Fauna e Flora setecentista: das entidades aos problemas de normalização

Helena Freire Cameron[1,2], **Fernanda Olival**[2], **Daniel Reyes**[2], **Renata Vieira**[2]
[1]Portalegre Polytechnic University, Portugal
[2]University of Évora, CIDEHUS
helenac@ipportalegre.pt, mfo@uevora.pt
daniel.a.g.reyes@gmail.com, renatav@uevora.pt

Resumo

Este artigo aborda tarefas do tratamento de fontes históricas do século XVIII, em língua portuguesa. O trabalho desenvolvido incidiu nos domínios específicos de fauna e flora. Por esta última característica, esperava-se um fraco nível de ambiguidade vocabular, mas assim não aconteceu. Por isso, apresenta-se um roteiro do processo de normalização ortográfica; descreve-se a constituição do *corpus* anotado de entidades e, sobretudo, discutem-se problemas ligados à variação lexical nestes *thesauri* de especialidade e os constrangimentos do processo. Desta forma, pretende-se contribuir para a reflexão sobre o que é o processo de normalização de fontes históricas e chamar a atenção para a importância das boas práticas neste quadro.

1 Introdução

A anotação de entidades tem possibilitado estudos de diversa natureza em variadas áreas. No entanto, para uma realidade pretérita, a anotação de EN reveste-se de uma complexidade acrescida, com exigência de uma definição mais diferenciada ao nível das categorias, uma vez que as comummente aplicadas não são suficientemente abrangentes para o universo específico aqui tratado: Fauna e Flora.

No campo das Humanidades tem-se insistido quase sempre no mesmo padrão de categorias e num leque reduzido de entradas (Rodríguez-Puente et al., 2019). Acresce que a anotação com vista à criação de *datasets* capazes de constituir padrão ouro para treino de modelos tem requisitos que devem ser cumpridos, especialmente em textos históricos, pois exigem categorias adaptadas ao domínio específico e às linhas estruturantes da época em estudo (Álvarez Mellado et al., 2021).

Neste trabalho, aborda-se uma fonte histórica do século XVIII que reúne dados relevantes sobre o território português, a sua ocupação e o seu enquadramento natural (serras, rios) nesse período. Duas dimensões diretamente associadas à ocupação do território são a fauna e a flora, frequentemente relacionadas não apenas com os hábitos e costumes das populações, mas também com as atividades económicas das diferentes regiões.

A constituição de *corpora* anotados a partir de textos históricos em português e com validação científica humana é ainda reduzida face a outras línguas igualmente de expressão mundial. Citem-se os trabalhos desenvolvidos por Aguilar et al. (2017); Grilo et al. (2020); Zilio et al. (2022); Santos et al. (2024); Nunes et al. (2025) entre outros. Estes estudos e recursos são muito necessários, não só como aplicação de processos de Processamento de Linguagem Natural (NLP) a textos pré-contemporâneos como pela constituição de *datasets* capazes de se constituírem como *gold standard*, por exemplo, para tarefas automatizadas de anotação de entidades.

Neste descreve-se a constituição de um *corpus* anotado em duas áreas específicas, Flora e Fauna. Analisam-se os dados relativos às subcategorias e às entidades, em cinco concelhos do Sul de Portugal no século XVIII, que serviram de amostra. Com base nos resultados obtidos da anotação, discutem-se tópicos que vão muito para além da normalização gráfica, nomeadamente questões de variação lexical e constrangimentos neste *corpus* anotado que, sendo em domínios específicos, se esperava, à partida, que fosse mais objetivo e isento de ambiguidades.

