfatórios. Assim, pretende-se, com este estudo, propor uma tipologia da variação gráfica nas *Memórias* a partir da observação da frequência de ocorrências num subcorpus mais pequeno. O objectivo final consiste em potenciar a automação.

O conjunto textual das Memórias Paroquiais contém informações valiosas sobre o território, as gentes, os costumes, o património de vária natureza, especialmente o edificado, e a organização da sociedade do Portugal de setecentos. A disponibilização dos textos em ortografia atual vai permitir trazer estes textos para a contemporaneidade e disponibilizá-los a públicos diversos, sem formação em Linguística, facilitando o acesso ao conteúdo para estudos noutras áreas do saber e noutras geografias.

Constituição do *corpus* das Memórias Paroquiais (1758)

Em 1755 ocorreu o grande terramoto de Lisboa, acompanhado de maremoto e seguido de um grande incêndio, que arrasou grande parte da cidade naquela altura, e que causou destruição também noutras zonas de Portugal. Tal como também faziam outros reinos europeus da altura, passados quase três anos, a Secretaria de Estado efetuou um novo inquérito a todas as paróquias de Portugal. Entre outros assuntos, tinha em vista apurar os danos do sismo e o grau de reconstrução, bem como possibilitar a reconstituição de dados para um Dicionário Geográfico de Portugal, que o oratoriano Luís Cardoso começara a publicar havia poucos anos. Este inquérito de 1758 continha 60 questões, organizadas em três grandes temáticas: terra, serra, rio.

> Da cidade desta Bispado a de Lisboa dista trinta e sinco legoas legoas [sic];Priuilegios que ha nesta Aldeya os seguintes o do tabaco da santíssima Trindade o de catiuos e o de santo Antonio. As agoas desta Pouoacao e Aldeia todas são boas. No tempo do terramoto não oue Roinna alguma so sim oue hum tremedeiro nas paredes das cazas sem perda alguma. No simo desta Aldeya se acha huma serra que tem o seu nascimento na Irmida de são Paullo da uila de castello dauide tem tres quartos de legoa em comprido chamase a serra da Portella e da senhora da penha e acaba ao ribeiro do pinheiro termo da villa de Maruão tem humas cantarias feias feias [sic] Nesta serra se cria huma erua chamada arcaria que singular para dor de dentes Inchasos he esta Serra fregedisima produs poucos pastos. /p. 1017/ E sendo assim feita esta deligencia por mandado do Excelentissimo Reuerendisimo Meu senhor Dom Ioão de Azeuedo Denissimo Bispo de Portalegre Freyre da ordem de são Bento daVis e dos mais Interrogatórios não achej que responder, so sim aos que uão assim aqui na escrita e sendo nececario o Iuro In uerba sacerdotis Carreiras 2 de abril 1758 Do Parocho das Carreiras Ioze Dias Mendes [assinatura autógrafa]//"
>
> (ANTT, *Memórias Paroquiais* (MP) – *Portalegre – São Sebastião das Carreiras*, vol. 9, nº 158, p. 1016)

O inquérito foi respondido pelos vários párocos, como no extrato acima. Cerca do ano de 1832, os textos manuscritos foram compilados em 41 volumes, com todas as paróquias/freguesias indexadas alfabeticamente, independentemente do concelho a que pertenciam na época. Terão sido encadernados nessa altura, juntamente com 2 volumes de acrescentos, pois faltavam respostas de alguns locais, eventualmente por se terem extraviado ou por outro motivo. Receberam ainda um volume de índices, que permitia recuperar a informação com mais eficácia. Nos anos 30 do século XIX, a coleção acabou à guarda do Arquivo Nacional da Torre do Tombo, sendo desde 1896 reconhecida pelo nome de *Memórias Paroquiais*.

Ainda no século XIX, mas sobretudo no século XX, este conjunto documental começou a ter grande procura por parte de investigadores. Por isso, cerca de 2002 foi microfilmado, e os volumes originais deixaram de poder ser consultados pelo público. Pouco depois, os microfilmes foram digitalizados e passaram a estar disponíveis *online*, no *site* da Torre do Tombo.

Face à dificuldade de captura e processamento do texto manuscrito, o CIDEHUS-Universidade de Évora[1], num sistema colaborativo que reuniu paleógrafos, historiadores, estudantes, e também socorrendo-se de cedências de transcritores particulares, conseguiu constituir um *corpus* digital completamente processável (Figura 1, na página seguinte). Neste momento estão disponíveis os textos relativos a boa parte das paróquias do sul de Portugal continental, em acesso aberto e com licenças *Creative Commons 4.0* (Santos *et al.*, 2020).

