Corpus Memórias Paroquiais: Avanços em Reconhecimento de Entidades

Renata Vieira[1] **Helena Cameron**[1,2] **Fernanda Olival** [1] **Joaquim Santos**[3]

[1]Universidade de Évora, CIDEHUS
Évora, Portugal

[2]Instituto Politécnico de Portalegre
Portalegre, Portugal

[3]UNISINOS
São Leopoldo, Brazil

{renatav,helenafc,mfo}@uevora.pt, netojoaquim@edu.unisinos.br

Abstract. *This paper describes recent developments in NER on the Parish Memories historical corpus. The corpus has received new annotation categories for describing fauna and flora. A study about the adaptability of the model for dealing with original data without normalization is also discussed.*

Resumo. *Este artigo aborda os avanços recentes em REN no contexto do corpus Memórias Paroquiais. O corpus foi enriquecido com anotações adicionais que introduzem categorias dedicadas à fauna e flora. Além disso, é discutido um estudo sobre a adaptabilidade do modelo para lidar com dados originais sem normalização.*

1. Introdução

O *corpus* textual das *Memórias Paroquiais*, uma fonte histórica portuguesa que contém as respostas de párocos a um inquérito da coroa enviado para todo o país, 3 anos após o sismo de 1755, tem sido trabalhado ao longo de décadas. Por volta de 2008, de modo a tornar esta coleção acessível a um público mais alargado, e já antevendo a possibilidade de processamento computacional, os textos da parte sul de Portugal continental começaram a ser transcritos para suporte digital e acabaram disponíveis *online*. Mais recentemente, além da transcrição e revisão dos textos, tem sido desenvolvido um trabalho de normalização gráfica, passando os textos para a ortografia contemporânea e, desta forma, possibilitando o acesso a estes por um público mais amplo.

Com o objetivo de melhorar a pesquisa no corpus e facilitar a compilação de dados e análise da fonte para estudos históricos, o *corpus* tem sido anotado com entidades nomeadas. Neste artigo, pretende-se apresentar o trabalho desenvolvido de anotação de um novo grupo de categorias de entidades, a FAUNA e a FLORA, que se adicionaram às já existentes (Vieira et al. 2021; Olival et al. 2023; Santos et al. 2024; Nunes et al. 2025). Estas categorias são muito produtivas nestes textos e permitirão obter dados com utilidade não só para historiadores, como também para biólogos, geógrafos e outros investigadores.

Os modelos de Reconhecimento de Entidades Nomeadas (REN) têm sido desenvolvidos com base em versões normalizadas das Memórias. Dada a morosidade da tarefa

manual de normalização e o objetivo de estender a anotação de forma automatizada, investigamos a real necessidade de normalização prévia. Assim, os modelos foram também treinados em versões não normalizadas e, neste fórum, discute-se o comportamento dos modelos nas duas versões, de modo a aferir o impacto da variação linguística no seu reconhecimento.

2. Trabalhos Relacionados

O REN representa uma tarefa frequente no Processamento de Linguagem Natural, com uma série de estudos já realizados para o português (Albuquerque et al. 2023; Silva 2023). Embora a maior parte das iniciativas em REN se concentre na língua contemporânea, com o avanço das humanidades digitais, o reconhecimento de entidades nomeadas está ganhando importância crescente em pesquisas que lidam com fontes históricas.

Em (Ehrmann et al. 2023), apresenta-se um estudo sobre reconhecimento e classificação de entidades nomeadas em documentos históricos considerando uma variedade de línguas. Abordagens que consideram particularmente o português histórico são apresentados em: (Grilo et al. 2020; Aguilar et al. 2017; Zilio et al. 2022; Nunes et al. 2025; Santos et al. 2024).

Há vários estudos anteriores realizados com o *corpus* Memórias Paroquiais (MPs) (Vieira et al. 2021; Cameron et al. 2022; Nunes et al. 2025). O que agora se desenvolve é o primeiro a incluir as categorias Fauna e Flora na anotação. Neste trabalho, foi aplicado o melhor modelo de REN desenvolvido para as MPs (Nunes et al. 2025) até o momento. Esse modelo, desenvolvido com base na versão anterior do *corpus*, foi retreinado com novas categorias. Investiga-se ainda sua adaptabilidade para o português legítimo da época, ou seja, a partir de sua transcrição fiel, não normalizada.

