Proposta recebida em Julho 2020 e aceite para publicação em Dezembro 2020.

Avaliando entidades mencionadas na coleção ELTeC-por

Assessing named entities in the ELTeC-por collection

Diana Santos
Linguateca & Universidade de Oslo
d.s.m.santos@ilos.uio.no

Eckhard Bick
South Denmark University
eckhard.bick@mail.dk

Marcin Wlodek
Linguateca
martimwlodek@hotmail.com

Resumo

Este artigo relata a preparação da anotação da coleção ELTeC-por com entidades mencionadas apropriadas ao género textual "romances e novelas publicadas entre 1840 e 1920", para possibilitar a leitura distante em português.

Em primeiro lugar apresentamos a coleção ELTeC-por, compilada no âmbito da ação COST "Distant Reading for European Literary History" para estudar a literatura europeia, e explicamos as diversas restrições e escolhas necessárias, fornecendo uma caracterização inicial segundo vários eixos: a origem e tamanho das obras, o seu (sub)género literário, o género do autor, o local de publicação e a existência ou não de mais edições.

Em seguida apresentamos o sistema PALAVRAS-NER, com o qual anotaremos a coleção, explicando detalhadamente o seu funcionamento.

Passamos então à descrição da criação de uma subcoleção de oito obras revistas, que servem, por um lado, para avaliar o desempenho do sistema de REM automático, e, por outro, para caracterizar o tipo de população esperada. As obras podem classificar-se segundo dois eixos diferentes: romances históricos vs. romances contemporâneos; e obras com grafia original ou grafia modernizada. Além disso, algumas obras são obviamente canónicas, outras não.

Além da descrição quantitativa do resultado de anotação e revisão, apresentamos algumas considerações qualitativas sobre o processo.

Também fornecemos uma análise detalhada de algumas categorias, tentando mostrar como os lugares, profissões e gentílicos mais mencionados podem ser indicadores numa leitura distante.

Concluímos comparando com o trabalho internacional feito na análise de entidades mencionadas de obras literárias, explicando as diferenças e sugerindo trabalho futuro.

Palavras chave

leitura distante, reconhecimento de entidades mencionadas, português, literatura portuguesa, humanidades digitais, compilação de corpos

Abstract

This paper reports on the NER annotation of the ELTeC-por collection, a collection of hundred Portuguese novels published between 1840 and 1920, compiled in the scope of the COST action "Distant reading for European literary history".

In addition to discussing its compilation, the choices taken and what remains to be done, we provide an initial characterization of the novels according to size, subgenre, publication place, author gender and which edition was used.

Then we present PALAVRAS-NER, the NER system which we use to annotate the collection, explaining the way it works.

We then focus on a subcollection of eight novels fully human revised, which we use to both evaluate the performance of the automatic system, and to characterize the population of the full collection. These novels can be further subdivided according to two different features: historical versus contemporary novels, on the one hand, and original vs. modernized ortography, on the other. Also some works are canonical while others are not.

In addition to the quantitative analysis of the annotation results and process, we present some qualitative description of the human revision as well.

We offer a detailed analysis of some categories, demonstrating how the most mentioned places, professions and demonyms can be good indicators for distant reading.

We end the paper comparing briefly with other work using named entities for literary texts and suggesting future work.

Keywords

distant reading, named entity recognition, Portuguese, Portuguese literature, digital humanities, corpus compilation

1. Introdução

Este trabalho foi desenvolvido no âmbito da ação COST "Distant reading for European literary history" (CA16204)[1], que tem por objetivo re-

[1] Veja-se https://www.distant-reading.net/

DOI: 10.21814/lm.12.2.336

volucionar a história da literatura na Europa (ou, pelo menos, a história do romance e novela) através da aplicação de métodos empíricos a uma coleção multilingue de várias literaturas. Uma destas é a portuguesa, e daí o presente artigo.

Apesar de termos batalhado por uma coleção que refletisse a literatura lusófona (Santos et al., 2018), acabámos por construir duas coleções: uma que contivesse obras apenas de literatura portuguesa, para seguir o padrão do ELTeC, a chamada coleção nuclear, ELTeC-por, tal como foi feito para o espanhol e para o inglês, e a coleção ELTeC-por-ext, ou seja, uma coleção alargada, que contém (ainda em andamento) também obras brasileiras, e obras que por outros motivos não se enquadram nos critérios da coleção padrão, quer por terem dimensões demasiado reduzidas ou por serem escritas pelos mesmos autores (relembramos que cada autor pode ter no máximo três obras). A coleção alargada é, de facto, o conjunto da ElTeC-por e da ElTeC-por-ext, mas por razões óbvias não duplicamos as obras.[2]

Com efeito, as regras para a construção das coleções mínimas, que contêm ou deverão conter cem obras[3], são as seguintes (ver a documentação oficial do ELTeC (2018)):

- A coleção apenas contém romances ou novelas publicadas na Europa entre 1840 e 1919[4], para que as obras pertençam ao domínio público.
- Devemos procurar um equilíbrio entre os seguintes parâmetros: vintena em que a obra foi publicada (1840–1859; 1860–1879; 1880–1899; 1900–1920); tamanho da obra — tendo sido definidos os seguintes intervalos com base no número de palavras: pequena (entre 10.000 a 50.000 palavras), média (entre 50.000 e 100.000 palavras) e grande, com um tamanho maior do que 100.000 palavras; e canonicidade: além de obras que pertencem ao cânone, devem constar muitas que não o fazem. Escusado será dizer que esta questão provocou muita celeuma, visto que há muitas formas de compreender o cânone, e que acabou por ser parafraseada pelo critério objetivo "tem mais do que uma reedição no período 1980–2010", que pode ter o valor sim ou não.
- No máximo 11 e no mínimo 9 autores devem ter três obras na coleção.
- Entre 10 a 50% das obras devem ser escritas por mulheres.

Dentro destes parâmetros, cada grupo dedicado a uma literatura[5] tem de fazer as escolhas que lhe pareçam mais apropriadas, visto que datas de publicação, tipo de escritores, e critérios de pertença ao cânone, são diferentes em cada comunidade literária (ou linguacultura).[6]

Esta ação COST está dividida em quadro vertentes (correspondentes a grupos de trabalho), enquadrando-se o trabalho que descrevemos aqui nas duas primeiras, denominadas respetivamente *Scholarly resources* e *Methods and tools*. O primeiro grupo de trabalho lida com a constituição e validação das coleções, enquanto o segundo tem como objetivo investigar e desenvolver métodos e ferramentas que possam ser usadas na criação (da anotação) das coleções, e para o seu processamento.

No seio do segundo grupo, um subgrupo dedicou-se à questão das entidades mencionadas, de que tratamos em detalhe no presente artigo.

2. Caracterização da coleção

Visto que a primeira coleção se encontra já razoavelmente terminada, enquanto a alargada ainda está no seu início e ainda nem todos os (novos) critérios para esta última foram definidos, neste artigo apenas descrevemos cabalmente a coleção ELTeC-por, fazendo apenas referência pontual a casos presentes na coleção ELTeC-por-ext.

Conforme indicado na documentação associada a esta coleção, já pública[7], como compiladores[8] deparámo-nos com uma grande escassez de obras publicamente disponíveis. Podemos mesmo dizer, sem perigo de errar, que praticamente só autores canónicos tinham sido digitalizados em português. As poucas exceções à regra vinham

[2] Mais informação sobre as variadas coleções encontra-se em https://www.distant-reading.net/eltec/.

[3] À data da escrita do presente artigo, apenas 5 coleções contêm 100 obras, das 10 publicadas por Odebrecht et al. (2020).

[4] Embora o período do COST seja de 1840 a 1920, não usámos o ano 1920 para que o quarto período também cobrisse exatamente vinte anos, como os outros. Mas isso não é consistente em todas as coleções ELTeC.

[5] Usamos o termo "grupo dedicado a uma literatura" e não país porque por exemplo as literaturas em inglês, francês, ou alemão abarcam vários países da Europa. Além disso, os grupos na ação não são automaticamente definíveis por país-literatura, visto que vários participantes na ação representam um país e dedicam-se a outra literatura, como é o caso de Christof Schöch, alemão especialista em literatura francesa, ou Jan Rybicki, polaco especialista em literatura inglesa, ou a primeira autora deste artigo, que representa a Noruega mas trabalha sobre o português.

[6] Sobre este assunto leia-se Herrmann et al. (2020).

[7] Em https://github.com/COST-ELTeC/ELTeC-por

[8] A responsabilidade da compilação é da primeira autora e dos representantes de Portugal no COST: Raquel Amaro, Isabel Araújo Branco e Paulo Silva Pereira.

do projeto Gutenberg, e/ou de projetos que disponibilizam versões modernas de livros antigos, como o LusoLivros.

Foi por isso necessário proceder ao trabalho de revisão do reconhecimento ótico de caracteres (ROC) feito por projetos estrangeiros e com ferramentas certamente muito pouco apropriadas à tarefa (ou seja, com sistemas antigos que não tinham sido treinados ou pensados para o português, e muito menos para a ortografia portuguesa do século XIX). Isso levou a que um livro digitalizado, por exemplo, pelo googlebooks levasse em média de 10 a 20 horas para limpar.

Tivemos também a ajuda da Biblioteca Nacional portuguesa para digitalizar (desta vez com um sistema mais adaptado ao português) algumas obras das quais só havia versão em papel. A revisão do reconhecimento ótico de caracteres (ROC, em inglês OCR) destas obras levou certamente menos tempo, mas foram relativamente poucas devido a o nosso projeto não ser evidentemente prioritário para essa instituição, à qual estamos de qualquer maneira muito agradecidos. Na tabela 1 apresentamos a proveniência das obras.

Origem	Quantidade
Gutenberg	32
Archive.org	30
Biblioteca Nacional	16
Luso-Livros	12
Projeto Adamastor	2
Googlebooks	3
Bibliotrónica	1
Wikimedia	1
NuPill	1
Arq. Mun. Sines	1
Hathitrust	1

Tabela 1: Donde vêm as obras (Santos (2020))

Vemos que grande parte delas foi necessário rever no âmbito desta ação. Apenas as obras disponibilizadas pelo Gutenberg e pelos sites Luso-Livros, Bibliotrónica Portuguesa e Projeto Adamastor (totalizando 47 obras) já se encontravam revistas, e a sua grafia, exceto no caso do projeto Gutenberg, atualizada. Quando fomos nós a rever, mantivemos a grafia original. Donde na coleção ELTeC-por temos apenas 15 casos de grafia modernizada. Trataremos aliás dessa questão mais adiante neste artigo (secção 3.4), em que analisamos obras com as várias grafias.

É no entanto também importante indicar que nem todas as obras foram digitalizadas a partir da sua primeira edição (e repare-se que, exceto no caso das pedidas à Biblioteca Nacional, não tivemos qualquer influência nessa escolha). Assim, em 23 casos a edição digitalizada não é a primeira. Mas a obra é datada e classificada no período COST com base na sua primeira edição, exceto no único caso em que não se conhece a data da primeira edição.[9]

Em relação ao sexo do autor, infelizmente apenas conseguimos 17 obras de escritoras, 12 delas pequenas. Outras obras inicialmente elencadas, como *Severina* de Guiomar Torresão, acabaram por se revelar pequenas demais. Parece assim possível afirmar que as escritoras femininas nesse período geralmente produziam obras de dimensão mais reduzida, muitas delas novelas ou contos.

Na figura 1 mostramos a distribuição das obras por década. Como não será de espantar, as obras antigas (das primeiras duas décadas) são as menos abundantes. A primeira década do século XX é a que tem maior número de livros, o que poderá ser explicado pela primeira guerra mundial na segunda década, mas não podemos evidentemente concluir nada desta pequena amostra. Para isso, teríamos de ter dados sobre todo o universo de publicação e não só sobre os romances e novelas escolhidos.