[1] http://www.cidehusdigital.uevora.pt/

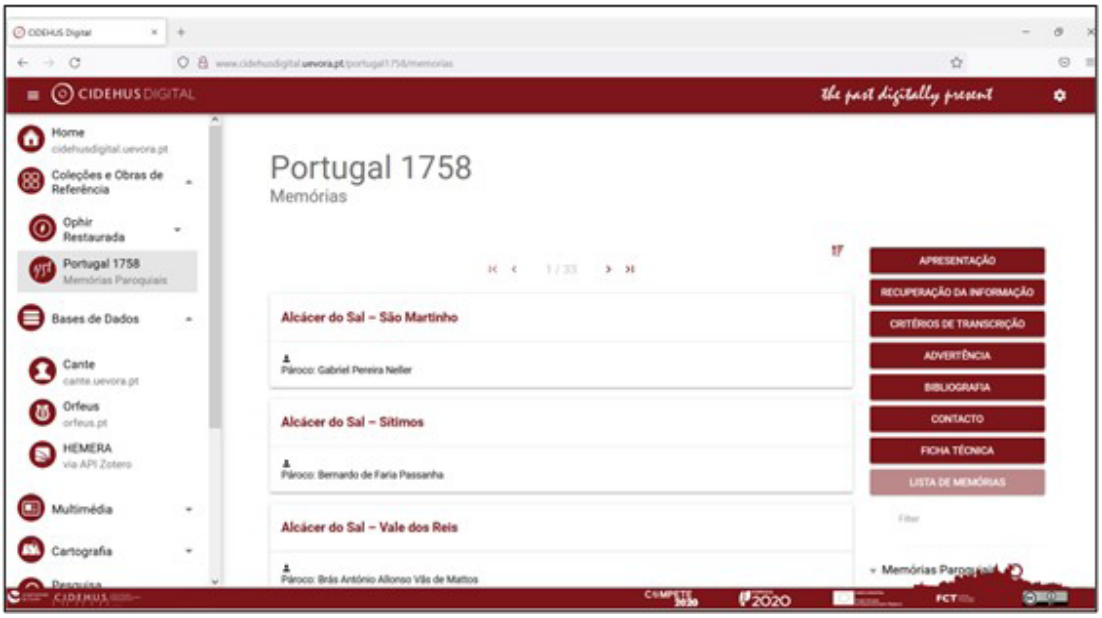

Figura 1 – As Memórias Paroquiais no CIDEHUSDigital
Fonte: http://www.cidehusdigital.uevora.pt/portugal1758/memorias

Este conjunto textual é de uma riqueza linguística inegável. Foi escrito a várias mãos, por párocos de várias idades, com várias formações e condições sociais. Se, para a Paleografia, este variado *corpus* traz grandes desafios e requisitos, para os estudos linguísticos esta variedade textual conduziu a uma expressão da variação linguística e gráfica consideravelmente maior do que a que possa ser encontrada num texto escrito por um só autor. Estas características constituíram motivos fundados para o estudo aqui proposto.

Tomámos como objeto de análise o *corpus* textual disponível no CIDEHUSDigital, que contém os textos oriundos de 366 paróquias do sul de Portugal, organizadas pelos atuais concelhos, a que chamámos convencionalmente **MP_P**. As 366 memórias do *corpus* foram processadas com recurso ao Programa de Concordâncias AntConc©[2] (Figura 2, na página seguinte).

Obteve-se uma lista não lematizada de 34.181 palavras diferentes, com 632.498 ocorrências. A lista foi ordenada alfabeticamente, facilitando o acesso ao *corpus*, e por frequência descendente, que nos esclarece o número de ocorrências de cada palavra. Cada palavra pode ser consultada em contexto, com os ocorrentes à direita e à esquerda (Figura 3, na página seguinte).

Dada a dimensão do *corpus* MP_P e a morosidade de se efetuar uma anotação manual na totalidade do *corpus*, constituiu-se um subcorpus de estudo, **MP_Portalegre**, com as 14 *Memórias Paroquiais* respeitantes às freguesias/paróquias correspondentes ao atual

[2] AntConc – A freeware *corpus* analysis toolkit for concordancing and text analysis. Anthony, L. (2022). AntConc (4.0.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from: https://www.laurenceanthony.net/software