2 A Constituição do *Corpus* anotado

As *Memórias Paroquiais* são um conjunto textual de grande relevância, reunindo informações de cada uma das freguesias do Portugal de setecentos, após o sismo de 1755, que teve impacto sobre boa parte do território. A coroa portuguesa fez chegar a todos os párocos um questionário com 60 perguntas sobre a Terra (população, edificado, usos e costumes, etc.), a Serra (características do território,

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 275–281

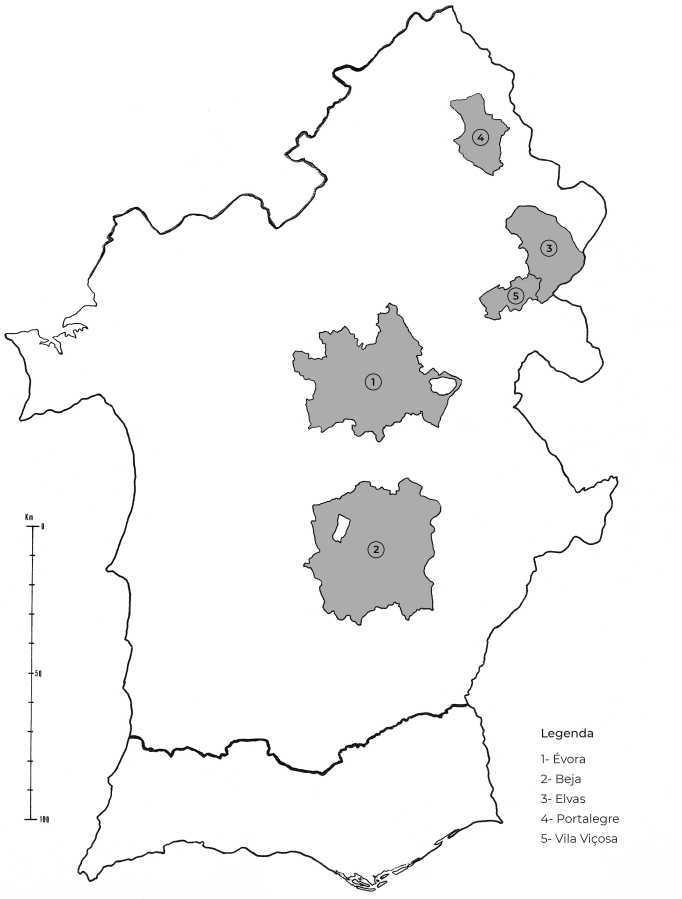

Figura 1: Região sul de Portugal, com as áreas dos concelhos estudados

plantas e animais, entre outros) e os Rios (existências, navegabilidade, rentabilidade económica, propriedades das águas, entre outros). As respostas dos padres, que visavam ser reunidas num futuro Dicionário Geográfico, foram coligidas posteriormente. As digitalizações dos originais manuscritos estão disponíveis *online* no Arquivo Nacional da Torre do Tombo. Os textos relativos ao Sul de Portugal foram transcritos por paleógrafos e estão disponíveis no repositório do CIDEHUSDigital[1]. Foi a partir destes que se desenvolveu esta investigação.

Os textos históricos anteriores ao século XX caracterizam-se por uma grafia não padronizada. Uma palavra, por pequena que fosse, podia ser redigida de múltiplas formas, por vezes no mesmo documento. Foi apenas na segunda década do séc. XX que foi introduzida a ortografia uniforme que todos deviam seguir. Por conseguinte, para recuperar informação de forma eficaz, a realidade anterior representa hoje um grande constrangimento. Assim, uma tarefa essencial é tentar normalizar a grafia, seja de modo manual ou tentando automatizar. Este último desiderato é um objetivo ainda em construção, conforme Cameron et al. (2023).

Para este estudo, os textos foram normalizados manualmente para a ortografia contemporânea, preservando toda a variação lexical. Para não desvirtuar a realidade histórica e linguística, optou-se por uma intervenção conservadora, limitada à atualização gráfica para o padrão hodierno: regularizaram-se os ditongos nasais (am → ão), eliminaram-se consoantes pseudoetimológicas (e.g. -th-) e reduziram-se consoantes duplas não etimológicas (e.g. -ll-, -bb-). Manteve-se a variação linguística (e.g. oiro/ouro), bem como formas antigas ainda em uso (El-Rei, cousa, mui). Ainda assim, o processo revelou-se complexo: na ausência de automatização para a normalização gráfica, todo o trabalho foi feito manualmente, o que implicou um elevado investimento de tempo e de recursos humanos.