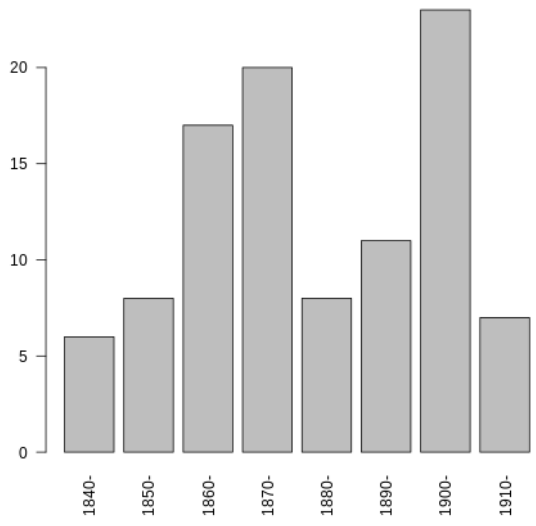


Figura 1: As obras do ELTeC-por por década

Uma questão que merece alguma atenção foi a quantidade de obras de cariz histórico que foram encontradas, e que levam à possibilidade de considerar que o género romance histórico foi muito cultivado na época a que nos reportamos. Com

[9] *O conde de Castel Melhor*, de D. João da Câmara, 2ª edição de 1903.

efeito, podemos identificar, nas 100 obras presentes, 33 romances históricos.

Outra questão interessante é a quantidade de romances ou novelas cujo título é apenas o nome de uma mulher (sete casos), ou de um homem (catorze casos).[10] Se considerarmos casos de títulos correspondendo à descrição de uma personagem masculina, como *O cristão novo* ou *Transviado*, ou feminina, como a *A divorciada* ou *A filha do Cabinda*, temos onze casos de títulos correspondendo a uma personagem masculina, e nove correspondendo a uma feminina. Finalmente, contando também os casos em que o título menciona ou inclui personagens femininas, temos dois casos em *O juramento da condessa Ester* e *Um conto português: episódio da guerra civil: a Maria da Fonte* e oito casos que incluem uma personagem masculina, como *A Confissão de Lúcio* ou *O crime do Padre Amaro*. Não há dúvida, com estes números, que o protagonismo masculino é predominante.

Quanto ao local de publicação da primeira edição ou da edição usada, a tabela 2 apresenta a distribuição das 119 obras da coleção ELTec-por-ext, em que não nos pareceu relevante distinguir entre as diferentes edições. O caso de local de publicação desconhecido reporta-se em geral às primeiras edições a que não tivemos acesso.

Local	Quantidade
Lisboa	67
Porto	27
Coimbra	7
Ponta Delgada	1
Guimarães	1
Rio de Janeiro	1
Funchal	1
desconhecido	14

Tabela 2: Local de publicação do conjunto das obras, na coleção alargada (118 obras)

O caso do Rio de Janeiro, correspondente à obra *A mulata*, é um exemplo de uma situação em que sabemos com certeza que o autor era português, devido à reedição muito mais tardia e à explicação de toda a história da obra no prefácio. O facto de ter sido reeditado em 1975 (80 anos depois da primeira publicação) torna a obra especialmente interessante, mas acabámos por decidir não a incorporar no ELTeC-por visto que não foi publicada em Portugal até dez anos da sua publicação inicial. Foi, contudo, uma das obras tratadas no presente artigo – o que aliás demonstra que a fixação dos critérios para a decisão final da coleção foi algo que levou muito tempo e deliberação.

A questão da publicação na Europa é algo que poderia dar a possibilidade de incluirmos já na coleção ELTeC-por livros de autores brasileiros, visto que é sabido que era comum estes publicarem em Portugal ou em França, provavelmente devido aos encargos da publicação no Brasil (Barbosa & Wyler, 2009). Por exemplo, Tristão de Alencar Araripe Júnior é um escritor brasileiro que publicou livros em Portugal com o pseudónimo de Cosme Velho, um deles inicialmente incluído no ELTeC-por mas depois transferido para o ELTeC-por-ext.

Existem também escritores portugueses que publicaram fora de Portugal, por exemplo na Itália, ou no Brasil. Pelas regras do COST, apenas os textos publicados na Europa devem fazer parte da coleção nuclear. Já a questão da primeira publicação em livro seguir-se à publicação em jornal ou periódico ser desaconselhada, esse requisito não foi seguido, por nos parecer impedir várias obras importantes de pertencerem à coleção, como o primeiro romance policial português, *O Mistério da Estrada de Sintra*, de Ramalho Ortigão e Eça de Queirós (aliás também o único em co-autoria na coleção portuguesa).

Relembrando as regras descritas na secção anterior, usámos nove autores com três obras. Os autores são, por ordem alfabética, Abel Botelho, Alberto Pimentel, Alexandre Herculano, Ana de Castro Osório, António Francisco Barata, Camilo Castelo Branco, Eça de Queirós, Júlio Dinis e Raul Brandão.

O equilíbrio entre os vários critérios não é sempre possível, o que significa que, se tentarmos apresentar todos na mesma visualização (veja-se a figura 2), é visível o viés para textos curtos (na bitola do ElTeC) e para autores masculinos, e mais recentes.

Em relação a não termos conseguido cumprir o critério de pelo menos 20 obras longas, e sem minimizar a grande dificuldade de obter textos longos, podemos relatar que dois dos textos inicialmente considerados nessa categoria tiveram de ser ou retirados por ainda não se encontrarem no domínio público, ou reclassificados como de tamanho médio, por a digitalização conter várias páginas repetidas.

[10] Isto vai contra a nossa impressão inicial de ser muito mais comum um nome de mulher como título, mas uma observação mais cuidada revela que os nomes de mulheres são quase sempre primeiros nomes, enquanto os de homem são sempre o nome completo, excetuando *Eurico, o presbítero* e nomes de personagens religiosas, que, como é sabido, são em português tratados pelo primeiro nome.

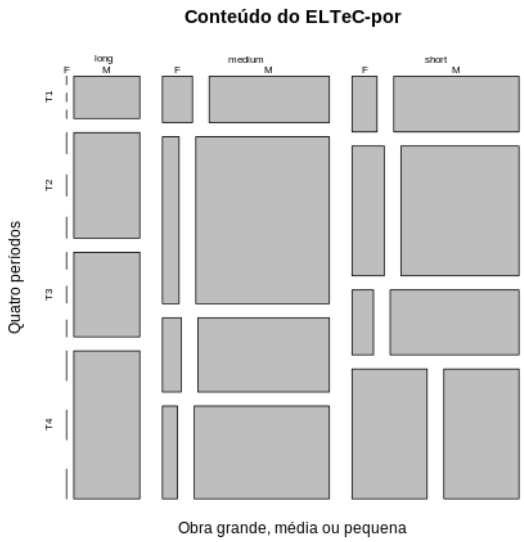


Figura 2: Visualizando o ELTeC-por de acordo com o período, o tamanho e o género do autor

3. Entidades mencionadas em textos literários

A coleção será tanto mais útil para a leitura distante da literatura lusófona (e mundial) quanto mais informação contiver sobre os próprios textos. Uma primeira e óbvia anotação é a das entidades mencionadas, em particular as pesssoas, os locais e as obras que são mencionadas num livro, assim como informação relacionada com uma caracterização sociológica da(s) realidade(s) descritas nos romances.

Após alguma discussão entre participantes de várias línguas, e a construção de uma coleção dourada inicial com base em excertos de romances em dez línguas diferentes (veja-se o trabalho de Stanković et al. (2019) e Frontini et al. (2020)), considerámos que dois campos geralmente não cobertos pela área do reconhecimento de entidades mencionadas (REM) seriam interessantes no estudo comparativo da literatura europeia: gentílicos e profissões.

Vale lembrar que a definição de entidade mencionada está geralmente ligada à caracterização morfossintática de nome próprio, que é diferente em línguas diferentes — e que é diferentemente sinalizada em termos gráficos por línguas diferentes. Assim, é conhecido que em inglês os gentílicos são marcados com maiúscula, e os nomes de profissões em alemão também apresentam maiúscula — visto que todos os substantivos o fazem. Por isso não seria possível para uma coleção multilingue usar uma abordagem baseada na língua (como fizemos para o HAREM (Santos et al., 2006), baseando-nos no português), e tivemos que concordar num conjunto de categorias (ELTeC, 2019) que todas as línguas tinham de identificar, para depois compararmos as literaturas. Nesse contexto, é importante chamar a atenção para o facto de que muito provavelmente não existirá nenhum sistema automático que faça essas decisões e só essas, visto que este é um conjunto de categorias de certa forma arbitrário.

Para o português, contudo, tanto profissões como nacionalidades já faziam parte do arsenal usado para a análise sintáctico-semântica do PALAVRAS (Bick, 2000), correspondentes às marcas `prof` e `Hnat`, veja-se Bick (2006, 2007), por isso bastou transformar estas marcações em categorias de entidades mencionadas, na saída do PALAVRAS-NER, e omitir (ou não considerar) os outros tipos de categorias semânticas identificadas por este sistema. Ou seja, concentrámo-nos em pessoas, lugares, organizações, obras, abstrações, acontecimentos, profissões e gentílicos.

Um dos objetivos do presente estudo é avaliar o desempenho desta anotação (neste contexto específico), e indagar, de forma preliminar, sobre o que ela nos pode trazer sobre a literatura anotada.

3.1. O PALAVRAS-NER

Incluído no analisador sintático PALAVRAS, o PALAVRAS-NER é primordialmente baseado em regras, como todos os outros níveis de anotação, e a informação lexical e gramatical para reconhecer entidades mencionadas está integrada nos léxicos e gramáticas gerais. A estrutura do sistema é um conjunto de módulos em cadeia ("pipeline"), cada um focando numa tarefa específica, mas usando a etiquetagem já construída pelos módulos anteriores, e preparando o terreno para os módulos subsequentes ao enriquecer o conjunto de categorias (desambiguado) que é entrada para as regras contextuais.

3.1.1. Identificação de entidades mencionadas e sua segmentação

O reconhecimento de entidades mencionadas pode ser dividido em duas tarefas: (a) identificação e (b) classificação das entidades. A identificação é geralmente (mas nem sempre) executada primeiro e inclui a atribuição da categoria gramatical "nome próprio" (PROP) ou a deteção de outras categorias gramaticais usadas como nomes, geralmente em maiúsculas, às quais é atribuída uma categoria gramatical secundária `<prop>`. Na tarefa de identificação

também está incluído o reconhecimento de uma cadeia de unidades como a ocorrência de uma entidade (por exemplo primeiros nomes e apelidos, nomes de instituições), e a identificação de abreviaturas como entidades mencionadas. Algumas expressões com várias palavras são tão frequentes que foram dicionarizadas, mas na maior parte dos casos a identificação das entidades multipalavra é feita dinamicamente, da seguinte forma: a anotação morfológica é feita primeiro, e regras gramaticais subsequentes atribuem classificação de partes de nome próprio (`@prop1` para a primeira parte e `@prop2` para as seguintes), o que tem óbvias vantagens em relação à alternativa de usar um pré-processador com reconhecimento de padrões:

1. Permite que a análise morfológica estabeleça o número e o género das entidades a partir dos seus constituintes e da sua estrutura
2. A gramática de REM pode mudar a composição de um nome próprio removendo, adicionando ou substituindo etiquetas de início ou continuação

Assim, o comprimento de um grupo de palavras reconhecido como uma unidade pode ser aumentado incrementalmente de uma forma sensível ao contexto e gramaticalmente motivada, por exemplo adicionando coordenações (as últimas duas palavras em *Doenças Infecciosas e Parasitárias*) ou sintagmas preposicionais (idem em *Câmara Municipal de Leiria*). Como as partes de entidades mencionadas são neste estádio visíveis como nomes ou outras categorias, até a valência sintática pode ser utilizada.