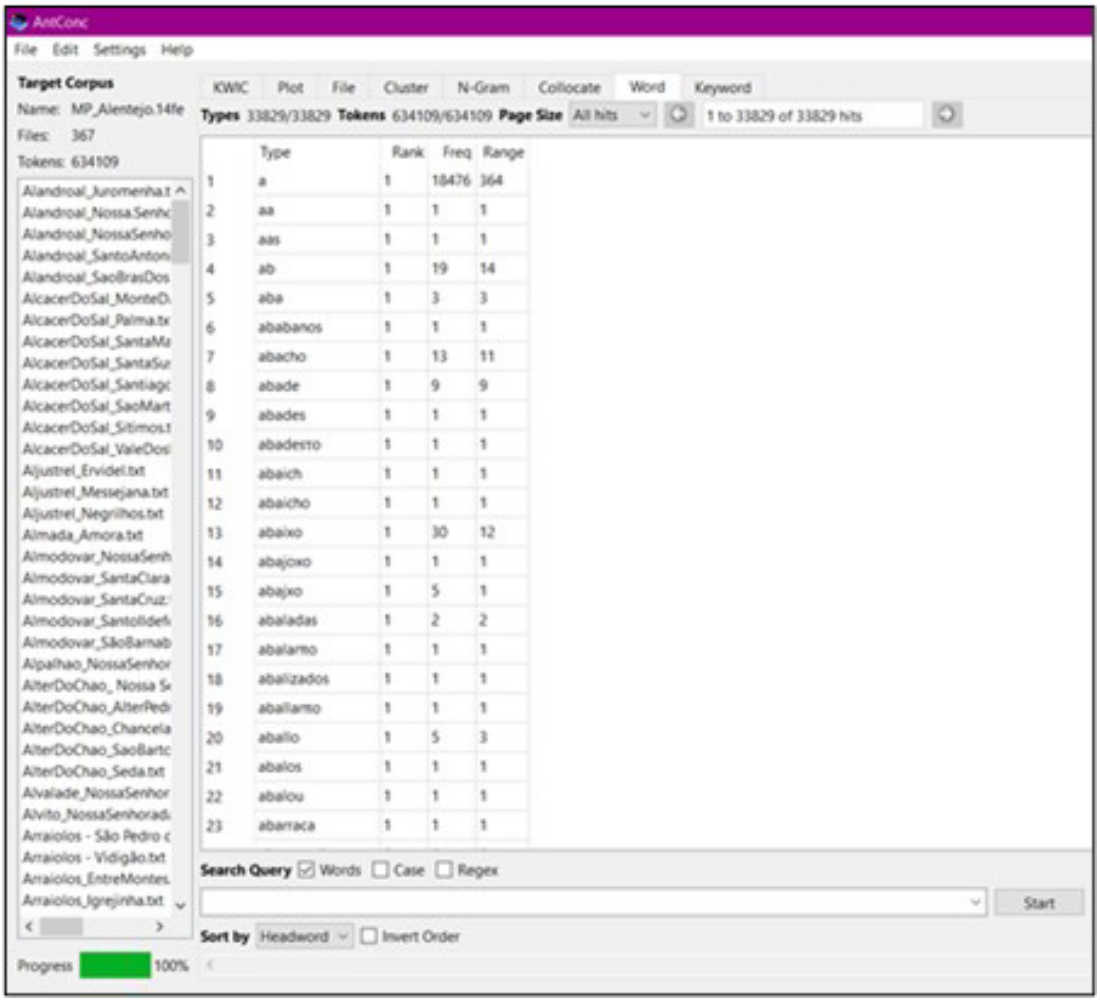

Figura 2 – Lista alfabética de palavras do *corpus* na ferramenta AntConc
Fonte: elaboração própria.

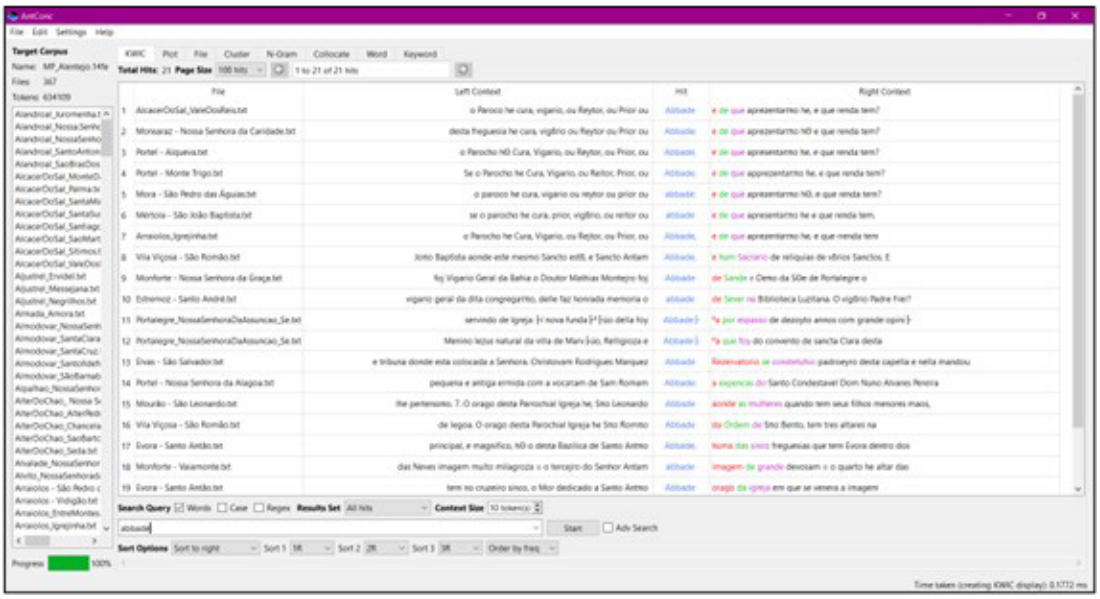

Figura 3 – Palavra em contexto – AntConc©.
Fonte: Elaboração própria.

concelho de Portalegre, cidade que é capital de um dos três distritos da região do Alentejo, em Portugal. O subcorpus **MP_Portalegre** é composto por 2 427 palavras diferentes com 15.597 ocorrências.

Previamente à anotação manual das variantes gráficas do *corpus*, foram estabelecidas *Guidelines* específicas pela equipa multidisciplinar do projeto. As palavras "mui", "cousa" e "el-Rei", ainda que lexicograficamente marcadas como "antigas", não foram consideradas variantes, uma vez que o seu uso ainda se verifica atualmente, em contexto de Português Europeu. Os pronomes clíticos em adjacência verbal foram normalizados de acordo com o uso no PE em sede de anotação. Foi normalizado o uso das maiúsculas de acordo com o postulado no Acordo Ortográfico em vigor. Os nomes próprios de pessoas e nomes geográficos com grafias diferentes da atual foram tratados como variantes gráficas, propondo-se os correspondentes na ortografia atual. A pontuação foi mantida. As palavras latinas mantiveram-se inalteráveis.

O *corpus* MP_Portalegre foi anotado manualmente com recurso à ferramenta de anotação INCEpTION[3] (*cf.* Vieira *et al.*, 2021). Embora este dispositivo esteja orientado para a anotação semântica, utilizámo-lo para o reconhecimento e recolha de variantes gráficas. A ferramenta foi configurada expressamente, criando-se uma Layer específica para a Variação. Os textos são percorridos pelo anotador, na ferramenta e, para cada palavra considerada como variante é realizada uma anotação, indicando, sobrescrito à palavra, uma etiqueta com a forma em grafia atual correspondente. Trata-se de um processo minucioso, que implica conhecimento linguístico e histórico em contexto, suportado em processamento computacional orientado para a resolução do problema. Assim, a anotação gráfica era frequentemente discutida em equipa multidisciplinar, de modo a não desvirtuar nem o contexto histórico, nem computacional.