Concluída esta tarefa, foi feita uma ano-

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

tação manual de entidades de Fauna e Flora na plataforma INCEpTION[2]. Os dados obtidos (formato CONLL) foram pós-processados, tendo sido lematizados manualmente (por entidade), constituindo assim o dataset. A anotação manual nos domínios de Fauna e da Flora foi feita em 87 textos relativos às três capitais de distrito do Alentejo (Évora, Beja e Portalegre), na região do sul, que correspondia a cerca de um terço de Portugal. Às localidades invocadas juntaram-se mais dois concelhos: Vila Viçosa, por ter sido sede da Casa de Bragança até 1640, e Elvas, pela sua posição fronteiriça. Na Figura 1, pode observar-se a localização geográfica de todos os concelhos tratados.

Évora era o concelho com maior relevância à época. Era composto por 22 freguesias, fazendo parte de uma ampla zona rural. Évora constituía, à data, uma urbe com importância económica e político-administrativa: era sede de arcebispado e de um dos três tribunais do Santo Ofício do espaço metropolitano português; tinha universidade. Até 1640 fora a segunda cidade portuguesa, em termos políticos.

Beja, no Baixo Alentejo, era uma zona predominantemente rural, com propriedades de grande extensão. Contabilizava, em 1758, 28 freguesias, urbanas e rurais.

Elvas situava-se no Alto Alentejo e reunia, à época, 17 freguesias.

O que é hoje o concelho de Portalegre congregava, no século XVIII, 10 freguesias, urbanas e rurais. Este concelho apresenta uma geografia muito distinta dos restantes por se situar numa zona fortemente arborizada, em plena Serra de S. Mamede, no Alto Alentejo.

Vila Viçosa englobava, na época, 6 freguesias, urbanas e rurais. Situa-se no Alentejo Central e geograficamente contém várias serras, sendo cruzada por um rio, afluente do Guadiana.

3 Fauna, Flora e sub-categorias

As categorias Fauna e Flora, pela sua abrangência, foram fracionadas em unidades mais pequenas, que pudessem descrever melhor as várias tipologias dentro de cada uma destas. Assim, **Fauna** foi subdividida em sete subcategorias (em inglês, para maior comparabilidade): *Fish* (peixes), *Bird* (aves), *Mammal* (mamíferos), *Reptile* (répteis), *Insect* (insetos), *Other* (outros animais não incluídos nos itens anteriores), *Product* (produtos derivados, como couro, etc.). No que respeita a **Flora**, esta foi igualmente dividida em sete subcategorias: *Herb* (ervas silvestres e cultivadas), *Tree* (árvores), *Cereal* (cereais), *Vegetable* (verduras e legumes), *Fruit* (frutas), *Other* (outros elementos de Flora não incluídos nos itens anteriores), *Product* (produtos derivados ou transformados, como cortiça, vinho, azeite, etc.)

Nos oitenta e sete textos foram anotadas 1068 ocorrências. Estas foram lematizadas e correspondem a 208 entidades distintas. Este procedimento de lematização é também uma tarefa fundamental, de modo a obter dados de ocorrências mais fiáveis, permitindo anular as flexões em género e número, neste caso.

Em **Fauna**, anotaram-se 53 entidades, que pertencem a três subcategorias: Bird, Mammal, Fish e *Product*. Não se encontraram nos textos em apreço outras. A subcategoria Mammal descreve mamíferos, produzidos tanto para consumo de carne/ leite (ovinos, caprinos, bovinos, suínos e leporídeos). como outros associados à pastorícia, como javali, lebre e veado; inclui também animais selvagens, como lobo, raposa e gineto/gato bravo. A subcategoria Fish contém 20 entidades, constituídas por peixes de rio.