3.1.2. Classificação de entidades mencionadas

A tarefa de classificação atribui categorias semânticas às unidades simples ou complexas identificadas pela tarefa de identificação. Nos casos mais simples basta consultar almanaques ("gazeteers"). O PALAVRAS tem dicionários com cerca de 26.000 entradas, além de verificar nomes internacionais numa base de dados ainda maior para o inglês. Contudo, a categoria nome próprio não é uma classe fechada em nenhuma língua e é portanto preciso reconhecer e classificar os nomes próprios de outra maneira (em alguns textos, isto abrange a maioria dos nomes próprios). Usando uma estratégia "local" (interna à entidade), podemos usar padrões sobre cadeias de caracteres e pistas morfológicas. Em entidades com várias palavras, a classe semântica do núcleo do sintagma (ou de outro constituinte) geralmente fornece um pista vital, cf. *Sociedade/Ministério/Praça/Prêmio de...* ou *Sra....* Usando uma estratégia "global", podemos usar regras contextuais para desambiguar a categoria de uma entidade mencionada: por exemplo a preposição *em* num contexto adverbial pode ser usada para projetar LUGAR no seu dependente, assim como entidades mencionadas que são sujeitos de um verbo de fala ou de um verbo cognitivo são provavelmente PESSOA. Para nomes de pessoas em particular, uma lista de primeiros nomes internacionais, lista de apelidos comuns em português e partículas de ligação comuns (*da/do*, *von/van*, *bin*) permite o reconhecimento parcial de nomes de pessoas, que são depois propagados para a entidade mencionada completa.

3.1.3. O conjunto de categorias

O conjunto de categorias identificado pelo PALAVRAS-NER tem um grupo nuclear de seis classes comuns à maioria dos sistemas de REM: Pessoa (`<hum>`), Organização (`<org>`), Local (`<top>`), Acontecimento (`<occ>`), Obra (`<tit>`) e Marca (`<brand>`). Além disso, tem outras categorias menos comuns, adicionadas por serem funcionalmente ambíguas: Assim, cidades e países são classificados como `<Ltown>` e `<Lcountry>` ou com uma classificação subespecificada `<civ>` em vez de `<top>`, porque podem ser usados como lugares (viver em X) ou como pessoas ou organizações (X lançou/criou etc.). Da mesma forma, o PALAVRAS-NER usa `<media>` para designar algo que pode funcionar como título de uma obra ou como organização; e `<inst>` para cinemas, lojas ou embaixadas, por exemplo, para lidar com a ambiguidade lugar/organização. Tal é consistente com a filosofia do PALAVRAS de distinguir entre forma e função: a classificação imediata refere-se à forma semântica, deixando a função semântica ser atribuída numa camada superior que distinguirá entre Pessoa ou Lugar ao conceder os papéis semânticos de agente ou experienciador ao primeiro caso e de Localização ou Destino no segundo.

Outras categorias cobrem classes menores como prémio (`<prize>`), doença (`<disease>`), astro (`<astro>`) para estrelas e planetas e veículo (`<V>`) para carros e barcos. Como já mencionado anteriormente, os limites do que constitui uma entidade mencionada variam de língua para língua, e por isso, para permitir uma análise comparativa entre várias línguas, faz sentido marcar palavras que não são consideradas nomes próprios em português, tal como meses, nacionalidades e profissões, que o PALAVRAS pode "elevar" ao estatuto de entidade mencionada fazendo uso da marcação semântica dos nomes co-

```
<word id="1" form="José_das_Dornas" base="José_das_Dornas"
   postag="PROP" morf="M S" extra="*" head="2" deprel="SUBJ&gt;" ner="NER:hum"/>
<word id="2" form="era" base="ser" postag="V"
   morf="IMPF 3S IND VFIN" extra="fmc vK mv" head="0" deprel="FS-STA"/>
<word id="3" form="um" base="um" postag="DET" morf="M S" extra="arti"
   head="4" deprel="&gt;N"/>
<word id="4" form="lavrador" base="lavrador" postag="N" morf="M S" sem="Hprof"
   extra="cjt-head prop" head="2" deprel="&lt;SC" ner="NER:Hprof"/>
<word id="5" form="abastado" base="abastado" postag="ADJ" morf="M S" sem="jh"
   extra="np-close" head="4" deprel="N&lt;"/>
```

Figura 3: Análise da frase *José das Dornas era um lavrador abastado*, em que *José das Dornas* e *lavrador* estão marcados como entidades mencionadas, em formato MALT.

muns, marcação esta que inclui 200 categorias prototípicas numa ontologia ordenada hierarquicamente e com poucos níveis de profundidade.

Um terceiro tipo de entidade mencionada considerado pelo PALAVRAS-NER são expressões numéricas como data (`<dato>`) e ano (`<year>`). As moradas também incluem números de forma sistemática, e são marcadas com `<address>` em vez de simplesmente `<top>`. Obviamente, todos estes casos usam emparelhamento de padrões e a construção dinâmica de constituintes (adicionando unidades marcadas `@prop2`) em vez de simples consulta ao dicionário.

O sistema completo do PALAVRAS-NER compreende um módulo final de filtragem que transforma o formato interno do PALAVRAS em texto corrido, re-contraindo as contrações originais, juntando os enclíticos e colocando espaços nas expressões com várias palavras. Só as entidades mencionadas mantêm a sua classificação, usando etiquetas `<NER>` e `</NER>` em que a categoria gramatical (PROP, N ou NUM) é um atributo da entidade, veja-se o seguinte exemplo:

```
A mobilia, de um estofo azul e assetinado,
rivalisa em symetria com os mais encantados
jardins de <NER="PROP,civ,Ltown">Gra-
nada</NER>.
```

Também é possível obter toda a anotação em XML (formato MALT (Hall & Nilsson, 2005)) usando um novo campo para a classificação em entidades mencionadas, como mostra a figura 3.

3.1.4. A sequência de comandos do PALAVRAS-NER

Aqui mostramos como as tarefas associadas ao REM estão distribuídas na sequência de comandos do analisador. As duas gramáticas de REM contêm 1400 regras de CG, enquanto o resto do sistema (todos os níveis) compreende cerca de 7 mil regras. O léxico geral contém cerca de 70.000 lemas e o almanaque ("gazeteer") 26.000 nomes.

Primeiro pré-processador Atomização e reconhecimento de entidades mencionadas baseado em reconhecimento de padrões.

Segundo pré-processador Consulta a almanaques incluindo a ontologia internacional, e verificação de expressões com várias palavras

Analisador morfológico (incluindo apoio dicionarístico) atribuição de categoria gramatical, afixação e inflexão, e categorias semânticas para nomes próprios e subpartes de entidades mencionadas

Desambiguação morfológica (regras de CG) trata de nomes próprios ambíguos, por exemplo em posição inicial de frase

Primeira gramática de REM (regras de CG) verifica e corrige os limites das EM, e faz classificação local e contextual

Função sintática (regras de CG de "mapping" e de desambiguação) explora as etiquetas de EM para decisões de semântica

Segunda gramática de REM (regras de CG) explora as relações sintáticas e outra informação de outros módulos para adicionar, modificar e desambiguar as classes de REM

Gramática de estrutura sintática (regras de CG) adiciona marcadores para ligação de complementos a curta e longa distância, para a estrutura do sintagma verbal, e para coordenação

Gramática de dependências conjunto de regras para construir as árvores sintáticas completas

Pós-processador Transforma o resultado num formato textual ou no formato XML MALT

3.1.5. Tradução do PALAVRAS-NER para este projeto

Apresentamos aqui brevemente a "tradução" das categorias produzidas pelo PALAVRAS-NER para as classificações que usamos neste trabalho, na Tabela 3.

PALAVRAS-NER	Aqui
hum, groupind	Pessoa
civ, top, inst, site	Local
official	Profissao
org, admin	Organizacao
date, periodo	Data
tit, brand	Obra
Hnat	Demonimo
-	Outro

Tabela 3: Tradução das categorias do PALAVRAS-NER para a grelha do COST

3.2. A anotação e sua revisão

Para anotar quer do princípio quer como revisão usámos o sistema BRAT[11], que permite a anotação através da internete (ver figura 5), e que usa uma codificação em termos de posição no ficheiro, veja-se a Figura 4.

```
T1      Pessoa 153 158  Maias
T2      Pessoa 194 208  Eça de Queirós
T4      Pessoa 344 349  Maias
T5      Local 368 374   Lisboa
T6      Data 379 385    Outono
T7      Local 418 446   Rua de S. Francisco de Paula
T8      Local 492 509   Casa do Ramalhete
T9      Pessoa 886 904  senhora D. Maria I
T10     Local 1016 1025 Ramalhete
T11     Pessoa 1464 1483        monsenhor Buccarini
T12     Pessoa 1495 1508        Sua Santidade
T13     Profissao 1858 1867     Monsenhor
T14     Profissao 2205 2214     Monsenhor
T15     Obra 2231 2245  Vénus Citereia
T16     Pessoa 2369 2375        Vilaça
```

Figura 4: As entidades mencionadas codificadas pelo BRAT

Infelizmente a anotação através deste sistema mostrou-se muito lenta (chegou a 20 segundos entre a mudança de uma análise e a visualização da mesma!), e tivemos que dividir os ficheiros em vários fragmentos e apenas ter uma obra acessível do servidor de cada vez. Os tempos indicados neste artigo são portanto consequência desta situação.

Para transferir entre XML e o formato BRAT, usamos os vários conversores desenvolvidos e disponibilizados pela Universidade de Belgrado[12], veja-se a figura 6.

3.3. Textos processados

Os textos escolhidos para esta experiência – muitos deles acabando por não fazer parte da coleção final – foram os seguintes:

Tripeiros Antonio José Coelho Lousada. *Os tripeiros: Crónica do século XIV*, 1857, ortografia não atualizada. (POR0040)

Pupilas Júlio Dinis. *As pupilas do Senhor Reitor*, 1867, edição dos anos 1990, ortografia atualizada.

Viscondessa S. de Magalhães Lima. *A senhora viscondessa*, 1875, ortografia não atualizada. (POR0028)

Maias Eça de Queirós. *Os Maias*, 1888, edição dos anos 1990, ortografia atualizada.

Febo J.P. Oliveira Martins. *Febo Moniz*, 1867, edição dos anos 1980, ortografia atualizada. (POR0067)

Mulata Carlos Malheiro Dias. *A Mulata*, 1896, edição dos anos 1970, ortografia atualizada.

Ambições Ana de Castro e Almeida. *Ambições*, 1903, ortografia não atualizada. (POR0099)

Viagens Almeida Garrett. *Viagens na minha terra*, 1846, ortografia não atualizada. (POR0004)

As Pupilas e os Maias, canónicos, beneficiam de uma ortografia atualizada (anos 90), depois temos dois (Febo e Mulata) semi-canónicos[13] que têm a ortografia atualizada (anos 70 ou 80 do século XX), e mais três não canónicos (Tripeiros, Viscondessa e Ambições) com a ortografia da época em que foram publicados (1857, 1875 e 1903). Finalmente, juntámos um canónico (as Viagens) com a ortografia original, de 1846.

Por uma questão de simplicidade de escrita, referiremos no que se segue cada texto pelo nome curto da lista acima.

3.4. Primeiras tarefas

Para ter a noção de quanto tempo levaria a fazer a anotação humana de um texto usando o

[11] http://brat.nlplab.org/index.html

[12] http://nerbeyond.jerteh.rs/

[13] Obviamente que a canonicidade é algo que depende de uma definição complexa, que neste momento não é consensual. Mas estes dois textos, por terem sido repescados nos anos 70 ou 80 para não caírem no esquecimento, têm evidentemente mais direito a um estatuto de semi-canónicos do que os que nunca mais foram editados ou analisados.