A Figura 4, na página seguinte, elucida o processo de anotação na plataforma referida para um dos textos do subcorpus.

No subcorpus de estudo foram anotadas 3 976 formas que diferem na sua grafia face à ortografia em vigor. A partir da plataforma de anotação, extraiu-se um ficheiro .tsv, que foi pós-processado manualmente. Foi construído um *DataSet* com três colunas, *MP_PTG.all*: na primeira, podemos localizar a variante gráfica no texto de origem/freguesia onde esta aparece; a segunda coluna é preenchida com as variantes do século XVIII retiradas do *corpus*, com a indicação das formas atuais correspondentes, na terceira coluna (Quadro 1, na página seguinte).

[3] INCEpTION – A semantic annotation platform, UK Lab – Technische Universität Darmstaad, https://inception-project.github.io ; software is licensed under the Apache License 2.

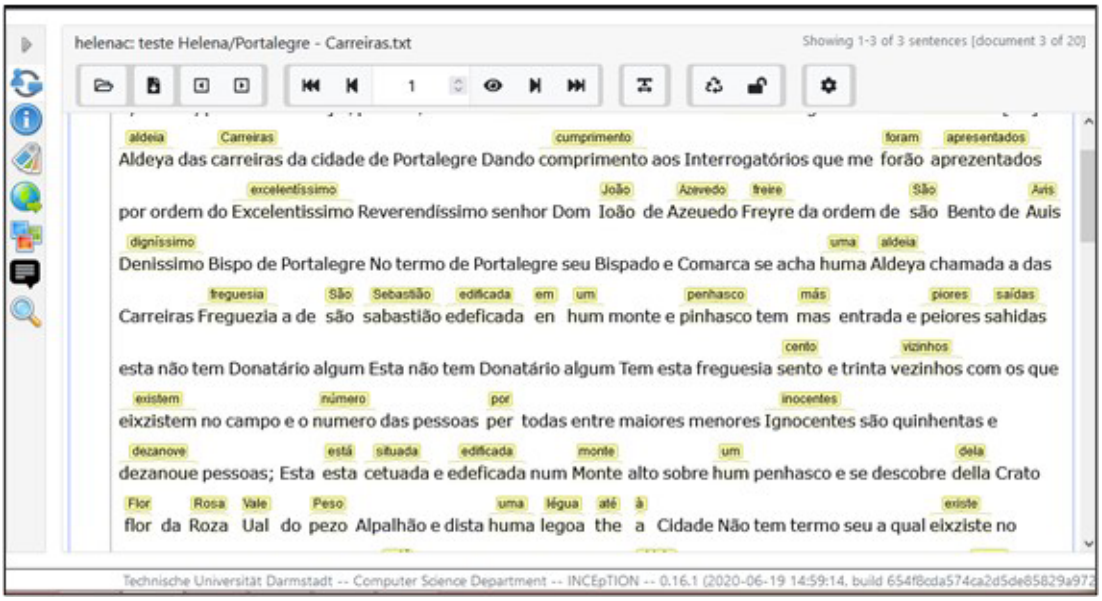

Figura 4 – Anotação manual da paróquia Portalegre-Carreiras na plataforma INCEpTION
Fonte: elaboração própria.

Quadro 1 – Extrato de DataSet Memórias de Portalegre – variantes gráficas.

Freguesia	Forma presente no texto	Forma atual
Portalegre_NossaSenhoraSaAssuncao_Se	Assima	Acima
Portalegre_NossaSenhoraSaAssuncao_Se	assima	acima
Portalegre_RibeiraDeNisa	acçoez	ações
Portalegre_SaoLourenco	administrádo	administrado
Portalegre_RibeiraDeNisa	adimitio	admitiu
Portalegre_NossaSenhoraSaAssuncao_Se	advirtir	advertir
Portalegre_NossaSenhoraSaAssuncao_Se	affirma	afirma
Portalegre_NossaSenhoraSaAssuncao_Se	affirma	afirma
Portalegre_Reguengo	afirmão	afirmam
Portalegre_Fortios	afluencia	afluência
Portalegre_NossaSenhoraSaAssuncao_Se	Affonço	Afonso
Portalegre_NossaSenhoraSaAssuncao_Se	Affonço	Afonso
Portalegre_RibeiraDeNisa	Affonso	Afonso

Fonte: Elaboração própria

A partir do DataSet, foi construído um segundo *DataSet*, *MP_PTG.formasAtuaisComVariantes*, com indexação pelas formas em ortografia atual, não lematizadas, o que permitiu reunir 1129 palavras atuais diferentes, que têm variantes século XVIII no *corpus*. Numa primeira análise quantitativa, podemos ver que cada forma atual reunia, em média, três variantes gráficas no *corpus* de estudo.

A partir dos dados obtidos, propôs-se uma tipologia da variação gráfica, que servirá como suporte para uma futura automatização do processo de normalização dos textos para grafia atual. Este conjunto de dados, que reúne um léxico de variantes, irá contribuir para o treino de um modelo automático ou semi-automático para uma normalização ortográfica de textos em português do século XVIII, que permitirá, espera-se, facilitar o acesso e estudo destes textos de elevado valor patrimonial a públicos especialistas não linguistas. Por outro lado, o processamento computacional destes textos com vista à extração automática de dados para obtenção de conhecimento (por exemplo, o Reconhecimento de Entidades Nomeadas) torna-se muito difícil, e as variantes gráficas revelam-se como verdadeiros obstáculos que muito diminuem a precisão dos sistemas.