No que respeita à **Flora**, foram anotadas 174 entidades, nas subcategorias Cereal, Fruit, Herb, Tree, Vegetable, Products e Other. A subcategoria que tem maior número de entidades é Herb (64), reunindo um conjunto de várias espécies, quer medicinais, quer o que hoje se chamam aromáticas, ou mesmo plantas selvagens. Este agregado fornece uma importante "radiografia" não só da existência e cultivo das espécies vegetais como dos usos que se faziam das plantas, por exemplo o tratamento de enfermidades (e.g. erva da erisipela), a sua direta implicação em atividades económicas (carrasco, que servia para a produção de tintas), práticas supersticiosas (e.g. arruda), etc. No que diz respeito à subcategoria Fruit, foram anotadas 43 entidades, entre frutos frescos para a alimentação (melão, ameixa, romã, melancia, maçã, etc.), frutos secos (passas de figo, amêndoa, noz) e leguminosas (feijão branco, feijão frade).

4 Análise dos Dados

Neste itinerário de trabalho, com tarefas sequenciais bem delimitadas, conhecer os dados é igualmente importante para definir domínios de atuação, tendo em vista normalizar com mais eficácia.

[2]https://inception-project.github.io/

Os dados do *corpus* anotado foram ordenados por subcategorias, de modo a poder ver-se quais as dominâncias e/ou casos isolados.

Na categoria **Fauna** só se encontraram três subcategorias: *Mammal*, *Fish* e *Bird*, conforme a tabela 1.

Categoria	Ocorrências	Entidades distintas
FAUN_MAMMAL	131	31
FAUN_FISH	129	20
FAUN_BIRD	21	2

Tabela 1: Distribuição de dados sobre Fauna

Na categoria **Flora**, ao invés, foram encontradas todas as sete subcategorias, hierarquizadas na tabela 2:

categoria	ocorrências	entidades distintas
FLORA_CEREAL	362	6
FLORA_FRUIT	110	43
FLORA_HERB	86	64
FLORA_TREE	71	25
FLORA_VEGET	66	18
FLORA_PROD	55	7
FLORA_OTHER	34	11

Tabela 2: Distribuição de dados sobre Flora

No que respeita às entidades, estas foram ordenadas por frequência decrescente, de que a Figura 2 é elucidativa:

Figura 2: Word cloud da totalidade de entidades anotadas

A subcategoria *Cereal* tem apenas 6 entidades distintas, mas é aquela que regista maior número de ocorrências (362). A segunda subcategoria mais frequente é *Mammal*, com 131 ocorrências e 31 entidades distintas. A terceira mais frequente é *Fish*, com 129 ocorrências. A elevada frequência destas entidades é reveladora da importância das atividades piscícolas nas bacias hidrográficas da região, tendo sido anotadas 20 entidades distintas. As denominações dos peixes foram uma das vertentes deste estudo que causaram maiores constrangimentos, conforme se discutirá adiante.

Fruit é a quarta subcategoria mais frequente, tendo sido anotadas 43 entidades distintas, que bem demonstram a variedade de espécimes existentes nestes concelhos.

A quinta subcategoria mais recorrente é *Herb*. Esta é a que contém maior número de entidades distintas e foi onde também o processo de normalização não foi fácil, como se discutirá mais à frente. A subcategoria com menor número de ocorrências é **Flora** *Other*.

Vejam-se agora as subcategorias em relação aos concelhos onde foram anotadas.

Em 1758, quase todas as freguesias ou paróquias do Alentejo teriam uma parte urbana e outra rural. A localização geográfica e características naturais influenciaram, certamente, o maior ou menor registo de entidades nestes domínios específicos da **Fauna** e **Flora**.

Observando ao nível dos concelhos quais as subcategorias mais frequentes verifica-se que, nos concelhos de Évora, Beja, Portalegre e Elvas, a subcategoria mais frequente é *Cereal*. A única exceção era constituída por Vila Viçosa, onde a subcategoria mais referida foi *Herb*, conforme Figura 3.

Inequivocamente, trigo é o cereal mais referido nestes textos, revelando a sua importância nestas regiões. Os cereais anotados, com o respetivo número de observações, resumem-se a centeio (55), cevada (99), pão (12), tremês (2) e trigo (194). Note-se que, na época, "pão" era equivalente a "cereal".