3. *Memórias Paroquiais* e o sul de Portugal

As *Memórias Paroquiais* são constituídas por cerca de dois mil e seiscentos textos manuscritos e contêm as respostas dos párocos a um inquérito que versava sobre a Terra, a Serra e o Rio, e também sobre os efeitos do terramoto de Lisboa de 1755, cuja magnitude causou destruição a muitas centenas de quilómetros. Cada pároco respondia sobre a sua paróquia e dava informações preciosas sobre o território, a sua organização, as gentes e os costumes de Portugal em meados do século XVIII (1758-1761).

Este conjunto documental, de grande relevância histórica e linguística, sofreu algumas alterações ao longo da sua história custodial e até na sua composição. Primeiro, no século XIX, os vários textos foram organizados em formato de dicionário e encadernados; como se tinham perdido algumas Memórias, foram acrescentados dois volumes de suplemento (vol.s 42-43) e um de índices do conjunto (vol. 44), com um prólogo subscrito em 1832. No século XX, já no Arquivo Nacional da Torre do Tombo, os textos foram microfilmados (entre 1993 e 2003) e, mais recentemente, digitalizados a partir do microfilme (2005). Os volumes físicos deixaram de estar disponíveis para consulta, o que acarreta desafios acrescidos para os leitores, quando se confrontam com imagens incompletas, qualidade insuficiente ou repasses de tinta dificultando a leitura humana. Ainda que as tentativas de transcrição e edição tenham

começado cedo (Madahil 1937), no século XXI apareceram as de maior envergadura: (Capela 2003; Cosme and Varandas 2009; Rodrigues and Neto 2012).

Os textos das *Memórias* relativas à parte a sul do rio Tejo, que abarcava a Península de Setúbal, a província do Alentejo e o Reino do Algarve, representam cerca de 20% do total dos textos, muito embora correspondam a cerca de 42,8% do território continental de Portugal. Trata-se de zonas com freguesias/paróquias de grande extensão em superfície, mas fracamente povoadas, especialmente no Alentejo e no Algarve rurais, o que explica este desfasamento. Em cada texto, o pároco indica detalhes da organização administrativo-eclesiástica de cada paróquia, o local preciso de localização da mesma no âmbito concelhio e comarcão, bem como dados demográficos e outras características da terra. Estas informações são um verdadeiro desafio para historiadores, já que boa parte desta realidade já não existe, por força de mudanças administrativas implementadas desde então.

3.1. Fauna e Flora

Numa primeira fase de anotação das MPs, foram estabelecidas grandes categorias, adequadas a uma realidade pretérita, que respondiam às questões gerais do historiador: Quem (PER; AUTWORK)?, Quando (TIM_CRON)?, Onde (PLC; ORG)? (Cameron et al. 2022). Foram estabelecidas Guidelines e, nesta fase, anotaram-se apenas nomes próprios. Anotou-se sempre a maior expressão, permitindo incluir títulos e ocupações que frequentemente acompanhavam os nomes. O processo de anotação manual foi desenvolvido de forma síncrona entre todos os membros da equipe. Posteriormente, uma das historiadoras prosseguiu com a tarefa, trazendo as dúvidas para a discussão multidisciplinar. Numa segunda fase, dada a relevância dos produtos naturais e agropecuários para a economia da região, anotaram-se mais duas grandes categorias: FAUNA e FLORA, anotando nomes comuns. Tal como em outros casos, estes dois grandes descritores foram desdobrados em subcategorias. A categoria FAUNA foi subdividida em sete subcategorias, sendo seis relativas a animais (FAUN_FISH, FAUN_BIRD, FAUN_MAMAL, FAUN_REPTIL, FAUN_INSECT, e FAUN_OTH, para anfíbios, ou outras espécies que não estejam incluídas nos itens anteriores) e uma sétima (FAUN_PROD), relativa a produtos de origem animal transformados, como por exemplo ”lã”, ”mel”, ”couro”, etc. No que diz respeito à grande categoria FLORA, esta foi desdobrada igualmente em sete subcategorias, seis que classificam plantas (FLORA_HERB, FLORA_TREE, FLORA_CEREAL, FLORA_VEGET, FLORA_FRUIT, para ervas, árvores, cereais, vegetais, frutos, e também FLORA_OTH, para algum elemento que não consiga ser classificado nas subcategorias anteriores), e uma sétima subcategoria para produtos transformados (FLORA_PROD), como por exemplo ”azeite”, ”vinho”, etc.