Para uma tipologia da variação gráfica nas *Memórias Paroquiais*

Se, no estudo de textos contemporâneos, as Humanidades Digitais são já uma realidade, a sua aplicação ao estudo de documentos textuais históricos traz desafios acrescidos (Gonçalves e Banza, 2013). Este novo território de estudos requer uma abordagem interdisciplinar, *ab initio*, que é construída pelos vários campos do saber envolvidos interligados. O estudo das *Memórias Paroquiais* (1758) que temos vindo a desenvolver tem sido um exemplo privilegiado de uma abordagem em Humanidades Digitais, que nos tem levado dos manuscritos aos textos processáveis. Neste artigo tenta-se descrever um dos passos dessa abordagem, a normalização gráfica, partindo de uma observação da variação gráfica neste período tão rico da língua portuguesa.

A descrição do fenómeno da variação gráfica tem sido objeto de estudo de vários linguistas como Galves, Gonçalves, Cardeira, Marquilhas, Kemmler, entre outros igualmente relevantes e pertinentes.

Neste artigo cingimo-nos, apenas, à descrição da variação gráfica neste *corpus* de estudo bem como ao processo efetuado com vista à sua anotação computacional e posterior utilização em processos computacionais automatizáveis.

Os períodos pré-contemporâneos da língua portuguesa são marcados por uma expressiva renovação lexical (Verdelho, 1987) e por uma grande variação ortográfica (Kemmler, 2001). Os primórdios de normalização apenas começaram a partir do século XVI (Mateus e Cardeira, 2007) e no século XVIII houve intensa discussão sobre fenómenos de estabilização ortográfica, sem que tenha havido uma verdadeira proposta de modelo ortográfico (Gonçalves, 2020, p.653). A variação ortográfica é, assim, uma característica da língua portuguesa escrita até aos séculos XIX e XX (Gonçalves, 2020).

No século XVIII, a ortografia portuguesa era ainda bastante diferente da atual norma ortográfica em vigor (Cardeira, 2006), (Cardeira & Mateus, 2008).

O conjunto textual das Memórias Paroquiais, escrito a várias mãos, é revelador dessa diversidade ortográfica descrita na Literatura. O subcorpus construído reúne um conjunto de

variantes gráficas, que descrevemos brevemente neste ensejo. Com o processo computacional que desenvolvemos, descrito anteriormente, que nos permitiu ter pares de formas lexicais em grafia século XVIII e século XXI, pretendemos estudar a variação gráfica em contexto, de modo a tentarmos descrever este fenómeno com vista a um futuro treino mais eficiente de ferramentas computacionais que possam vir a produzir uma normalização gráfica de textos portugueses do século XVIII estruturada em critérios linguísticos e históricos.

Atualmente, as ferramentas que viabilizam o tratamento automático de textos em português em estádios pré-contemporâneos têm limitações e tal processamento não permite, ainda, tanto quanto sabemos, obter resultados em grandes volumes de massas textuais. (Cameron, Gonçalves & Quaresma, 2020). Para outras línguas europeias, têm sido feitos esforços com vista ao desenvolvimento de abordagens computacionais capazes de normalizar graficamente um texto pré-contemporâneo de modo automático. Após esta etapa, o texto, em ortografia contemporânea, pode não só ser lido por públicos não linguistas como também pode ser processado computacionalmente com eficácia e eficiência muito mais elevadas que o processamento computacional diretamente efetuado sobre o texto transcrito pré-contemporâneo. Entre outros, refiram-se os trabalhos desenvolvidos por Baron & Rayson (2008, 2011, …) para o inglês, Bollman (2011, 2018, …) para o alemão, e Gabay (2020) para o francês, que visam, com diferentes técnicas e abordagens, automatizar o processo de transposição de um estágio pré-contemporâneo de uma língua natural para o respetivo estágio contemporâneo.

Nas *Memórias Paroquiais*, a variação gráfica é muito expressiva. Tem uma elevada incidência de variantes no texto, como é exemplo a frase seguinte:

> (1) "Tem termo Seu e Comprehende as Aldeyas Seguintes: a Saber, Alagoa, Carreyras, Fortios, Ribeyra de Niza, Reguengo, E Vrra, e distáo meia Legoa, E uma Legoa e duas Legoas".
>
> (ANTT, MP, *Portalegre – São Lourenço*, Vol. 29, nº 223c, pp. 1533)

Veja-se o exemplo do termo "circunstância", que é escrito com nove variantes. Apresentamos, para cada uma destas, o número de ocorrências no *corpus* MP_P:

circunstância	5
circunstancia	13
circunstancias	9
circunstanças	3
çircunstanças	1
circunstansias	1
circunstansias	1
sircunstança	1
sircunstancias	5

Calculamos que aproximadamente 70% do total das palavras do *corpus* MP_P sejam variantes.

A variação gráfica no *corpus* textual das *Memórias Paroquiais* não tem uma incidência uniforme ao nível da frequência de ocorrências no *corpus*. Algumas características linguísticas da variação são muito frequentes no *corpus* enquanto outras aparecem raramente.