Ainda referente aos espécimes vegetais, os frutos são uma subcategoria com elevada frequência e grande número de entidades. Anotam-se diversos frutos para consumo humano. Destaca-se, ainda, a expressão "pêros de guarda", que se reportaria a uma espécie de maçã de maior durabilidade.

Os produtos derivados evidenciam a atividade económica destas regiões do sul de Portugal, algumas ainda hoje subsistentes. Em 1758, o azeite era o recurso com maior número de ocorrências (29), seguindo-se o vinho (16), a farinha (4) e a tinta (1).

As entidades da subcategoria **Flora** *Tree* têm elevadas frequências nos concelhos de Évora e Portalegre, e baixas frequências nos restantes concelhos. Englobam 26 entidades distintas. São sobretudo árvores com aproveitamento económico, como sobreiro, oliveira e azinho, ainda hoje de grande importância.

Os frutos surgem com elevada frequência em Elvas, contrastando com a baixa incidência em Évora e Beja. Ainda no presente, a “ameixa de Elvas”

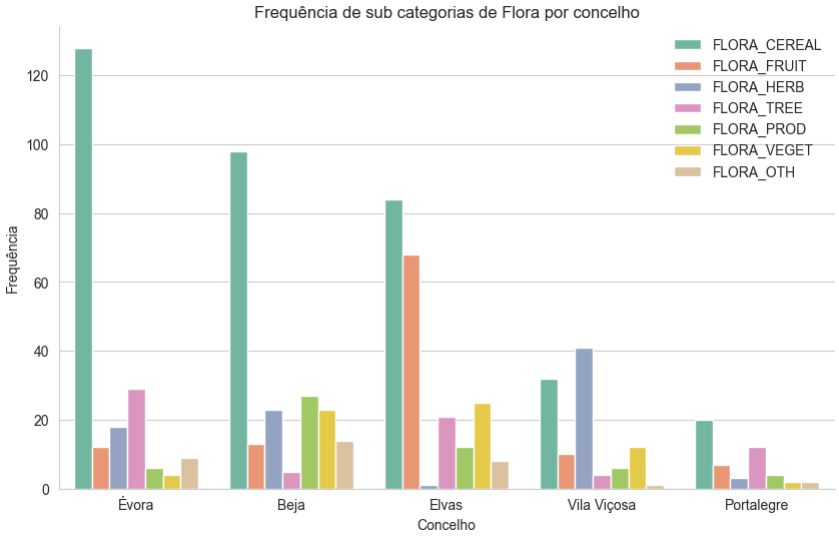

Figura 3: Subcategorias de Flora por concelho

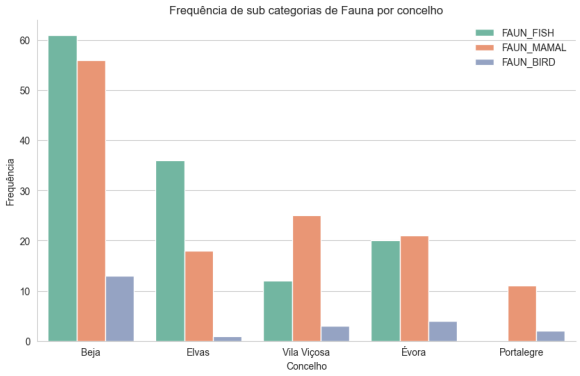

Figura 4: Subcategorias de Fauna por concelho

mantém-se relevante, afirmando-se como um produto de reconhecido valor económico.

O "vinho", atualmente com grande impacto na economia de todo o Alentejo, tinha um modesto número de registos no que foi o município de Portalegre em meados do século XVIII.

Relativamente aos espécimes animais, destacam-se os animais de criação, como “cabra”, “coelho” e “porco”, este último ainda hoje com elevado valor económico. Registam-se também animais selvagens, como a “perdiz”, espécie cinegética que continua a ser muito procurada. Entre os peixes de rio mais referidos, salientam-se o “bordalo” e a “boga”. Todos os concelhos têm anotações de Fauna, embora com diferente número de ocorrências, como se pode ver na Figura 4.