3.2. Normalização gráfica

A normalização gráfica é uma tarefa exigente, que requer sólidos conhecimentos de linguística histórica, aprofundado conhecimento da língua e domínio do período histórico em questão. Contudo, é uma tarefa que parece não ser considerada essencial para alguns públicos: os linguistas conseguem ler em qualquer estádio da língua e os historiadores também lidam com formas pré-contemporâneas com facilidade. Mas, quer para historiadores ou mesmo linguistas em fases mais precoces do seu percurso formativo e, portanto, com menos experiência, ou para investigadores oriundos de outras áreas científicas, o

acesso a textos históricos em ortografia da época pode representar um embaraço; dificulta-lhes o acesso ao conteúdo na sua plenitude. Com efeito, a normalização gráfica pode ser um importante contributo para uma maior divulgação científica, permitindo que qualquer pessoa leia textos pretéritos em ortografia contemporânea. Contudo, não pode ser encarada de ânimo leve, sob pena de desvirtuar as características do discurso original. Deverá apenas limitar-se a uma atualização da ortografia e a uma normalização do uso das maiúsculas segundo a norma em vigor. Nos casos que requeiram uma pontual normalização lexical, como por exemplo a atualização de nomes próprios ou comuns desusados, esta deverá ainda ser mais prudente, devendo sempre manter toda a variância linguística da língua (e.g. ouro/oiro). As palavras antigas que estão ainda em uso (e.g. mui, el-rei, cousa, ...) não devem ser normalizadas, uma vez que ainda integram o discurso contemporâneo em domínios e contextos específicos.

Para o PLN, a normalização é um processo intermédio de pré-processamento que pode contribuir para uma maior precisão nos resultados, diminuindo as numerosas variantes ortográficas características do século XVIII, ou padronizando o uso de maiúsculas, que podem aparecer no meio de palavras, ou desdobrando palavras que, aquando da transcrição, estavam incompletas ou ilegíveis. Têm sido feitos alguns esforços de desenvolvimento de ferramentas capazes de automatizar a normalização de textos pré-contemporâneos em algumas línguas europeias. Citem-se alguns: (Baron and Rayson ; Burns 2013; Amoia and Martinez 2013; Pettersson et al. ; Samardžić et al. 2015; Bollmann and Søgaard 2016), entre outros igualmente pertinentes. Para a língua portuguesa, tanto quanto se sabe, não foi ainda desenvolvida nenhuma ferramenta capaz de normalizar textos pré-contemporâneos com resultados considerados satisfatórios. A normalização é, assim, uma tarefa que tem sido feita de forma manual, com grande investimento humano.

4. Experimentos

CATEG	P	R	F1
AUTWORK	70.00	66.67	68.29
ORG	64.29	65.45	67.86
TIM_CRON	65.28	70.15	67.63
PER_AUT	83.33	93.75	88.24
PER_CAT	53.33	100.00	69.57
PER_DIV	87.80	90.00	88.89
PER_NAM	68.59	76.98	72.54
PER_OCC	70.37	76.00	73.08
PER_PGRP	69.57	76.19	72.73
PER_SAINT	77.21	78.36	77.78
PLC_AQU	80.00	74.29	77.04
PLC_FAC	67.14	64.38	65.73
PLC_GPE	78.45	78.11	78.28
PLC_LOC	65.38	76.40	70.47
PLC_MOUNT	80.00	92.31	85.71

Tabela 1. Resultados reportados em (Nunes et al. 2025)

Em (Nunes et al. 2025) foram testadas alternativas de modelos de aprendizado para o *corpus* MPs, sendo que o modelo baseado em Albertina PT-PT (Rodrigues et al. 2023) foi o que apresentou melhores resultados, ver Tabela 1.

Albertina PT-PT é um modelo de linguagem da família BERT, desenvolvido para o português europeu. Esse modelo com melhor desempenho foi então aplicado para treino na nova versão do *corpus* MPs que inclui as anotações de Fauna e Flora, que se discutem neste artigo.