A variação da representação gráfica das sibilantes é a característica linguística com maior número de ocorrências neste *corpus*. Entre outros exemplos, também expressos nos excertos aqui escolhidos, destaque-se o termo "alicerce", grafado de seis formas diferentes:

> (2) "[…] segundo se achaõ alguns antigos **alicerçes**, que parece o foraõ".
>
> (ANTT, *MP*, *Torrão – Odivelas*, vol 26, nº 7, p. 74)
>
> "Do ditto convento se descobrem alguns vestigios, como **alicersses**, sepulturas […]".
>
> (ANTT, *MP*, *Évora – São Miguel de Machede*, vol. 22, nº 16, p. 89)
>
> "[…] jgreja propria, que há muitos anos se conserva só com os **alicersez**."
>
> (ANTT, *MP*, *Vila Viçosa (São Bartolomeu)*, vol. 40, nº 271ª, p. 1665)
>
> "Em contorno da Villa, forao achados muntos **alisserces**, columnas e bazes […]".
>
> (ANTT, *MP*, *Alandroal-Juromenha*, vol. 18, nº 48, p.315)
>
> "[…] abrindoçe os **aliçerçes** da nova capella que á mesma Senhora de Ayres tem fabricado os seuz devotos […]".
>
> (ANTT, *MP*, *Viana do Alentejo – Viana do Alentejo*, vol. 39, nº 150, p. 895)
>
> "[…] ao redor huma parede mais larga arruinada, e quazi posta no **alicerce** […]".
>
> (ANTT, *MP*, *Nisa – Alpalhão*, vol. 3, nº 16, p. 144)

A segunda característica linguística da variação mais frequente no *corpus* é o uso da grafia pseudo-etimológica em dígrafos sem valor linguístico, como por exemplo o dígrafo "-th-", com notória existência no *corpus*, com 211 ocorrências, em palavras de diversas categorias morfológicas. O seguinte exemplo ilustra este registo:

> "Relaçam Da freguezia de Santa **Catharina** de Pardais **thermo** de Villa Viçoza."
>
> (ANTT, *MP*, *Vila Viçosa – Pardais*, vol. 27, nº 82, p. 523)

Veja-se que a forma gráfica coincidente com a forma da ortografia atual nem sempre é a que tem maior número de ocorrências. A forma "**até**" tem 85 ocorrências no *corpus* e as variantes gráficas "**athe**" e "**athé**" têm, respetivamente, 305 e 144 ocorrências.

O uso de consoantes duplas sem valor linguístico, característica bem conhecida deste período da língua, tem igualmente uma elevada frequência de ocorrências no *corpus* MP_P. Veja-se o caso do dígrafo "-ll-", a título de exemplo. Encontrámos 965 formas simples grafadas com este dígrafo que correspondem a 33.829 ocorrências. Estes dígrafos surgem em todos os textos do *corpus*, sem que possa ser percebida alguma tendência de uso ou eventual critério. A incidência do uso destas consoantes duplas é muito grande. Por vezes, numa só frase, encontram-se diversas palavras assim grafadas, como na seguinte:

> "Na **capella** mor está **collocada** a Jmagem do Sancto: no **collateral** direito […]".
>
> (ANTT, *MP*, *Elvas – Santa Eulália*, vol.14, nº 110, p. 803)

Uma característica da variação com elevado número de ocorrências no *corpus* é o uso de "y" e de "j" com valor de "i" em ditongos. A representação da vogal com "y" tem 412 ocorrências. O exemplo seguinte evidencia este uso:

> "[…] della he certo **veyo** a freguezia para esta que hoje existe e he a quarta de que ha noticia: está quazi no **meyo** do povo […]".
>
> (ANTT, *MP*, *Elvas -Campo Maior*, vol. 8, nº 80, p. 557)

A representação da vogal no ditongo era feita com "y" ou "j" sem que seja percetível, neste *corpus*, algum critério.

> A **mayor** abundancia de fructos, que a villa tem, he Castanha […]
>
> (ANTT, MP, Portalegre – Alegrete, vol. 19, nº 41, p. 274)
>
> A **major** abundançia de frutos que produs a mesma freguezia conforme os annos, conciste em vianda de azinho sevaro
>
> (ANTT, MP, Portalegre- Caiola, vol 8, nº 37, p 209)

De igual modo, o uso de "u" com valor consonântico tem uma elevada frequência de ocorrências neste *corpus*.

> E sendo assim feita esta deligencia por mandado do Excelentissimo **Reuerendisimo** Meu senhor Dom Ioão de **Azeuedo**
>
> (ANTT, Portalegre – Carreiras, vol. 9, nº 158, p. 1017)

Estas características da variação foram agrupadas num grupo convencional, a que chamámos tipo 1.

O segundo grupo convencional com maior número de ocorrências tem a ver com o uso das maiúsculas e da acentuação gráfica. Chamámos-lhe Tipo II. Muitos nomes comuns eram grafados com maiúscula e alguns nomes próprios não o eram. Não se consegue notar, neste *corpus*, um critério uniforme que possa ser percetível no que respeita ao uso da letra maiúscula. O extrato seguinte, tal como muitos outros igualmente pertinentes, elucida este uso:

> "Satisfazendo aos Jnterrogatorios que me foraõ mandados por Voça Excelencia Reverendissima digo Chamada Santa Anna. Que a minha freguezia está cituada na Provincia do Alentejo, Arcebispado de Evora, Termo da Villa de Arraiollos, Commarca de Villa Viçosa da Serenissima Caza de Bargança. Entre Montes e Aldeya tem cento e dezanove vizinhos, e pessoas saõ quinhentas e sincoenta e sete."
>
> (ANTT, *MP*, *Arraiolos – Entre Montes*, Vol. 13, nº E 23, p. 201)

Dada a imprevisibilidade do uso das maiúsculas, alguns paleógrafos, ao transcreverem do original manuscrito, intervieram na normalização das maiúsculas. Outros respeitaram o original. Apenas a consulta e processamento do texto manuscrito pode revelar, com exatidão, a incidência desta variação no texto.