Relativamente às entidades da subcategoria Fish, estas têm maior número de ocorrências nos concelhos de Évora, Beja e Elvas. Em Vila Viçosa tem um número de ocorrências reduzido face aos concelhos anteriores; em Portalegre, é mesmo inexistente, o que é inverosímil e dever-se-á à falta de atenção dos párocos que responderam ao inquérito.

5 A variação lexical e os seus problemas

Em domínios de especialidade, como denominações geográficas, nomes vulgares de plantas e animais, há grande discrepância entre as formas do século XVIII e os equivalentes atuais para efeitos de normalização.

No *corpus* anotado que serviu de base para este estudo, pode observar-se que há uma relação direta entre maior número de entidades diferentes numa subcategoria e maiores constrangimentos em sede de normalização. As baixas frequências de vocabulário foram as que trouxeram maiores dificuldades. Estes constrangimentos podem ser devidos a vários fatores. O primeiro pode ser assumido como provável erro, ou do escrevente, ou do transcritor, obrigando a uma consulta do original manuscrito para verificação do termo.

O segundo constrangimento tem a ver com as próprias denominações. Os padres respondentes não seriam, certamente, botânicos, nem biólogos, pelo que as denominações de plantas e peixes nas *Memórias Paroquiais* poderão ser designações regionais, sem correspondência na atualidade.

A consulta a especialistas trouxe, igualmente, grandes desafios. Perante a inexistência de imagens quer das plantas quer dos animais, uma vez que a fonte é apenas textual, a classificação taxonómica dos espécimes torna-se mais complexa. Veja-se o exemplo de duas denominações de peixes: "combo

beiçudo" e "cabecinha". Estas poderão ser espécies de barbos, peixe muito frequente na bacia do rio Guadiana, mas estas denominações não existem na atualidade. Igualmente, algumas denominações podem referir-se ao estado juvenil destes peixes e não propriamente a uma subespécie.

No que respeita à Flora, algumas espécies de plantas não cultivadas carecem igualmente de verificação por especialistas. Um exemplo é a multiplicidade de entidades designadas por cardos, como cardo abrolho, cardo alvarinho, cardo arzol, cardo corredor, cardo rasteiro, cuja denominação atual (nome vulgar) tem de ser validada, uma vez que a correspondência destas denominações para a atualidade não é imediata.

Foi encontrado um registo de uma planta, "saisso", cuja regularização ortográfica não conseguiu ser validada para um possível registo ortográfico atual, já que esta denominação, tal como está, não existe em dicionários nem é conhecida pelos especialistas consultados.

Uma outra dificuldade, que é paralela à normalização gráfica mas que, em sede de enriquecimento de dados, tem de ser acautelada, tem a ver com variantes lexicais que designam a mesma espécie. Nos textos, no que respeita a animais, encontramos chibo/cabra, carneiro/borrego, gineto/gato bravo, raposa/zorra, designações que, no par, são consideradas equivalentes. Contudo, também aqui pode haver um uso regional ou uso da forma mais "culta" em detrimento da popular, por exemplo em chibo/cabrito. Para as espécies vegetais, são usados como variantes: bolota/lande, pão/cereal, azinho/ azinho-sobro/ azinho-sôvero, maçã/pêro, designações que, igualmente, podem admitir um uso regional.

Pensando-se que, sendo esta uma área de especialidade, o vocabulário seria mais limitado e mais preciso, mas verificou-se exatamente o contrário. Os padres, não sendo especialistas em botânica e zoologia, usariam termos comuns. Por outro lado, o registo escrito poderá estar ligado à oralidade pelo que poderá conter erros, ou usos regionais. Assim, em domínios de especialidade, a normalização (orto)gráfica não consegue ser apenas uma simples atualização da grafia requerendo uma intervenção de especialistas de domínio para confirmação do registo escrito normalizado a adotar, que não alterará nunca a variação lexical. No processo aqui descrito, esta verificação foi realizada na nomenclatura de dicionários autorizados de língua portuguesa e por especialistas nos domínios da Fauna e da Flora ou da Geografia. Ainda assim, algumas denominações, sobretudo de peixes e de plantas silvestres, não conseguiram reunir, nesta fase, a unanimidade dos especialistas, constituindo isto uma limitação ao trabalho desenvolvido.