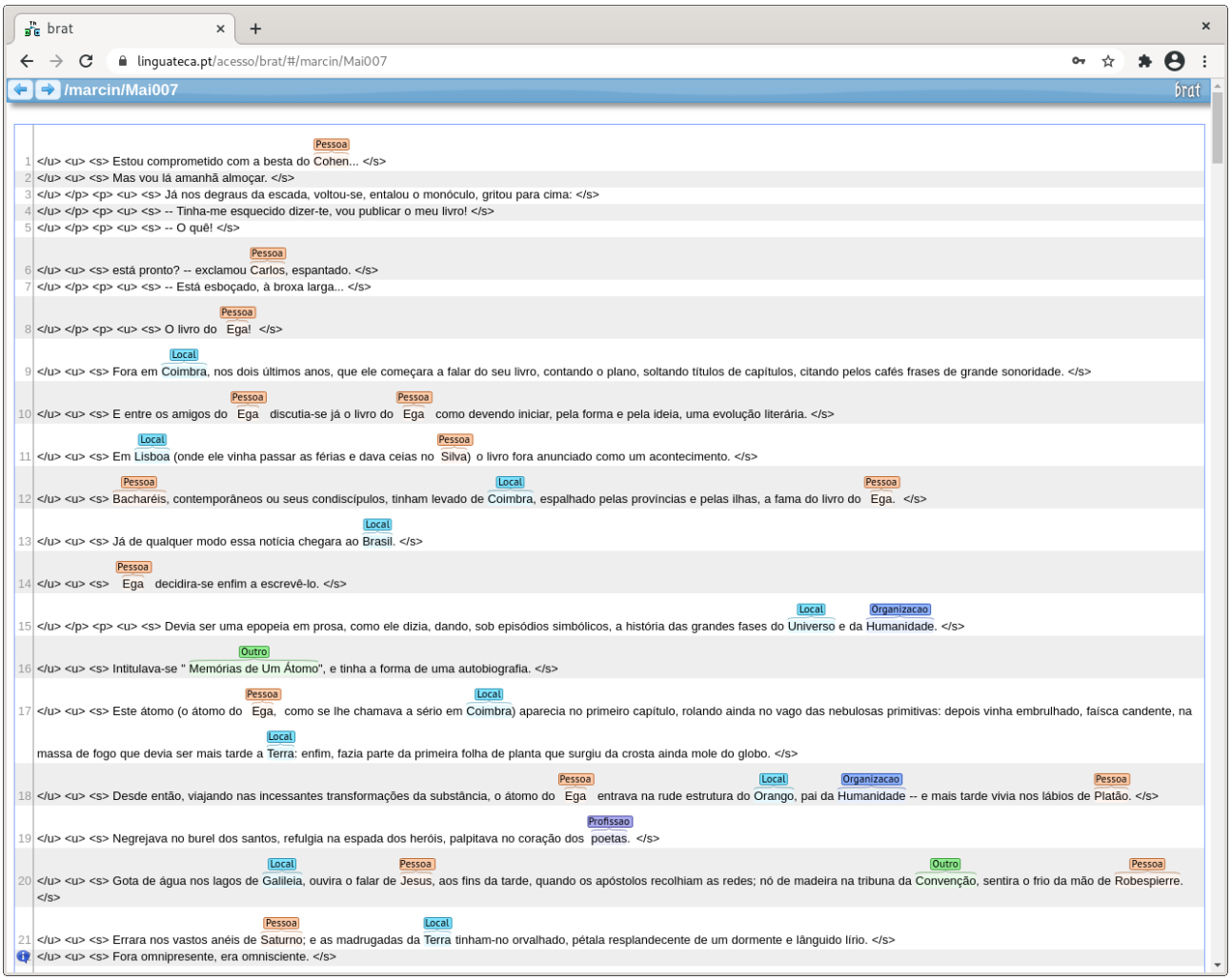


Figura 5: Sistema BRAT de revisão de anotação, com uma parte de *Os Maias*

BRAT, fizemos essa experiência. O texto escolhido, Tripeiros, tinha 38.173 palavras (contadas pelo programa `wc` do Linux) e foram encontradas 2.149 entidades mencionadas. A sua anotação levou 10 horas (mas refira-se a lentidão do sistema, visto que todo o ficheiro estava acessível ao mesmo tempo).

Depois fizemos a experiência de quanto levaria a anotar apenas adicionando profissões e gentílicos (que também chamaremos demónimos neste texto), a uma obra automaticamente anotada e já humanamente revista em relação a pessoas, lugares e obras (no âmbito da nossa anotação de personagens (Santos & Freitas, 2019)). O texto escolhido, Pupilas, tinha 96.448 palavras com 2400 entidades mencionadas. O resultado da revisão extra e da adição de gentílicos e profissões levou a 3453 entidades, em 11 horas.

Passámos depois à tarefa mais natural, e aquela que pretendemos utilizar na construção da coleção ElTeC-por: a revisão humana de textos automaticamente anotados pelo PALAVRAS-NER.

A esse respeito, consideramos dois tipos de textos: aqueles que têm uma grafia atualizada, e que se espera, portanto, que o sistema automático trate melhor, e aqueles com grafias antigas e provavelmente mais problemáticas.

Os dois primeiros textos anotados, os Maias e a Viscondessa, tinham os seguintes tamanhos em palavras: 218.665 e 26.305 e os seguintes números de entidades reconhecidas automaticamente: 11.862 (13.739) e 664 (846).[14] Só em si já uma diferença assombrosa. Como a diferença podia ser devida à canonicidade ou à grafia (Os Maias fazem parte do cânone e apresentavam uma grafia modernizada, a Viscondessa foi esquecida e tinha grafia antiga), tratámos em seguida dos outros textos, para ver se era a canonicidade a responsável pelo número maior de entidades, ou se isso seria uma característica da obra ou do autor.

[14] O primeiro número refere-se aos casos sem demónimos nem profissões, o segundo incluindo estes.

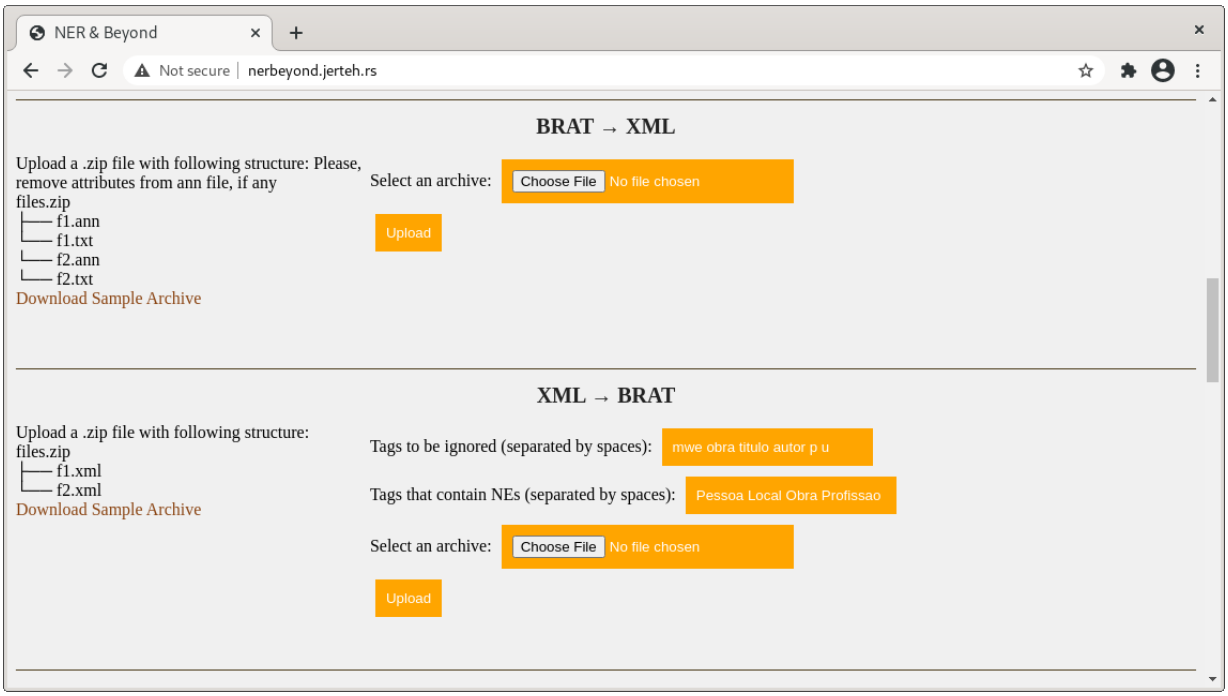


Figura 6: Parte do sistema para conversão acessível do servidor sérvio

A tabela 4 mostra os tamanhos de todas estas obras, assim como o número de entidades detetadas, antes e depois da revisão humana.

Vemos aqui já uma grande variabilidade, tanto nas dimensões dos textos como no número de entidades mencionadas que contêm. Nas oito obras escolhidas, de oito autores diferentes, a densidade de entidades mencionadas varia de 2,7 entidades por 100 palavras a 6,4.

Mas muito mais pode ser estudado com a informação que coligimos. Começamos por detalhar o tipo de entidades existentes e a sua forma (no sentido de serem compostas por uma ou várias palavras). Na tabela 5 apresentamos as quantidades para cada tipo de entidade.

Nota: visto que os números refletem um trabalho que foi modificando as caraterísticas para obter uma maior eficiência, é preciso explicar que o texto Viscondessa e um terço dos Maias foram anotados com a versão 13892 do PALAVRAS-NER, que ainda não detetava explicitamente profissões ou gentílicos enquanto que o resto do material já foi anotado com profissões e demónimos automaticamente, com a versão 13930.

Também podemos olhar para o tamanho das entidades: quantas vezes correspondem a uma palavra só, ou a muitas palavras, como a Tabela 6 mostra.

Antes de analisarmos este material, faz sentido descrever o próprio processo de revisão, dando voz ao revisor.

3.5. Relato de um trabalho intelectual

Apresentamos nesta secção alguns comentários que nos parece importante salientar, visto que em geral não é dada voz às pessoas que fazem a própria revisão ou anotação humana.

A primeira observação que é extremamente importante considerar é que o tempo e esforço de uma revisão não depende muito de o texto estar já anotado ou não, visto que o anotador/revisor tem sempre de fazer uma leitura atenta de toda a obra, e dirigir a sua atenção para o significado das palavras que tem de anotar. Por isso, em termos de tempo e esforço, não é significativa a diferença entre um texto virgem e um texto já anotado. O tempo que se usa a corrigir (muitas vezes duas ações) é quase o mesmo que se perde a anotar de novo (uma ação).[15]

Também e de um ponto de vista subjetivo, não foi encontrada diferença entre textos com grafia antiga ou moderna. Mas podemos indicar que, quando se tratava de um romance histórico, foi por vezes necessário identificar sentidos de palavras que já caíram em desuso, como é o caso de *correio*: "indivíduo que precede viajantes de distinção para lhes preparar aposentos, etc." (Dicionário Priberam).

[15]Para demonstrar isto não bastam os poucos números que temos, que além disso foram certamente influenciados pela ordem da revisão e pelas diversas alterações nos servidores para tornar o processo de revisão mais rápido, além de que o carácter de cada obra terá uma influência não despicienda, mas aqui ficam: Tripeiros, 15 horas; Pupilas, 11; Viscondessa, 5; Maias: 33.

Id	tamanho	EM antes	EM depois	densidade
Tripeiros	38.173	0 (2.080)	2.149	5,6
Viscondessa	26.305	664 (846)	727	2,7
Pupilas	96.448	2.400 (4.116)	3.453	3,5
Maias	218.665	13.739	14.094	6,4
Febo	69.683	3.747	3.559	5,1
Mulata	103.676	3.169	2.995	2,9
Ambições	78.933	2.370	2.523	3,2
Viagens	71.843	3.139	2.956	4,1

Tabela 4: Descrição das obras anotadas com entidades mencionadas. Densidade é definida como o 100*número de EM/número de palavras.