As categorias compreendem, obra de autor, organização, tempo cronológico, pessoa e local, os dois últimos com várias especializações. Em pessoa foram consideradas menções específicas a autores, categoria social, divindade, nome próprio, ocupação, grupos e santos, enquanto que para local, foram consideradas as menções a aquíferos, edificações, entidades geo políticas, lugares em geral e montanhas, indentificados respectivamente na ordem da tabela.

4.1. Experimento 1 - Aprendizado de FAUNA e FLORA

Tipo de Entidade	Treino	Teste	Dev	Total Geral
AUTWORK	107	23	7	137
ORG	273	82	37	392
TIM_CRON	220	78	19	317
FAUN_BIRD	17	3	3	23
FAUN_FISH	48	28	9	85
FAUN_MAMAL	67	29	16	112
FLORA_CEREAL	203	46	25	274
FLORA_FRUIT	31	9	1	41
FLORA_HERB	36	2	40	78
FLORA_OTH	15	3	7	25
FLORA_PROD	31	6	4	41
FLORA_TREE	36	10	4	50
FLORA_VEGET	34	4	1	39
PER_AUT	95	25	9	129
PER_CAT	32	11	6	49
PER_DIV	113	41	23	177
PER_NAM	503	151	65	719
PER_OCC	92	20	11	123
PER_PGRP	131	36	28	195
PER_SAINT	417	134	64	615
PLC_AQU	168	38	19	225
PLC_FAC	206	48	38	292
PLC_GPE	796	197	88	1081
PLC_LOC	303	90	56	449
PLC_MOUNT	50	17	3	70
TOTAL DE B-TAGS	**4024**	**1131**	**583**	**5738**

Tabela 2. Distribuição em treino, desenvolvimento e teste

Neste trabalho, as novas categorias de Fauna e Flora são o objeto de análise.

Primeiramente, observamos a distribuição para treino e teste (split) para todas as categorias, ver Tabela 2. A tabela apresenta a distribuição, com base na contagem de B-Tags, ou seja, as etiquetas que indicam o início de uma expressão representando uma entidade, e ilustra a pouca disponibilidade de exemplos para as novas categorias de fauna e flora. No entanto, esta é uma característica natural do *corpus*, e para o aumento de exemplos seria necessário ampliar o número de textos anotados.

CATEG	P	R	F1
AUTWORK	57.89	47.83	52.38
ORG	60.81	54.88	57.69
TIM_CRON	72.09	79.49	75.61
PER_AUT	88.00	88.00	88.00
PER_CAT	80.00	72.73	76.19
PER_DIV	82.50	80.49	81.48
PER_NAM	78.57	79.61	79.08
PER_OCC	65.22	75.00	69.77
PER_PGRP	74.29	72.22	73.24
PER_SAINT	84.62	73.88	78.88
PLC_AQU	77.50	81.58	79.49
PLC_FAC	60.00	75.00	66.67
PLC_GPE	77.99	82.74	80.30
PLC_LOC	60.19	72.22	65.66
PLC_MOUNT	73.68	82.35	77.78

Tabela 3. Experimento reproduzido no corpus atualizado

CATEG	P	R	F1	Suporte
FAUN_BIRD	100.00	66.67	80.00	2
FAUN_FISH	90.00	96.43	93.10	30
FAUN_MAMAL	74.29	89.66	81.25	35
FLORA_CEREAL	79.31	100.00	88.46	58
FLORA_FRUIT	100.00	33.33	50.00	3
FLORA_HERB	0.00	0.00	0.00	0
FLORA_OTH	33.33	33.33	33.33	3
FLORA_PROD	75.00	100.00	85.71	8
FLORA_TREE	60.00	90.00	72.00	15
FLORA_VEGET	50.00	100.00	66.67	8

Tabela 4. Resultados para as categorias Fauna e Flora

A Tabela 3 apresenta os resultados do modelo desenvolvido em (Nunes et al. 2025) para o novo *corpus* para as categorias previamente existentes. Nota-se diferença de valores que pode ser explicada pelas variações nas partições realizadas no *corpus* ou pela influência das novas categorias. Na tabela 4 apresentam-se as categorias de interesse específico desse artigo. O suporte representa o número de casos identificados no teste.Temos uma distribuição não balanceada das classes, mas apenas três classes apresentam F1 menor ou igual a 50%, todas com poucos casos de suporte (3

ou menos). Se considerarmos as categorias com no mínimo 15 casos de suporte, o menor F1 é de 72%. Uma das categorias não obteve casos de teste e ficou com avaliação nula.