Outro constrangimento teve a ver com a validação possível das espécies para uma futura constituição de datasets enriquecidos. Os especialistas tiveram acesso a dados apenas textuais, sem existência de nomes científicos, o que dificultou muito o trabalho de validação científica das denominações. Acresce que, atendendo a que apenas se analisaram freguesias de cinco concelhos, a existência de designações locais face a outras de outras regiões não conseguiu ser completamente comprovada neste estudo, constituindo igualmente uma limitação. Todavia, ficou o alerta. Em futuros trabalhos, com maior número de freguesias, estas entidades serão revalidadas. Não está posto de parte o recurso a trabalho colaborativo, aberto a quem conhece a realidade local, por vezes de escala micro, para ajudar a identificar estes recursos ou micro-topónimos; pode ser uma alternativa, embora algumas espécies possam já ter desaparecido.

6 Conclusão

Como se demonstrou, pensar em normalização automática de textos históricos, com grande abrangência temática, incluindo vocabulário de setores específicos, representa um grande desafio. Implica estar aberto a integrar especialistas em domínios que se situam muito para além das Humanidades. Isto torna-se bem evidente quando se pensa em abarcar especialidades como Fauna e Flora.

Normalizar está longe de ser uma tarefa de resultados imediatos. Exige ampla preparação para ser consistente. Normalizar textos históricos é mais do que um mero ajuste ortográfico. É um processo que exige um trabalho interdisciplinar prévio para desambiguar com segurança. Reveste-se de uma importância crucial para uma melhor compreensão dos próprios textos e da realidade histórica em questão. Para além disso, permite trazer para o presente textos de elevado valor patrimonial que, por via, passam a poder ser lidos por públicos generalistas e de especialidade.

Agradecimentos

Este trabalho é financiado por fundos nacionais através da Fundação para a Ciência e Tecnologia (FCT - Portugal), no âmbito do projeto

UID/00057/2025.

References

Gustavo Aguilar, Suraj Maharjan, Adrian Pastor López Monroy, and Thamar Solorio. 2017. A multi-task approach for named entity recognition in social media data. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 148–153.

Helena Freire Cameron, Fernanda Olival, and Renata Vieira. 2023. Planear a normalização automática: tipologia de variação gráfica do corpus das memórias paroquiais (1758). *LaborHistorico*, 9 (1):2359–6910.

Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The BDCamões collection of Portuguese literary documents: a research resource for digital humanities and language technology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 849–854, Marseille, France. European Language Resources Association.

Rafael Oleques Nunes, Joaquim Santos, André Spritzer, Dennis Giovani Balreira, Carla Dal Sasso Freitas, Fernanda Olival, Helena Freire Cameron, and Renata Vieira. 2025. *Assessing European and Brazilian Portuguese LLMs for NER in Specialised Domains*, volume 15412. Springer, Cham.

Paula Rodríguez-Puente, Cristina Blanco-García, and Iván Tamaredo. 2019. Annotation in the corpus of historical english law reports (chelar): Potential for historical genre analysis. *Journal of the Spanish Association for Anglo-American Studies*, 41 (2):63–84.

Joaquim Santos, Helena Freire Cameron, Fernanda Olival, Fátima Farrica, and Renata Vieira. 2024. Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 117–126.

Leonardo Zilio, Maria Jose Bocorny Finatto, and Renata Vieira. 2022. Named entity recognition applied to Portuguese texts from the 18th century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022) Virtual Event, Fortaleza, Brazil, CEUR Workshop Proceedings, v. 3128*.

Elena Álvarez Mellado, María Luisa Díez-Plata, Pablo Ruiz-Fabo, Helena Bermúdez, Salvador Ros, and Elena González-Blanco. 2021. Tei-friendly annonation scheme for medieval named entities: a case on a spanush medieval corpus. *Language Ressources and Evaluation*, 55 (2):525–549.