A acentuação gráfica está presente em diversas variantes em todo o *corpus*, como no exemplo:

> "[…] hum álto **naõ** muyto distante da cidade: a fundação **naó** he muyto antiga **naõ** se sabe porem o anno certo della […]".
>
> (ANTT, *MP*, *Portalegre – Nossa Senhora da Assunção*, vol. 29, nº 223, p. 1518)

Estabelecemos, ainda, outros grupos convencionais na proposta de tipologia da variação neste *corpus*. Eles são preenchidos pelo registo gráfico dos pronomes clíticos em adjacência verbal (tipo III), aglutinação na grafia de algumas palavras (tipo IV). Estes grupos têm menor número de ocorrências no *corpus*.

Existe, ainda, um tipo de variação gráfica devida a razões que não estão ligadas ao registo linguístico, tais como erros do próprio escrevente, ou alguma "criatividade" deste no registo escrito. Destacamos alguns exemplos que nos pareceram elucidativos, entre muitos outros igualmente pertinentes. Note-se que muitos destes termos variantes estão devidamente assinalados pelos transcritores com a expressão [sic], mostrando de forma evidente este afastamento da norma não explicado por nenhum fenómeno linguístico:

> A sua parroquia está **dentra** [sic] da aldeya a hum ládo della, e fora desta nam há mais lugares, nem aldeyas na freguesia.
>
> (ANTT, *MP*, *Odemira – São Teotónio*, vol. 36, nº 51, p. 321)
>
> Nam ha **tambel** [sic] hospital.
>
> (ANTT, *MP*, *Moura – Amareleja*, vol.3, nº 60, p. 472)
>
> Tem sinco altares, a saber: o altar mayor, Espirito Santo; o altar na derejta, Nossa Senhora do Carmo; o **culatrar** [sic] da esquerda, Nossa Senhora do Rozário;
>
> (ANTT, *MP*, *Mértola – Espírito Santo*, vol. 14, nº 77, p. 519)

Estas variantes foram classificadas na nossa proposta de tipologia com o tipo V. Estes grupos não conseguem ser descritos de forma regular, para poderem ser objeto de uma normalização automatizada, pelo que terão de ser descritos casuisticamente.

Rumo à normalização

Apesar de esforços como o apresentado em (Reynaert,& Marquilhas, 2012), não temos conhecimento de ferramentas atuais, disponíveis, capazes de fazer uma conversão automática ao português contemporâneo, e que considerem todas essas questões. Novas abordagens de tratamento computacional, baseadas em modelos de linguagem construídos com técnicas de redes neurais (Bollmann, 2018) podem ser aplicadas a partir da disponibilidade de *corpora* do português do século XVIII digitalizados. Essas abordagens requerem uma avaliação baseada em exemplos anotados. O trabalho realizado aqui serve de base para a construção de um conjunto de avaliação, que sirva quase de matriz (Quadro 2, na página seguinte). Contudo, enquanto a avaliação não for considerada adequada, é preciso retreinar os modelos até que se obtenham resultados satisfatórios, dada a complexidade do fenómeno, o elevado número de palavras em questão e a imprevisibilidade associada.

O esforço de descrição da variação com vista a uma futura automatização da normalização pode, no nosso entender, beneficiar desta proposta de tipologia, conjugada com uma descrição sob a forma de regras para os grupos da tipologia apresentada com maior número de ocorrências no *corpus*. Dada a arbitrariedade da grafia, apesar dos progressos das redes neuronais neste pelouro, a automatização poderia melhorar a acuidade assertiva com um sistema de regras. É uma espécie de reforço, num campo de muita aleatoriedade / pendor criativo do autor do texto ou de reduzido nível de domínio da escrita.

Quadro 2 – Quadro resumo da tipologia de variação no *corpus* das Memórias Paroquiais (1758)

tipo 1	tipo 2	tipo 3	tipo 4	tipo 5
consoantes duplas pseudo-etimológicas	letras maiúsculas	pronomes clíticos em adjacência verbal	aglutinação de palavras independentes	"criatividade" do autor
representação das sibilantes	acentuação gráfica			
"y"/"j" com valor de "i" nos ditongos				
"u" com valor consonântico				
dígrafos "ch" e "th"				

Fonte: elaboração própria

Considerações finais

As Humanidades Digitais, nesta experiência, combinam saberes de áreas distintas como história, linguística, paleografia e informática, e requerem, de cada uma das áreas envolvidas, adaptações. Esta interação não é passiva, nem menos ainda uma soma de conhecimentos. É transformativa e performativa e rapidamente apela por cruzamentos com outros saberes, num fluxo que pode ser imparável. Desde logo, na história e na paleografia, passa por uma modificação daquilo que é entendido como a publicação de fontes. Na transcrição de fontes importa cada vez mais a fidelidade absoluta ao original para servir os interesses do linguista. Estudar adequadamente as variantes exige isso mesmo. Contudo, importa chegar a algoritmos que possibilitem a conversão automática da transcrição numa lição que faça a atualização automática das formas gráficas para facilitar a recuperação da informação. Publicar fontes não pode continuar a ser uma atividade passiva. Importa que esta seja cada vez mais em versão apta à transformação digital e à conversão das fontes textuais em dados, que podem ser georreferenciados, visualizados de múltiplas formas, ligados a outros a partir de bases de conhecimento, entre muitas outras possibilidades que a cada dia surgem. São elas o produto da interação com a geografia, a estatística e outras áreas. E tudo isto coloca permanentemente novas exigências à computação. Na linguística, estudar as variantes ortográficas pretéritas torna-se cada vez mais tarefa imperativa.