4.2. Experimento 2 - Avaliação do modelo em textos não normalizados

O mesmo modelo treinado no experimento 1 foi aplicado a um novo conjunto de textos com as versões normalizadas e não normalizadas do *corpus* MPs, com o objetivo de investigar o impacto da normalização no reconhecimento. Aqui avaliamos a diferença no número de categorias produzidas para cada versão, em um conjunto de 16 textos, Tabela 5. Foram contabilizadas as etiquetas B, que indicam o início de uma expressão identificada na categoria, que pode ou não ser multi-palavra. O modelo comportou-se de forma satisfatória transversalmente às várias categorias. No exemplo abaixo pode verificar-se o reconhecimento de uma entidade da classe PLACE_FAC, que anota edificações, onde ocorre um termo com ortografia do século XVIII no meio da expressão:

```
Igreja      B-PLC_FAC
Parochial   I-PLC_FAC
de          I-PLC_FAC
São         I-PLC_FAC
Pedro       I-PLC_FAC
da          I-PLC_FAC
cidade      I-PLC_FAC
de          I-PLC_FAC
Elvas       I-PLC_FAC
```

Observa-se, igualmente, que as variações do contexto também não impediram o reconhecimento correto de duas entidades dessa categoria.

```
se
vay
seguindo
the
o
convento    B-PLC_FAC
de          I-PLC_FAC
Santa       I-PLC_FAC
Clara       I-PLC_FAC
e
igreja      B-PLC_FAC
dos         I-PLC_FAC
Terceyros   I-PLC_FAC
de          I-PLC_FAC
Sam         I-PLC_FAC
Francisco   I-PLC_FAC
,
sempre
subindo
asima
the
o
castello
```

CATEG	Normalizadas	Não normalizadas
B-FAUN_BIRD	2	1
B-FAUN_FISH	42	24
B-FAUN_MAMAL	21	13
B-FLORA_CEREAL	85	71
B-FLORA_FRUIT	48	53
B-FLORA_HERB	1	2
B-FLORA_PROD	9	5
B-FLORA_TREE	20	17
B-FLORA_VEGET	8	13
B-FLORA_OTH	5	0

Tabela 5. Número de entidades para Fauna e Flora identificadas na versão normalizada e não normalizada

Analisando as expressões não reconhecidas pelo modelo nas várias categorias, investigou-se uma eventual correlação entre elementos gráficos e ortográficos distintivos e o reconhecimento do modelo.

• **Maior número de expressões anotadas na versão normalizada**

Relativamente às duas categorias em análise no experimento 1, verifica-se que as subcategorias FAUN_FISH, FLORA_CEREAL e FLORA_TREE são as que têm maior número de expressões anotadas e com maior número de expressões anotadas na versão normalizada face à outra. Verificou-se que o modelo não consegue reconhecer as expressões ”pexe” e ”peyxe” (não normalizada) na categoria FAUN_FISH, facto que pode explicar a diminuição do número de expressões na versão não normalizada. Por outro lado, a expressão no plural ”pexes”, com mais um caracter e em posição de final de palavra, conseguiu ser reconhecida e recebeu a etiqueta correta. A existência de consoantes dobradas não pareceu ser um impedimento ao reconhecimento nesta categoria, em que ”bordallo” foi corretamente reconhecido como FAUN_FISH. Nas expressões com uma variância gráfica de 2 ou mais caracteres, o modelo parece não reconhecer o termo, como o nome do peixe ”samdiaez”, (”sandiais” na versão normalizada). Na subcategoria FLORA_CEREAL, verificou-se que expressões com alternância gráfica ”s/c” em início de palavra, por exemplo ”sevada” e ”senteyo”, não foram reconhecidas, pelo que a diferença gráfica entre a normalizada e a não normalizada parece, aqui, ter tido grande impacto. Quanto à subcategoria FLORA_TREE, não foi encontrado um padrão ortográfico explicativo, em que ”oliveyras” foi corretamente reconhecido e etiquetado mas ”azynheira” não o foi. Também se encontram casos em que as versões são iguais, por exemplo em ”olivais”, mas a expressão não foi reconhecida pelo modelo.