Id	Pessoa	Local	Profissão	Obra	Org.	Dem.	Abst.	Acont.
Tripeiros a.r.	1030	231	568	8	56	101	0	0
Tripeiros d.r.	1007	306	646	8	0	152	1	12
Pupilas a.r.	2737	154	1052	23	37	13	0	0
Pupilas d.r.	2498	73	806	32	3	18	13	10
Viscondessa a.r.	457	71	172	31	50	29	0	0
Viscondessa d.r	375	68	239	8	6	22	7	0
Maias a.r.	8065	2265	1463	361	591	254	0	0
Maias d.r.	8148	2553	1778	324	92	522	137	12
Febo a.r.	1977	540	703	64	73	210	0	0
Febo d.r.	1870	369	841	22	0	264	36	20
Mulata a.r.	1673	419	700	73	79	57	0	0
Mulata d.r.	2462	280	204	36	9	3	1	0
Ambições a.r.	1287	403	447	33	124	47	0	0
Ambições d.r.	1354	204	831	22	21	61	10	7
Viagens a.r.	1692	540	567	101	128	22	0	0
Viagens d.r.	1326	607	701	103	32	123	8	10

Tabela 5: Que tipos de entidades mencionadas: a.r. significa antes da revisão (ou seja, automaticamente analisado pelp PALAVRAS-NER), d.r. depois da revisão.

Id	1	2	3	4	5	6	7+
Tri	1481	463	153	45	5	1	1
Pup	2774	363	282	31	2	1	0
Vis	636	43	39	8	1	0	0
Mai	11919	1281	624	245	18	5	2
Feb	2707	540	274	35	3	0	0
Mul	2462	280	204	36	9	3	1
Amb	2020	348	129	18	11	2	0
Viag	2538	176	211	20	8	3	0
Tot	26537	3494	1916	438	57	18	4

Tabela 6: Qual o comprimento das entidades mencionadas, depois de revistas

Assim como *procurador do povo* (no sentido de representante nas Cortes) por oposição a *procurador* (no sentido de profissão judicial moderna).

Estas duas observações (sobre a dicotomia revisão/anotação e grafia antiga/moderna) são especialmente interessantes porque vão contra as nossas expectativas iniciais: ou seja, que a revisão seria consideravelmente mais rápida do que a anotação total; e que textos com grafia moderna seriam significativamente mais fáceis de rever ou anotar.

Do processo de anotação manual, concluímos que há certas classificações que é quase impossível conseguir automaticamente. Exemplos são casos de gentílicos que descrevem uma moda de barba (*suíças*), nomes de pessoas que se referem a uma época (*relógio Luís XV*), ou nomes de obras que incluem profissões e locais, como o *Barbeiro de Sevilha*.

Particularmente difíceis, confirmamos, são os casos de locais com nomes de pessoas: uma vivenda chamada *Vila Balzac* ou uma freguesia denominada *São Domingos*.

Finalmente, deveria ser possível corrigir erros sistemáticos em relação a personagens de uma obra, como é o caso de *Carlos da Maia*, sempre interpretado pelo PALAVRAS-NER como uma

pessoa (Carlos) de uma organização (Maia) no livro Maias, ou *Tomé*, sistematicamente classificado como Lugar em vez de Pessoa na obra Febo. (Embora isto seja possível de fazer, não foi contemplado no processo de revisão que relatamos aqui, e é possível que tenha contribuído negativamente para a percepção do(s) revisor(es).)

3.5.1. Erro ou divergência?

A nossa experiência de revisão dos textos anotados automaticamente leva-nos a considerar que houve várias razões para uma divergência entre a opinião humana e a do sistema automático, que convém identificar e também "resolver" de um modo diferente.

Em primeiro lugar, uma das características mais salientes da divergência é aquilo que se pode considerar uma profissão do ponto de vista de ser uma atividade constante ou essencial, e aquilo a que poderemos antes chamar papel, e que pode ser temporário ou acessório. Para o PALAVRAS-NER – e eventualmente para outros analistas humanos – não faria sentido separar os dois casos, mas no nosso caso nós apenas estávamos interessados em profissões como categorias socio-profissionais, ou como títulos nobiliárquicos ou eclesiásticos (*prior do Crato*, *duque de Palmela*) ou hierárquicos (*presidente*, *grão-mestre*).

Isso levou a que muitas "profissões" reconhecidas pelo PALAVRAS-NER fossem por nós rejeitadas, como *emissário, leitor, representante, profeta, portador, educador, orador* e muitos outros. Também decidimos não considerar classe ou estatuto social como profissão, e portanto não anotar *povo, clero, nobreza, escravo, proprietário, ama.*

Alguns dos "erros" do PALAVRAS-NER deveram-se, por outro lado, a conflitos entre variedades do português: *rapariga* terá sido sinónimo de prostituta, uma profissão, em português do Brasil, mas significa apenas menina em português de Portugal, *Veja* é uma revista brasileira da atualidade, mas isso é irrelevante para a literatura do século XIX; ou entre diferentes épocas: como é o caso do *procurador* e do *correio* já mencionados; *cavaleiro* antigamente era uma posição social ou um lugar no exército, mas muitas vezes também indicava apenas alguém que se deslocava a cavalo (por isso não considerámos profissão), e *Prior do Crato* é agora um largo de Lisboa, mas em romance históricos refere-se a uma pessoa específica, e não a um local. Finalmente, enquanto *cenógrafo* é nos nossos dias uma profissão conotada com o teatro, no século XIX significava aparentemente cenário.

É, portanto, muito importante salientar que o processo de revisão não significa apenas corrigir erros, mas também adaptar ou "personalizar", para uma dada tarefa, um sistema mais genérico.

Passaremos a uma avaliação mais objetiva dos resultados nas secções que se seguem, depois de apresentar em mais pormenor a informação que obtivemos com a revisão.

4. Resultados detalhados

As três tabelas apresentadas antes estão, naturalmente, longe de esgotar a informação que podemos obter com este processo.

Podemos, por exemplo, indagar quais as entidades mais comuns em cada obra (Tabela 7). Podemos ver imediatamente quais as personagens centrais nas obras (que são sempre as entidades mais frequentes), assim como algumas profissões e gentílicos que podem dar uma idea do ambiente, e por vezes locais. Assim, fala-se de alcaides, mouras, cavaleiros e besteiros nos Tripeiros, mas de viscondessas, padres e operários na Viscondessa. As Pupilas têm reitor, lavrador e padre, enquantos os Maias têm marquês e condessa. Febo tem castelhanos, rei e Cardeal, e a Mulata criada, médico, artista e poeta. O que não é previsível é a presença de Deus, que é das mais frequentes no Febo, na Viscondessa, na Mulata e nas Viagens, e provavelmente reflete mais o discurso direto do que um viés religioso.

Mas podemos fazer uma análise semelhante por categoria. Em vez de nos concentrarmos no conjunto de todas as entidades mencionadas, podemos apreciar cada tipo separadamente.

Assim, começando pelos lugares, vemos na Figura 7, na distribuição em números absolutos que os Maias mencionam quase 2500 lugares, enquanto as Pupilas apenas nomeiam cerca de 100.

Tendo em conta que as obras têm dimensões bastante diferentes, e comparando valores relativos, vemos que as Pupilas têm a menor densidade de lugares, enquanto os Maias já não diferem muito dos Tripeiros ou das Viagens.

Mas além de uma análise quantitativa podemos também investigar que lugares é que são mais mencionados (Tabela 8), e quantos lugares diferentes por obra (Tabela 9).

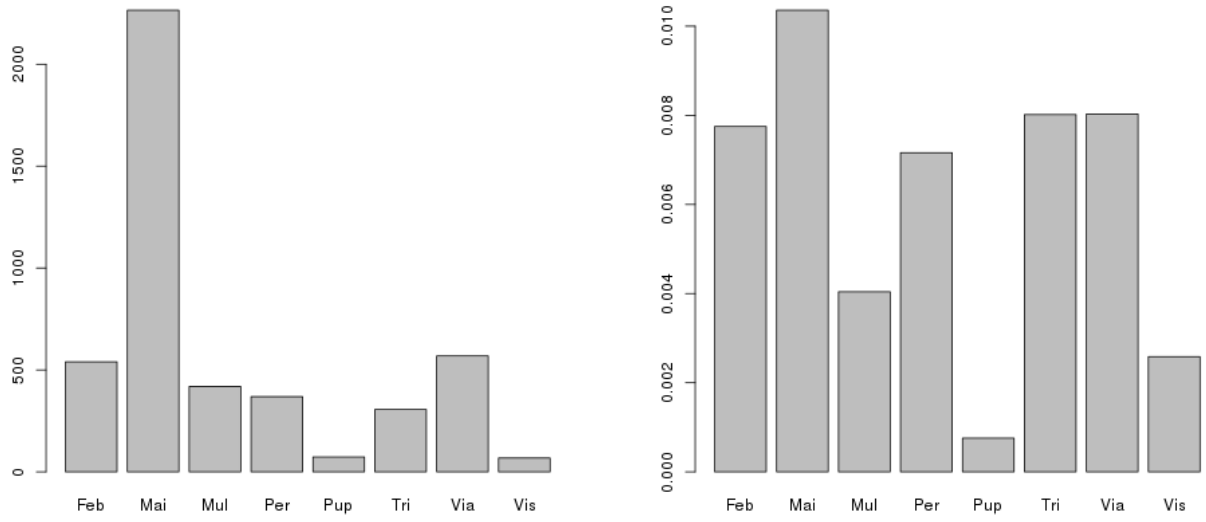
Vemos assim que aqueles textos que mais mencionam lugares, também são, talvez não surpreendentemente, aqueles que os repetem mais: Mais uma vez os Maias estão no topo. Exceção para o caso da Mulata que, talvez por ser passada no Brasil, menciona bastantes lugares mas pouco os repete.

Viscondessa		Pupilas		Tripeiros		Maias	
Viscondessa	76	Daniel	535	Fernando	76	Carlos	1714
Julio	73	Margarida	427	Irene	57	Ega	1069
Alfredo	70	Clara	334	João Bispo	54	Dâmaso	374
viscondessa	37	reitor	246	Garifa	51	Vilaça	283
Deus	26	Pedro	182	João	37	Maria	240
Cecilia	25	José das Dornas	138	Gonçalo Domingues	36	Craft	235
Felisbella	21	João Semana	105	alcaide	33	Afonso	202
Virginia	19	padre	84	moura	33	Alencar	198
padre	12	Guida	78	besteiro	28	Lisboa	195
operario	12	Joana	71	Mestre	28	Ramalhete	167
sr. Francisco Alves	11	Sr. Reitor	60	João Ramalho	26	Cruges	166
sr. Francisco	8	João da Esquina	58	Gaia	26	marquês	152
creado	8	Sr.ª Teresa	47	Rui Pereira	26	condessa	137
creada	8	Clarinha	41	conde	24	Maria Eduarda	116
Maria	7	lavrador	39	cavaleiros	23	Paris	111
Febo		**Mulata**		**Ambições**		**Viagens**	
Ana	203	Edmundo	515	João	161	Carlos	191
Deus	164	Honorina	193	Bella	100	Joanninha	170
Maria	128	Emílio	83	Candida	87	Deus	147
castelhano	111	Deus	79	Viscondessa	65	frade	137
Tomé	110	Julião	72	Visconde	55	Santarem	82
Febo	109	criada	40	Vihegas	56	fr. Diniz	76
D. Alonso	109	senhora Maria	31	Pillar	48	Georgina	50
Fernão	108	médico	26	Isabella	44	Lisboa	50
Cardeal	95	artista	25	Telles	41	Portugal	43
Marcos	90	criado	22	Lisboa	37	poeta	40
Febo Moniz	79	senhor Edmundo	19	abbade	36	Julia	36
rei	66	Emília	15	dr. Ramalho	35	Joanna	35
Margarida	58	turco	15	Maximiano	35	Cartaxo	29
Lisboa	55	poeta	15	Maria Helena	35	frades	26
castelhanos	55	Rio Grande	14	doutor	35	Laura	26