O desenvolvimento de sistemas computacionais aptos a tratarem a língua natural é recente e estes sistemas comumente tomam por base os textos contemporâneos. Para serem capazes de auxiliar em processos de fontes históricas, necessitam de se adaptar para as diferenças de codificação, ou diferenças linguísticas existentes entre as diferentes épocas. A linguística, em sede de Humanidades Digitais, pode, assim, ter como função conciliar a

língua natural com as novas possibilidades desenvolvidas nas áreas de inteligência artificial e processamento de linguagem natural. No desafio proposto neste trabalho, a linguística relaciona o português do século XVIII com o contemporâneo, um desafio para as abordagens computacionais. E esse repto não terá solução se as linhas de programação forem escritas de forma isolada. Gerar novos e eficientes produtos não dispensa estas múltiplas interações que esbatem as fronteiras disciplinares.

Fonte manuscrita

ARQUIVO NACIONAL TORRE DO TOMBO (ANTT). *Memórias Paroquiais*, disponível em: https://digitarq.arquivos.pt/details?id=4238720 e transcrito em http://www.cidehusdigital.uevora.pt

Referências

BANZA, A. P., GONÇALVES, M. *Roteiro de História da Língua Portuguesa*. (U. C.-H. Heritage, Ed.) Évora: Universidade de Évora, 2018.

BARON, A., RAYSON, P. "VARD 2: A tool for dealing with spelling variation in historical corpora". *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Birmingham: Aston University, 22 May 2008.

BARON, A. *Dealing with spelling variation in Early Modern English texts*. PhD Thesis, Lancaster University, 2011.

BOLLMANN, M., PETRAN, F., DIPPER, S. Rule-Based Normalization of Historical Texts. In *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, pages 34-42, Hissar, Bulgaria. Association for Computational Linguistics, 2011.

BOLLMANN, M. *Normalization of historical texts with neural network models*. Universitätsbibliothek Johann Christian Senckenberg. (Dissertation), 2018.

CAMERON, H.F., GONÇALVES, M.F., QUARESMA, P. "Linguistic and orthographical classic Portuguese variants. Challenges for NLP". In: Maria José Finatto, Renata Vieira, Senja Pollak and Saturnino Luz (ed.), *Proceedings of the Workshop on Digital Humanities and Natural Language Processing*, co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020), vol. 2607. Évora (Portugal): CEUR-WP org, 43-48, 2020.

CARDEIRA, E. *O essencial sobre a História do Português*. Alfragide: Editorial Caminho, 2006.

CARDEIRA, E., MATEUS, M. H. *Norma e Variação*. Alfragide: Editorial Caminho, 2008.

EDMOND, J. (ed). *Digital Technology and the Practices of Humanities Research*. Cambridge, UK: Open Book Publishers, 2020, disponível em: https://doi.org/10.11647/OBP.0192

EUROPEAN COMMISION. *Commission recommendation of 10.11.2021 on a Common European data space for cultural heritage*, Brussels, 10.11.2021 – C(2021) 7953 final, disponível em: https://digital-strategy.ec.europa.eu/en/news/commission-proposes-common-european-data-space-cultural-heritage

GABAY, S., BARRAULT, L. *Traduction automatique pour la normalisation du français du XVIIème siècle*. TALN 2020, ATALA, Jun 2020, Nancy, France.

GONÇALVES, M. F., BANZA, A. P. (ed.). *Património Textual e Humanidades Digitais: da antiga à nova Filologia*. Évora: Publicações do CIDEHUS, 2013, disponível em: https://books.openedition.org/cidehus/1073

GONÇALVES, M.F.. Orthography and Orthoepy, in Lebsanft, Franz and Tacke, Felix. *Manual of Standardization in the Romance Languages*. Berlin; Boston: De Gruyter, 2020. p. 651-678. Disponível em: https://doi.org/10.1515/9783110458084

MCGILLIVRAY, B., MIHÁLY, G., *Applying Language Technology in Humanities Research*, Cham: Palgrave Macmillan – Springer Nature Switzerland, 2020.

REYNAERT, M., HENDRICKX, I., & MARQUILHAS, R. Historical spelling normalization. A comparison of two statistical methods: TICCL and VARD2. *Proceedings of ACRH-2*, 87-98, 2012.

SANTOS, I., OLIVAL, F., SEQUEIRA, O. «Excavating the data pit: the Portuguese Parish Memories (1758) as a gold standard», in *DHandNLP 2020: Digital Humanities and Natural Language Processing: Proceedings of the Workshop on Digital Humanities and Natural Language Processing (DHandNLP 2020)* co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2020). ed by M. José Finatto; Renata Vieira; Senja Pollak; Saturnino Luz, Évora, v. 2607, 2020. Disponível em: http://ceur-ws.org/vol-2607/.io

SCHREIBMAN, S., SIEMENS, R., UNSWORTH, J. (ed.). *A companion to Digital Humanities*, Oxford: Blackwell, 2004.

VENTURA, A. (dir.) As Memórias Paroquiais de 1758 do actual Concelho de Portalegre", in *A Cidade – Revista Cultural de Portalegre*, nº 10 (nova série), 1995, p. 93-136. Disponível em: https://www.bdalentejo.net/bdaobra/bdadigital/obra.aspx?id=253#

VIEIRA, R., OLIVAL, F., CAMERON, H.F., SANTOS, J., SEQUEIRA, O. and SANTOS, I., 2021. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data*, 7, p.20. disponível em: http://doi.org/10.5334/johd.43