• **Maior número de expressões anotadas na versão original não normalizada**

Vejam-se os casos de FLORA_FRUIT e FLORA_VEGET em que, ao contrário das restantes subcategorias, o número de expressões reconhecidas aumenta na versão não normalizada. Face ao reduzido número de ocorrências, não se avaliou FLORA_HERB. Há expressões anotadas na versão normalizada com vários *tokens*, alguns com um separador dentro da expressão, como a vírgula. Nestes casos, o modelo nem sempre reconhece a expressão inteira, usando a vírgula como separador, pelo que o modelo devolve a anotação

separada em duas ou mais expressões. Estes casos podem explicar o aumento do número de expressões anotadas. Mas também se encontram expressões não anotadas, como ”ortaliça”, ”amecha” e ”figueyra”, face aos termos normalizados ”hortaliça”, ”ameixa” e ”figueira” anotados, com elementos distintivos gráficos: o ”h” em início de palavra e, em ”amecha”, mais do que dois caracteres diferentes face ao termo normalizado. Quanto a ”figueyra”, tentou verificar-se um eventual impacto da grafia no reconhecimento de palavras com a letra ”y” dentro do *token*. Observou-se que este uso é transversal às várias categorias, encontrando-se casos de palavras com ”y” que são muitas vezes reconhecidas, como por exemplo ”ribeyro”, com a etiqueta PLC_AQU, face a outras semelhantes que não são reconhecidas, como no exemplo dado.

5. Discussão

O alargamento da anotação de EN às categorias FAUNA e FLORA apresenta características um pouco diferentes, face aos estudos prévios de REN. O modelo foi treinado com, respetivamente, sete subcategorias, que permitem classificar os animais e as plantas de acordo com uma tipologia genérica mas distintiva: peixe, pássaro, mamífero, réptil, inseto, e outros que não consigam ser classificados nas subcategorias anteriores e, para as plantas, ervas, árvores, cereais, vegetais, frutos e outros não incluíveis nas anteriores. Em ambas as categorias, foi incluída uma sétima subcategoria, que etiqueta produtos transformados, derivados ou produzidos a partir de animais e de plantas. Estas etiquetas anotam expressões que não são consideradas nomes próprios, o que não acontecia nos estudos prévios. Por outro lado, estamos perante expressões com menor ambiguidade no discurso, o que também não acontecia noutras categorias, em que um mesmo nome próprio poderia ser o nome de um santo, o nome de uma freguesia e o nome de igreja, etc.

Este experimento teve como dificuldade o menor número de expressões anotadas no treino, uma vez que nem todos os textos continham expressões anotáveis com estas categorias. Ainda assim, verificou-se que o modelo conseguiu ter um comportamento considerado satisfatório, com uma precisão que varia entre 33% (FLORA_OTH) e 100% (FAUN_BIRD). O Recall mais elevado (100%) foi atingido em duas subcategorias de FAUNA, CEREAL e PROD, e o valor mais baixo em FLORA_FRUIT. O F1 varia entre 33,33% de FLORA_OTH e 93,10% de FAUN_FISH. Estes valores permitem sustentar um alargamento do REN nestas duas categorias à totalidade do corpus MPs, permitindo vir a obter mais dados de elevado valor histórico, geográfico e no âmbito da Biologia e Botânica, capazes de contribuir para um maior conhecimento desta região naquela época.

O segundo experimento teve como objetivo inicial medir a eventual interferência das diferenças ortográficas no REN, permitindo avaliar o comportamento do modelo em ambas as versões, normalizada e não normalizada.

- **Nº de caracteres diferentes no token**

Verificou-se que a diferença entre as versões parecia ter tendencialmente maior impacto no início da expressão, no B, especialmente quando se verificava que o número de caracteres diferentes no *token* era de 2 ou mais. Nos casos em que a diferença, no B, era de apenas um caracter, podemos distinguir várias situações: no caso de consoante dobrada, o modelo reconheceu boa parte das expressões, etiquetando corretamente, por exemplo, ”bollotas” (FLORA_FRUIT) ou ”bordallos” (FAUN_FISH) mas este reconhecimento não foi sistemático, não etiquetando por exemplo ”cavallos” (FAUN_MAMAL).