Tabela 7: As 15 entidades mais frequentes em cada obra

Obra	Lugares
Tripeiros	Gaia 26, Lisboa 20, Porto 19, Castela 18, Leça 15, Olival 10, S. Domingos 9, Monsaraz 8, Portugal 8, Miragaia 7, Douro 7
Pupilas	Porto 15, Lua 8, Terra 8, Sol 6, Coimbra 4
Viscondessa	Alcantara 7, Portugal 4, França 4, Lisbôa 3, theatro de D. Maria 3
Maias	Lisboa 195, Ramalhete 167, Paris 111, Sintra 106, Santa Olávia 103, Olivais 66, Portugal 61, Rua de S. Francisco 59
Febo	Lisboa 55, Portugal 35, Santarém 24, Espanha 19, Almeirim 14, Tejo 12
Mulata	Rio Grande 14, São Paulo 10, Rio 9, Juiz de Fora 9, Pascoal 8, Tijuca 8, Rua do Ouvidor 8, Botafogo 7, Roma 7, Largo do Paço 7
Ambições	Lisboa 37, Paris 24, Portugal 13, Inglaterra 8, Coimbra 7, casa do Maximiano 5
Viagens	Santarem 74, Lisboa 49, Portugal 40, Cartaxo 26, Tejo 15, Inglaterra 14, Azambuja 12, pinhal da Azambuja 10

Tabela 8: Quais os lugares mais mencionados por obra



(a) Números absolutos **(b)** Números relativos

Figura 7: Distribuição dos lugares

Obra	Diferentes	Locais	Repetição
Tripeiros	123	306	2,49
Pupilas	29	73	2,52
Viscondessa	43	68	1,58
Maias	500	2553	5,11
Febo	116	369	3,18
Mulata	228	411	1,80
Ambições	73	204	2,79
Viagens	229	607	2,65

Tabela 9: Lugares diferentes e repetição

Devemos também comentar o facto de que a Lua, o Sol e a Terra foram considerados lugares na anotação das Pupilas. Se os tivéssemos relegado para Outro (planeta, abstração), ainda teríamos menos lugares mencionados na referida obra — nas outras obras ou nem lua nem sol foram grafados com maiúsculas[16], ou seja como for não chegaram a ser dos mais mencionados: Terra é mencionada 25 vezes nos Maias e 1 vez na Mulata, mas não chegam para atingir as posições de topo na lista de lugares.

Se considerarmos agora o número de pessoas mencionadas, mais uma vez são os Maias a obra que usa mais nomes de pessoas (mais de 8000), ver Figura 8, mas a nível relativo isso não é tão pronunciado. Desta vez as Pupilas não se destacam pela negativa, sendo a Mulata e a Viscondessa as que apresentam menos designações de pessoas. Isto está muito provavelmente relacionado com o pequeno número de personagens das duas obras.

Olhando agora para as profissões, deparamonos com outra situação. Embora ainda existam mais profissões nos Maias em termos absolutos, visto que é a obra mais longa, em termos relativos são claramente os Tripeiros que ganham.

Em relação aos gentílicos, são os romances históricos, provavelmente porque descrevem batalhas e lutas contra outros povos ou países, que levam a palma no uso de demónimos.

A última categoria que nos parece fazer sentido comparar é a das Obras, até porque um trabalho anterior (Stanković et al., 2019) pareceu apontar para uma diferença significativa entre textos canónicos e não canónicos exatamente no que se referia à menção (ou não) de nomes de obras, que seriam muito mais frequentes nos casos em que as obras pertenciam ao cânone. Os resultados no nosso (pequeno) universo não confirmam essa hipótese.

Aqui são os Maias, as Viagens e a Mulata que têm significativamente mais menções a obras, o que se pode explicar pelo facto de os protagonistas destes romances serem pessoas de cultura, lidos e habituados a comentar obras de outros no seu dia a dia. Se olharmos para as obras men-

[16] De facto, nas Pupilas há muitas outras ocorrências de sol e de lua também em minúsculas; estes lugares são devidos a uma explicação astronómica dada por Daniel a uma criança na quinta, e poderiam ser retirados da classificação de lugares. Referimos contudo esta situação aqui para mostrar a quantidade de incertezas e de decisões que são arbitrárias, mesmo numa anotação humana.

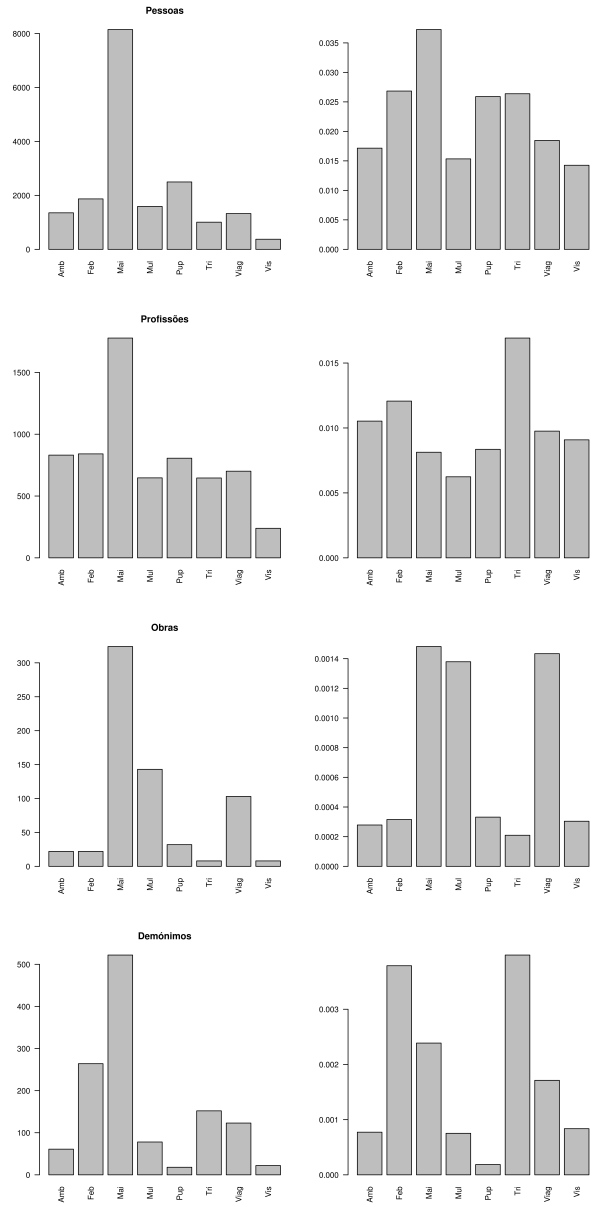


Figura 8: Distribuição em números absolutos, e relativos, das pessoas, das profissões, das obras e dos gentílicos

cionadas nestes livros, vemos que nos Maias são maioritariamente jornais (os dois mais frequentes são a *Corneta* e o *Figaro*), na Mulata romances contemporâneos (o *Barão de Lavos* e a *Dama das Camélias*), e nas Viagens clássicos (os *Lusíadas* e o *Fausto*). Isso espelha a vida quotidiana de Carlos da Maia, que gira à volta da política; de Edmundo, que é ou tenta ser escritor; e do próprio Almeida Garrett, poeta, romancista e dramaturgo.[17]

[17]Certamente o narrador de um romance não pode geralmente ser identificado com o autor, mas neste caso não parece haver muita diferença entre o "eu" das Viagens e o próprio Garrett.

5. Avaliação do desempenho do PALAVRAS-NER

O trabalho efetuado até agora, de revisão de uma anotação automática de nove romances, permitenos duas coisas diferentes:

1. avaliar o desempenho do PALAVRAS-NER para este tipo de tarefa, e consequentemente a necessidade de revisão, assim como estimar a taxa de erro expectável no caso de anotação sem revisão humana;
2. desenvolver parâmetros de estudo das obras literárias do período COST, a serem testados e avaliados em maior número de obras.

Tratamos aqui do primeiro aspeto, deixando o segundo para a próxima secção. Em relação especificamente ao desempenho do PALAVRAS-NER, depois de apresentar a precisão e a abrangência do mesmo em relação a alguns dos textos, tentaremos responder às seguintes perguntas:

- Há diferença significativa entre o desempenho em textos com grafia moderna e antiga?
- Há diferença significativa entre o desempenho em textos canónicos e não canónicos?
- Há diferença significativa entre o desempenho em textos correspondentes a romances históricos e textos modernos (ou melhor, contemporâneos da época em que foram escritos)?
- Há áreas específicas em que o PALAVRAS-NER se atrapalha? Por exemplo tipos de entidades?
- Existem regras fáceis de implementar no PALAVRAS-NER e que aumentem o seu desempenho para o resto dos textos?

A tabela 10 dá uma visão global do desempenho do PALAVRAS-NER em relação a estes textos, tomando a anotação manual como o correto. (Devido a uma diferença entre o formato que foi revisto nas Pupilas e ao formato obtido automaticamente, não podemos, infelizmente, avaliar este texto.)

As medidas de avaliação empregadas são as usuais neste tipo de tarefa de classificação, dadas pelas equações seguintes, em que *Corr* representa o número de classificações corretas produzidas pelo sistema, *Pres* indica o número de casos que deviam ser classificados, com base na classificação humana, e *Esp* indica o número de classificações erróneas (ou espúrias) produzidas pelo sistema.

$$Precisão = \frac{Corr}{Corr + Esp} \quad (1)$$

$$Abrangência = \frac{Corr}{Pres} \quad (2)$$

$$Excesso = \frac{Esp}{Corr + Esp} \quad (3)$$

As fórmulas são as mesmas quer se refiram à totalidade das entidades, ou apenas a uma categoria específica. Ao contrário do HAREM, não separamos a tarefa da identificação e da classificação, nem aceitamos classificações vagas: é apenas o desempenho da tarefa da classificação que medimos, assumindo que apenas uma classificação é pertinente.

	Corr	falta	Esp	P	A	E
Mai	10702	1866	1208	.797	.759	.090
Mul	2456	203	247	.808	.820	.081
Feb	2759	356	408	.764	.776	.113
Tri	1474	563	314	.783	.691	.167
Vis	502	176	206	.664	.691	.272
Amb	1626	535	377	.686	.643	.159
Via	2028	548	731	.646	.686	.233

Tabela 10: Desempenho do PALAVRAS-NER em sete textos revistos: P, A e E representam respetivamente a precisão, a abrangência e o excesso, nesta e em próximas tabelas. Por uma questão de formatação, nestas tabelas o ponto decimal é usado em vez da vírgula.

Imediatamente observamos que os textos de grafia antiga (os últimos quatro) têm pior desempenho, como seria de esperar, sobretudo na abrangência, mas não tão pronunciado como temíamos, mesmo entrando em conta com o facto de que a versão usada nos Tripeiros tinha inúmeros problemas de segmentação, que podem evidentemente impedir um programa de analisar convenientemente o texto.

Quanto aos textos de romances históricos (Tripeiros e Febo) poderem apresentar pior desempenho do que os "modernos", tal não é observado aqui. Vejamos mais em detalhe os casos das categorias novas, ou seja, os gentílicos e as profissões.

A análise dos demónimos mostra que esta categoria varia muito de obra para obra. Na Viscondessa, a obra com menos demónimos, o desempenho é péssimo, no Febo, que tem muitos, é o melhor. Impunha-se portanto uma análise mais aturada para identificar o porquê destas diferenças, que foi fácil de descobrir: enquanto que, na Viscondessa, dos 13 tipos de denónimos, oito tinham grafia antiga, como *francezes* ou *orientaes* ou *allemã*, impedindo o PALAVRAS-NER de os reconhecer, no Febo (com grafia moderna), a esmagadora maioria dos demónimos eram *castelhanos* e *portugueses*, fáceis portanto de identificar.

	Corr	falta	Esp	P	A	E
Mai	208	304	34	.860	.406	.140
Mul	46	26	11	.807	.639	.193
Feb	193	71	9	.955	.731	.045
Tri	57	94	2	.966	.377	.034
Vis	2	20	26	.071	.091	.929
Amb	22	39	25	.468	.361	.532
Via	12	105	9	.571	.103	.429

Tabela 11: Desempenho do PALAVRAS-NER em relação aos demónimos nos textos revistos

Na tabela 12 mostramos os gentílicos mais frequentes em cada obra, em que vemos, surpreendentemente, que os ingleses aparecem com muito peso em quatro das sete obras. Ter-se-á de esperar pela análise da coleção inteira para confirmar se têm de facto um papel importante na literatura portuguesa do período considerado, ou se foi apenas uma coincidência.