• **"Y" com valor de semi-vogal**

Uma das características linguísticas que se crê com impacto em REN é o uso de "y" com valor de "i" semi-vogal no meio do *token*, verificando-se que grande parte dos *tokens* com "y" não são reconhecidos pelo modelo, como por exemplo em "azeyte" (FLORA_PROD), "azinheyra" (FLORA_TREE), "peyxe" (FAUN_FISH), entre outros. Este problema é transversal a todas as restantes categorias, mesmo as com maior número de expressões anotadas, como por exemplo em "Reyno de Castela" (PLC_GPE), ou "aldeya de Tellena do termo de Badajos" (PLC_LOC), entre outros exemplos igualmente elucidativos que não foram reconhecidos pelo modelo.

• **Alternância gráfica "c/s", "s/z" e "g/j"**

Igualmente, a representação de "c/s", e "s/z" em início de palavra ou no meio do *token* no B parece ter grande impacto em REN, nas duas categorias do experimento 1 e transversalmente nas restantes. As expressões "sevada" (FLORA_CEREAL) e "senteyo" (FLORA_CEREAL) não são reconhecidas na versão não normalizada, o mesmo sucedendo noutras categorias, como "Marquez do Lavradio" (PER_NAM), "freguezia de São Pedro" (PLC_GPE), "religiozos dominicos" (PER_GRGP) entre outros exemplos.

A localização no token da disparidade gráfica é um aspeto a estudar de forma mais aprofundada em futuro estudo, também com uma base lexical mais alargada. Verifica-se que, quando a disparidade se localiza no início, como no exemplo dado, o modelo quase sempre não consegue associar as duas variâncias. Contudo, tal não é sistemático: no *corpus*, o modelo conseguiu reconhecer corretamente "orta de Dom Alvoro", apesar da variância do "h" inicial. Também a alternância "g/j" parece ter impacto em REN, em que "majestade", termo do B, não é reconhecido na versão não normalizada.

6. Conclusão

O alargamento das duas novas categorias de EN e o treino realizado do modelo permitem confirmar um desempenho do modelo satisfatório, com parâmetros de avaliação que podem ser considerados bons e, nalgumas subcategorias, muito bons.

No que respeita ao impacto da ortografia em REN, obtiveram-se dados muito interessantes que permitem sustentar que a diferença ortográfica entre as versões parece ter impacto no reconhecimento efetuado pelo modelo, mas não de modo uniforme, conforme apresentado. O que se revelou decisivo também foi a localização da diferença gráfica que, se no B, pode impedir o reconhecimento, mas não tendo impacto se dentro da expressão.

O impacto da diferença gráfica e ortográfica entre versões não é negligenciável, o que confirma, por um lado, a necessidade de normalização prévia ao processamento. Por outro lado, reforça a necessidade de desenvolvimento de ferramentas normalizadoras para o português, capazes de acelerar esta tarefa manual. Pretende-se o alargamento deste estudo a um *corpus* de maior dimensão com outras épocas pré-contemporâneas, que fornecerá elementos adicionais a este fenómeno, que pouco tem sido referido em estudos semelhantes.

Referências

[Aguilar et al. 2017] Aguilar, G., Maharjan, S., Monroy, A. P. L., and Solorio, T. (2017). A multi-task approach for named entity recognition in social media data. In *Proceed-*

ings of the 3rd Workshop on Noisy User-generated Text, pages 148–153.

[Albuquerque et al. 2023] Albuquerque, H. O., Souza, E., Gomes, C., Pinto, M. H. d. C., Ricardo Filho, P., Costa, R., Lopes, V. T. d. M., da Silva, N. F., de Carvalho, A. C., and Oliveira, A. L. (2023). Named entity recognition: a survey for the portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

[Amoia and Martinez 2013] Amoia, M. and Martinez, J. M. (2013). Using comparable collections of historical texts for building a diachronic dictionary for spelling normalization. In *Proceedings of the 7th workshop on language technology for cultural heritage, social sciences, and humanities*, pages 84–89.