Em relação às profissões (e títulos de nobreza), existem muito mais casos de palavras referindo-as em todas as obras, como se vê na tabela 13.

Inesperadamente, vemos que o reconhecimento de profissões tem um desempenho relativamente bom, e que portanto não podemos concluir que as duas novas categorias constituam o calcanhar de Aquiles da anotação. De qualquer maneira, é interessante reparar que a abrangência dos Maias está no grupo dos piores, juntamente com a Viscondessa e as Ambições. Por isso pode fazer sentido olhar para as profissões destes textos, ou melhor, de todos, na tabela 12.

Mas, ao contrário do que vimos no caso dos gentílicos, o simples rol das profissões mais frequentes não parece explicar as diferenças encontradas. Tivemos que investigar quais os casos de profissões que faltavam nos Maias: e a grande maioria dos casos são de títulos nobiliárquicos, como *marquês* ou *duquesa* que aparecem sozinhos, e que não são marcados como "profissões" pelo PALAVRAS-NER. Este é um caso óbvio de diferença entre as duas filosofias de anotação, e que será fácil de resolver no futuro, quer adicionando estes títulos ao PALAVRAS-NER como `Hprof`, ou não as considerando na análise manual.

Em último lugar, na Figura 9 apresentamos o desempenho (precisão e abrangência e F1) para as três categorias mais frequentes: Locais, Pessoas e Profissões, por obra, assim como os valores globais na mesma forma de visualização.

Obra	Demónimos
Mai	inglesa 48, inglês 44, português 24, brasileira 22
Mul	turco 15, cabocla 4, português 4, índia 3
Feb	castelhano 111, castelhanos 55, portugueses 20, português 14
Tri	moura 33, galegos 13, castelhanos 12, portuenses 10
Vis	portuguez 3, havano 3, mouro 2, francez 2
Amb	inglesa 12, brazileiro 7, inglês 4, portuguez 3
Via	inglez 11, portuguez 11, ingleza 8, portugueza 5
	Profissões ou títulos
Mai	marquês 152, condessa 137, conde 87, criado 80
Mul	criada 40, médico 26, artista 25, criado 22
Feb	Cardeal 88, rei 66, Prior 34, frade 33
Tri	alcaide 33, besteiro 28, conde 24, cavaleiros 23
Vis	Viscondessa 76, viscondessa 37, operario 12,padre 12
Amb	Viscondessa 64, Visconde 45, abbade 36, doutor 35
Via	frade 137, poeta 40, frades 26, barão 23

Tabela 12: Demónimos e profissões mais frequentes em cada obra

	Corr	falta	Esp	P	A	E
Mai	1026	742	253	.802	.580	.198
Feb	649	150	155	.807	.812	.193
Mul	585	61	82	.877	.906	.123
Tri	393	252	101	.796	.609	.204
Vis	105	130	53	.665	.447	.335
Amb	375	392	71	.841	.489	.159
Via	471	225	92	.837	.677	.163

Tabela 13: Desempenho do PALAVRAS-NER em relação às profissões nos textos revistos

Concluindo, a tarefa que definimos como útil para os textos literários exige certa adaptação e fixação de critérios, visto que não existe nenhum sistema que tenha sido desenhado para esta tarefa e para este tipo de textos.

O PALAVRAS-NER tem um desempenho razoável, mas que pode ser melhorado se adicionarmos informação de dois tipos:

- léxico (ou melhor dizendo, ortografia) de períodos antigos no que se refere a profissões e lugares
- léxico de títulos que queiramos que sejam marcados tal como profissões

Se, além disso, através de um análise rápida do seu resultado, conseguirmos identificar casos de erro sistemático que podemos corrigir (através da adição ao seu dicionário), pensamos que pode ser usado para leitura distante, anotando (semi)automaticamente as 100 obras da coleção ELTeC-por.

6. Hipóteses sobre os textos literários

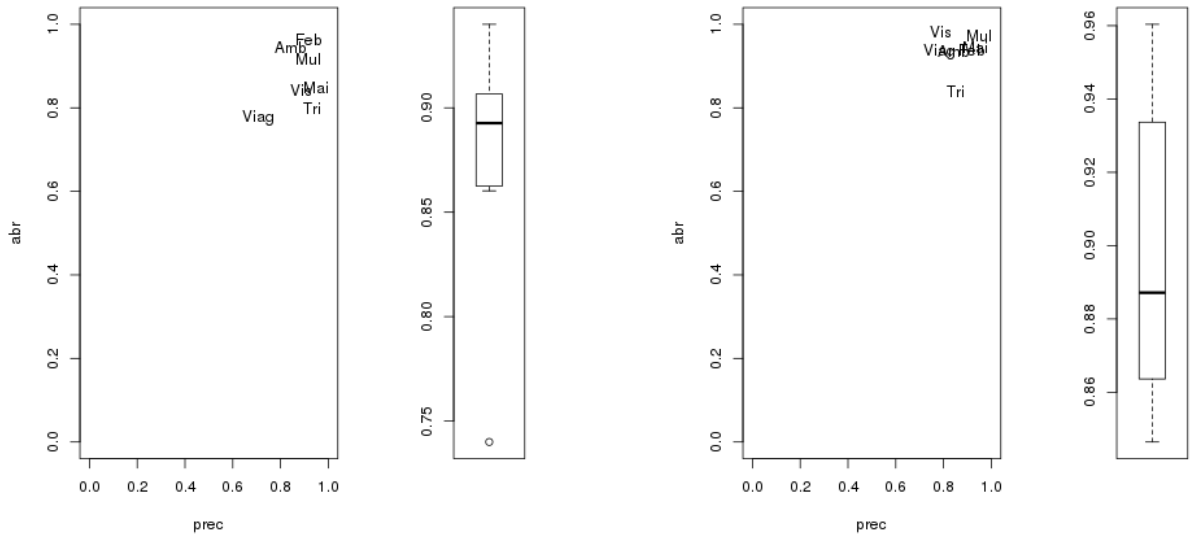
As seguintes perguntas parecem ser pertinentes:

- Existe alguma relação entre a população de entidades mencionadas e características exteriores de uma obra, como a sua canonicidade ou o (sub)género literário?
- idem em relação ao tipo de entidades, ao comprimento (em número de palavras) dessas entidades

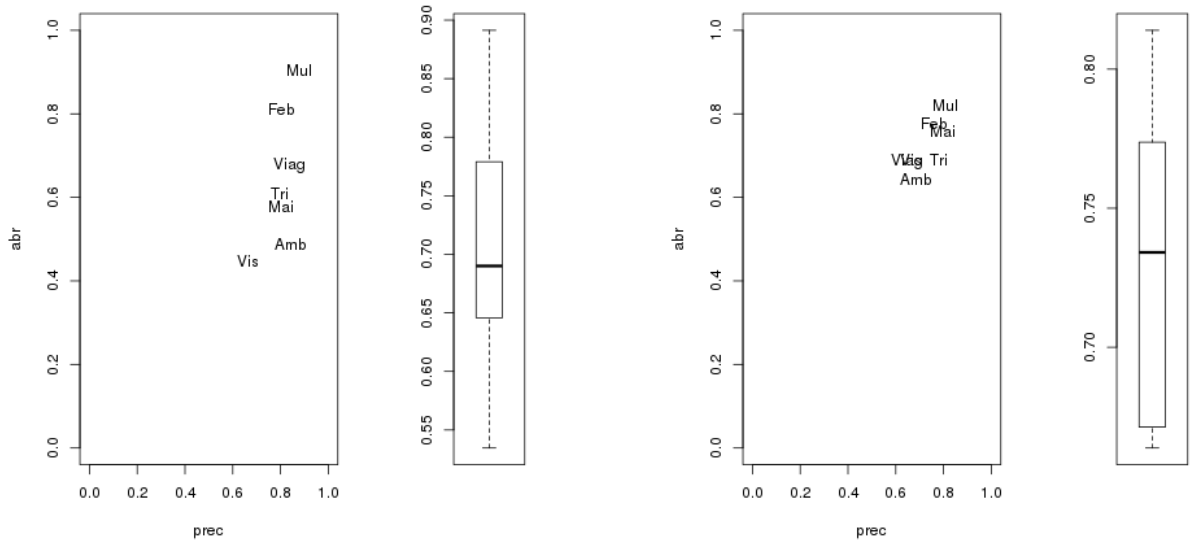
Não conseguimos responder positivamente a nenhuma destas perguntas. Em caso nenhum conseguimos encontrar uma propriedade que separasse as obras em duas classes obviamente distintas, quer entre canónicas e não canónicas, quer entre romances históricos ou contemporâneos.

De facto, as nove obras parecem ser demasiado díspares para podermos ver qualquer correlação. Uma (os Maias) é muito mais extensa do que as outras, o que pode levar a que tenha propriedades diferentes; outra (as Viagens) é escrita na primeira pessoa, o que também pode resultar em diferenças relevantes; finalmente apenas uma (Ambições) é escrita por uma mulher, tudo factores que podem ser fontes de variação adicional.

Por isso, concluímos que temos de classificar as cem obras de coleção para poder responder a estas e outras perguntas com um mínimo de confiança, embora não seja de rejeitar a hipótese de que a distribuição, tipo e quantidade das entidades mencionadas não tenha qualquer relação com a qualidade literária e ainda menos com a "canonização" de algumas obras em detrimento de outras.



(a) Locais **(b)** Pessoas



(c) Profissões **(d)** Todas as categorias

Figura 9: Avaliação do PALAVRAS-NER: Precisão, abrangência e F1

7. Comparação com outros trabalhos

Nos últimos tempos temos assistido a alguns estudos de REM sobre textos literários, que mencionamos aqui.

Bamman et al. (2019) anotaram cerca de 210 mil palavras de cem romances em inglês no LitBank data sheet, num período comparável ao do ELTeC-por (escolhendo as primeiras 20.000 palavras de cada obra) com as diretivas do ACE LDC (2005) (exceto a categoria armas), e utilizaram um sistema de REM treinado no material do ACE e treinado neste LitBank para mostrarem as diferenças de desempenho.

Como consequência dessa anotação, identificaram que os nomes de pessoas e de "facilities" (locais criados pelo Homem) eram muito mais referidos em texto literário, enquanto que organizações e entidades geopolíticas como países o eram muito menos.

Descobriram além disso um viés pronunciado contra o reconhecimento de personagens femininas: o sistema reconhecia sistematicamente melhor homens do que mulheres, mesmo retirando os casos óbvios de "Mr", "Miss" e "Mrs.".

No nosso caso, ao contrário deste trabalho, partimos de um conjunto de categorias que não tinha ainda sido testado para outros tipos de texto, por isso comparações diretas com os resultados do HAREM (Santos & Cardoso, 2007; Mota & Santos, 2008) não seriam justas. Poderíamos certamente comparar apenas nas categorias comuns, mas visto que o PALAVRAS-NER foi o vencedor do HAREM, não se esperaria um decrécimo de pontuação. Por outro lado, a diferença devida às novas categorias já foi medida com cuidado no presente artigo. Além disso, não estamos especialmente interessados em ver se o texto literário é diferente do conjunto de textos usados nos vários HAREM, visto que, ao contrário do ACE, esta avaliação foi pensada para vários géneros e não apenas para texto jornalístico.

Mas seria interessante como trabalho futuro identificar o género das pessoas mencionadas, e confirmar se existe ou não mais dificuldade de reconhecer um ou outro género, além de essa análise nos permitir ter uma ideia mais clara da população dos romances (e de certa forma corroborar, ou não, a observação feita na secção 2 baseando-nos apenas nos títulos).

Dekker et al. (2019) comparam quatro sistemas de REM para identificar personagens (melhor dizendo, redes de personagens) em romances antigos e modernos de ficção científica em inglês (vinte de cada): BookNLP, StanfordNER, Illinois NET e IXA-Pipe-NERC, e criam também uma coleção dourada com essas entidades e co-referências. Chamam a atenção para casos de nomes difíceis (por exemplo contendo apóstrofes como *D'Artagnan*), e para a possibilidade de ser útil manter diferentes designações da mesma personagem nas redes. Já Valaa et al. (2015) tinham usado alguns sistemas de REM para prever o número de personagens numa obra, mas o foco principal era compreender co-referências entre denominações de personagens.

de Does et al. (2017) estudam o REM da identificação de pessoas em romances em holandês, criando um sistema especificamente para esse domínio. Krug et al. (2018) criaram uma coleção dourada para texto em alemão, com 90 excertos de romances totalizando quase 400 mil palavras, em que todas as personagens estão marcadas e identificadas.

Convém, contudo, salientar que estes sistemas não marcam apenas nomes próprios quando se referem a personagens: pronomes e descrições nominais também são marcados. Por isso mais uma vez não podemos comparar facilmente os valores e os resultados com o estudo que apresentamos aqui. Relembramos aliás que, ao contrário do inglês, do holandês e do alemão, o português é uma língua de sujeito nulo, o que complica bastante esta contabilização, como apontado por Freitas et al. (2019).

Por outro lado, estes estudos quase sempre se dedicam apenas a pessoas — e, por vezes, mesmo apenas a personagens. Lee & Yeung (2012) são a exceção, obtendo tanto pessoas como lugares em redes com dois tipos de nós (locais e pessoas).

Frontini et al. (2020) fizeram uma primeira avaliação do desempenho de sistemas de REM para quatro línguas, com dois sistemas cada. Para o português, além do PALAVRAS-NER foi usado o spaCy[18]. Este trabalho é evidentemente o que mais se aproxima do que descrevemos aqui, visto que se baseia no mesmo tipo de anotação, contudo apenas trata pessoas e lugares. A coleção dourada para o português referida nesse artigo corresponde a excertos de 40 obras, as primeiras vinte (53.958 palavras) com grafia moderna e as segundas (55.429 palavras) com grafia original, ao todo correspondendo a 1999 pessoas e 607 lugares, enquanto o trabalho descrito no presente artigo refere-se a textos completos, totalizando 703.726 palavras, englobando 4.591 lugares e 19.040 pessoas.

8. Conclusões e trabalho futuro

Neste artigo apresentamos uma coleção de cem obras literárias portuguesas construída no âmbito de uma ação europeia de forma a comparar várias literaturas, coleção essa publicamente acessível[19].

Essa coleção será tanto mais útil quanto for passível de ser "lida" a distância, e um dos mecanismos para o fazer é identificar automaticamente que entidades lá são mencionadas.

Por isso, e como trabalho preliminar para esse objetivo, apresentamos um sistema de reconhecimento de entidades mencionadas, o PALAVRAS-NER, assim como oito textos revistos de acordo com diretivas especialmente pensadas para a literatura europeia. Essa anotação pode ser inspecionada[20], para que os leitores possam ajuizar as dificuldades do processo, assim como a (falta) de qualidade em termos dos materiais usados — desde erros de codificação do próprio texto até uma divisão automática de frases muito deficiente produzida pelo BRAT.

Neste artigo, além de descrevermos o processo de revisão, analisamos detalhadamente o desem-

[18] https://spacy.io/

[19] https://github.com/COST-ELTeC/ELTeC-por

[20] http://dinis2.linguateca.pt/brat-v1.3_Crunchy_Frog/#/Marcin/

penho do PALAVRAS-NER sobre cada obra e cada tipo de entidade, assim como tentamos explorar possibilidades de leitura distante das obras através da frequência, da distribuição das entidades, e dos seus casos mais frequentes.

Concluímos que o analisador era suficientemente bom para ser aplicado à coleção inteira, com eventuais adições aos seus léxicos para dar conta de ortografias anteriores.

Contamos, por isso, em breve anotar automaticamente toda a coleção com entidades mencionadas (além de análise morfossintática), de forma a caracterizá-la usando informação desta anotação.

Quanto à obtenção automática de redes de personagens, trabalho inicial usando uma identificação preliminar das várias maneiras de nomear uma mesma pessoa em algumas obras, foi já relatado por Santos & Freitas (2019). Será preciso desenvolver um processo de o realizar (semi)automaticamente, de forma a podermos comparar cem ou mais redes de forma análoga à de Dekker et al. (2019), mas isso terá de ficar para uma próxima ocasião.

Também pretendemos realizar a identificação dos lugares mencionados na literatura como um todo, como inicialmente discutido por Sanches et al. (2019), caracterizando os romances passados no campo ou na cidade, identificando os centros de poder e de referência, e eventualmente quais as conotações com eles relacionadas, na esteira de Cooper & Gregory (2011). Este é um trabalho — paralelo — em progresso, mas que já permitiu verificarmos que a maior parte das entidades mencionadas que aqui referimos como "lugares" não identificam lugares no mapa, real ou fictício, mas sim questões de identidade, abstractas, ou metonimicamente aplicadas.

Finalmente, como já referido na secção anterior, encontra-se também em curso uma investigação sobre se há diferenças sistemáticas na menção de pessoas (e personagens) de acordo com o seu género e, em caso positivo, quais.

Agradecimentos

Estamos gratos à ação COST "Distant reading for European literary history" sem a qual este trabalho não teria existido. A ação é financiada pela COST e pela União Europeia, através do programa Horizon 2020. Sem a equipa do COST português não teria sido possível criar a coleção. Nomeadamente: a Raquel Amaro, Paulo Silva Pereira e Isabel Araújo Branco, assim como às diversas revisoras contratadas. Além disso, foi muito apreciado o apoio dos líderes do grupo de trabalho 1 da acção COST, nomeadamente Lou Burnard e Carolin Odebrecht, e do responsável principal da ação, Christof Schöch.

Também agradecemos à Biblioteca Nacional de Portugal, na pessoa da Dra. Margarida Conceição Lopes, o ter prontamente digitalizado vinte obras a nosso pedido.

O terceiro autor agradece o apoio financeiro da Universidade de Oslo para a revisão da anotação dos textos aqui descrita.

Agradecemos à equipa de HPC da Universidade de Oslo, assim como ao grupo Sigma, o apoio computacional, a Tino Didrichsen o apoio em relação ao vislcg3, e a Ranka Stanković o apoio em relação ao sistema de conversão "NER and Beyond".

Finalmente, agradecemos a todos os elementos da Linguateca que leram e comentaram este artigo, e a Bruno Martins e a José João Dias de Almeida por um trabalho aturado de revisão com muitas sugestões pertinentes.

Referências

Bamman, David, Sejal Popat & Sheng Shen. 2019. An annotated dataset of literary entities. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 2138–2144. doi 10.18653/v1/N19-1220.

Barbosa, Heloisa Gonçalves & Lia Wyler. 2009. Brazilian tradition. Em Gabriela Saldanha & Mona Baker (eds.), *Routledge Encyclopedia of Translation Studies*, 326–332.

Bick, Eckhard. 2000. *The parsing system "Palavras": Automatic grammatical analysis of Portuguese in a constraint grammar framework*: Aarhus University. Tese de Doutoramento.

Bick, Eckhard. 2006. Functional aspects in Portuguese NER. Em Renata Vieira, Paulo Quaresma, Maria da Graça Volpes Nunes, Nuno J. Mamede, Cláudia Oliveira & Maria Carmelita Dias (eds.), *Computational Processing of the Portuguese Language (PROPOR)*, 80–89. doi 10.1007/11751984_9.

Bick, Eckhard. 2007. Automatic semantic role annotation for Portuguese. Em *Workshop em Tecnologia da Informação e da Linguagem Humana (TIL)*, 1715–1719.

Cooper, David & Ian N. Gregory. 2011. Mapping the English lake district: A literary GIS. *Transactions of the Institute of British Geographers* 36(1). 89–108. doi 10.1111/j.1475-5661.2010.00405.x.

Dekker, Niels, Tobias Kuhn & Mariekke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5(189). 10.7717/peerj-cs.189.

de Does, Jesse, Katrien Depuydt, Karina Van Dalen-Oskam & Maarten Marx. 2017. Namescape: named entity recognition from a literary perspective. Em *CLARIN in the Low Countries*, 361–370.

ELTeC. 2018. Sampling criteria for the ELTeC. Relatório técnico. COST Action CA16204 – WG1.

ELTeC. 2019. Annotation guidelines for named entities in ELTeC corpus. Relatório técnico. COST Action CA16204 – WG2.

Freitas, Cláudia, Elvis de Souza & Luísa Rocha. 2019. Quantificando e qualificando o sujeito oculto em português. Em *Jornada de descrição do Português*, s/pp.

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos & Ranka Stanković. 2020. Named entity recognition for distant reading in ELTeC. Em *CLARIN Annual Conference 2020*, 37–41.

Hall, Johan & Jens Nilsson. 2005. Converting dependency treebanks to MALT-XML. Relatório técnico. Computer Science, Växjö University.

Herrmann, J. Berenike, Carolin Odebrecht, Diana Santos & Pieter Francois. 2020. Towards modeling the european novel. introducing ELTeC for multilingual and pluricultural distant reading. 10.17613/tfbp-p625. Presentation at the Digital Humanities Conference.

Krug, Markus, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhausa, Frank Puppe & Fotis Jannidis. 2018. Description of a corpus of character references in german novels - DROC [Deutsches ROman Corpus]. Relatório técnico. Georg-August-Universität, Göttingen.

LDC. 2005. ACE (Automatic Content Extraction) English annotation guidelines for entities, version 5.6.1. Relatório técnico. Linguistic Data Consortium.

Lee, John & Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. Em *Pacific Asia Conference on Language, Information and Computation*, 209–218.

Mota, Cristina & Diana Santos (eds.). 2008. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo HAREM*. Linguateca.

Odebrecht, Carolin, Lou Burnard & Christof Schöch (eds.). 2020. *European Literary Text Collection (ELTeC)*. COST Action Distant Reading for European Literary History (CA16204). 10.5281/zenodo.4274954. Version 1.0.0, November 2020.

Sanches, Danielle, Daniel Alves & Diana Santos. 2019. O projeto BILLIG: aplicando sistemas de informação geográfica e linguística computacional ao estudo da literatura. Apresentação no Primeiro Encontro sobre leitura distante em português.

Santos, Diana. 2020. Coleção portuguesa de romances e novelas (ELTeC-por). 10.5281/zenodo.4271644. Versão v0.9.0, novembro de 2020.

Santos, Diana & Nuno Cardoso (eds.). 2007. *HAREM, a primeira avaliação conjunta de sistemas de reconhecimento de entidades mencionadas para português: documentação e actas do encontro*. Linguateca.

Santos, Diana & Cláudia Freitas. 2019. Estudando personagens na literatura lusófona. Em *Symposium in Information and Human Language Technology (STIL)*, 48–52.

Santos, Diana, Cláudia Freitas & João Marques Lopes. 2018. Ler e estudar a literatura lusófona como parte da literatura mundial: recursos para leitura distante em português. Em Suemi Higuchi & Cláudio José Silva Ribeiro (eds.), *I Congresso Internacional em Humanidades Digitais no Rio de Janeiro (HdRio2018)*, 375–383. CPDOC/FGV.

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. 2006. HAREM: an advanced NER evaluation contest for Portuguese. Em *Conference on Language Resources and Evaluation (LREC)*, 1986–1991.

Stanković, Ranka, Diana Santos, Francesca Frontinia, Tomaž Erjavec & Carmen Brando. 2019. Named entity recognition for distant reading in several european literatures. Apresentação na Digital Humanities Conference.

Valaa, Hardik, David Jurgens, Andrew Piper & Derek Ruths. 2015. Mr. Bennet, his coachman, and the archbishop walk into a bar but only one of them gets recognized: On the difficulty of detecting characters in literary texts. Em *Conference on Empirical Methods in Natural Language Processing*, 769–774. 10.18653/v1/D15-1088.