[Baron and Rayson] Baron, A. and Rayson, P. Vard2: A tool for dealing with spelling variation in historical corpora. *In Postgraduate conference in corpus linguistics*.

[Bollmann and Søgaard 2016] Bollmann, M. and Søgaard, A. (2016). Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. *arXiv preprint arXiv:1610.07844*.

[Burns 2013] Burns, P. R. (2013). Morphadorner v2: A Java library for the morphological adornment of English language texts. *Northwestern University, Evanston, IL*.

[Cameron et al. 2022] Cameron, H. F., Olival, F., Vieira, R., and Neto, J. F. S. (2022). Named entity annotation of an 18th century transcribed corpus: problems, challenges. In Trojahn, C., Finatto, M. J., de Paiva, V., and Vieira, R., editors, *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022*, volume 3128 of *CEUR Workshop Proceedings*, pages 18–25. CEUR-WS.org.

[Capela 2003] Capela, J. V. (2003). *Freguesias do Distrito de Braga nas Memorias Paroquiais de 1758*. Universidade do Minho.

[Cosme and Varandas 2009] Cosme, J. and Varandas, J. (2009). *Memórias Paroquiais (1758), v.1*. Caleidoscópio XVIII, 517pp edition.

[Ehrmann et al. 2023] Ehrmann, M., Hamdi, A., Pontes, E. L., Romanello, M., and Doucet, A. (2023). Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

[Grilo et al. 2020] Grilo, S., Bolrinha, M., Silva, J., Vaz, R., and Branco, A. (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.

[Madahil 1937] Madahil, A. R. (1937). Informações paroquiais do distrito de aveiro de 1721. In de Aveiro, A. D., editor, *Arquivo do Distrito de Aveiro, Vol. III*.

[Nunes et al. 2025] Nunes, R. O., Santos, J., Spritzer, A., Balreira, D. G., Freitas, C. M. D. S., Olival, F., Cameron, H. F., and Vieira, R. (2025). Assessing European and Brazilian Portuguese LLMs for NER in specialised domains. In *Brazilian Conference on Intelligent Systems*, pages 215–230. Springer.

[Olival et al. 2023] Olival, F., Cameron, H. F., and Vieira, R. (2023). As Memórias Paroquiais: do manuscrito ao digital. *Atas da Jornada de Humanidades Digitais do CIDEHUS, Universidade de Évora*.

[Pettersson et al.] Pettersson, E., Megyesi, B., and Tiedemann, J. An SMT approach to automatic annotation of historical text. *In Proceedings of the workshop on computa-*

tional historical linguistics at NODAL- IDA 2013; May 22-24; 2013; Oslo; Norway. NEALT Proceedings Series 18, 087, pages 54–69. Linkoping University Electronic Press.

[Rodrigues et al. 2023] Rodrigues, J., Gomes, L., Silva, J., Branco, A., Santos, R., Cardoso, H. L., and Osório, T. (2023). Advancing neural encoding of Portuguese with transformer Albertina pt. In *EPIA Conference on Artificial Intelligence*, pages 441–453. Springer.

[Rodrigues and Neto 2012] Rodrigues, M. R. S. and Neto, M. S. (2012). *Informações paroquiais e história local: a diocese de Coimbra (século XVIII)*. Palimage Editores.

[Samardžić et al. 2015] Samardžić, T., Scherrer, Y., and Glaser, E. (2015). Normalising orthographic and dialectal variants for the automatic processing of Swiss German. In *Proceedings of the 7th Language and Technology Conference*, pages 294–298. University of Zurich.

[Santos et al. 2024] Santos, J., Cameron, H. F., Olival, F., Farrica, F., and Vieira, R. (2024). Named entity recognition specialised for Portuguese 18th-century history research. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 117–126, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

[Silva 2023] Silva, A. V. (2023). Uma revis£o para o reconhecimento de entidades nomeadas aplicado † lngua portuguesa. *Linguam°tica*, 15(2):69–85.

[Vieira et al. 2021] Vieira, R., Olival, F., Cameron, H., Santos, J., Sequeira, O., and Santos, I. (2021). Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*, 7:20.

[Zilio et al. 2022] Zilio, L., Finatto, M. J. B., and Vieira, R. (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*.