Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 15, Número 1 (2023)

ISSN: 1647-0818

Volume 15, Número 1 – 2023

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Prólogo

Desafio de Identificação de Personagens

O DIP – desafio de identificação de personagens – foi uma avaliação conjunta organizada pela Linguateca, NuPiLL - Universidade Federal de Santa Catarina, Universidade Estadual do Maranhão e Universidade de Oslo para promover a criação de programas que identificassem personagens em texto literário em português.

Embora existam vários trabalhos sobre o assunto para outras línguas, foi a primeira iniciativa a nível internacional para desenvolver programas de leitura distante para identificação de personagens, assim como foi a primeira avaliação conjunta em literatura no mundo lusófono.

O DIP começou a ser organizado em outubro de 2021, e termina agora com este volume da Linguamática em junho de 2023, em que documentamos a organização, a avaliação, os recursos criados e os problemas resolvidos ou identificados, assim como a participação.

Tivemos apenas um sistema participante que conseguiu apresentar resultados, o PALAVRAS-DIP, que é apresentado aqui. Além disso produziu dados sobre cerca de trezentas obras em português, públicos, que permitem que estudiosos de literatura, e de outras áreas, se debrucem sobre a literatura lusófona.

Esperemos pois que a publicação desta coletânea de artigos possa aumentar o interesse nesta área e levar quer ao desenvolvimento de mais sistemas quer a maior interesse pela leitura distante em português.

Diana Santos
Roberto Willrich
Marcia Langfeldt
Cristina Mota
Emanoel Pires
Rebeca Schumacher Fuão

Revisão

A comissão científica da **Linguamática** pode ser consultada na página Web da revista, em https://linguamatica.com/index.php/linguamatica/about/editorialTeam.

Para esta edição, colaboraram os seguintes investigadores:

- **Alexandre Rademaker,** IBM Research e FGV/EMAp
- **Amália Mendes,** CLUL, Universidade de Lisboa
- **Ana Alves,** CISUC, Instituto Politécnico de Coimbra
- **Cristina Mota,** Linguateca & INESC-ID
- **Diana Santos,** Linguateca & Universidade de Oslo
- **Emanoel Pires,** Universidade Federal do Piauí
- **Fernando Batista,** INESC-ID & Instituto Universitário de Lisboa
- **Henrique Lopes Cardoso,** LIACC & Universidade do Porto
- **Isabel Araújo Branco,** CHAM, NOVA FCSH
- **José Paulo Leal,** Universidade do Porto
- **Liliana da Silva Ferreira,** Fraunhofer Portugal AICOS
- **Luísa Coheur,** INESC-ID, Instituto Superior Técnico
- **Marcia Langfeldt,** DIP/Linguateca
- **Rebeca Schumacher Fuão,** DIP/Linguateca
- **Ricardo Rodrigues,** CISUC, Instituto Politécnico de Coimbra
- **Roberlei Alves Bertucci,** Universidade Tecnológica Federal do Paraná
- **Roberto Willrich,** Universidade Federal de Santa Catarina
- **Roney Santos,** Universidade Federal do Piaui
- **Sandra Aluísio,** ICMC, Universidade de São Paulo
- **Sérgio Nunes,** LIAAD, INESC TEC, Universidade do Porto

DIP: Desafio de Identificação de Personagens

Proposta recebida em Abril 2023 e aceite para publicação em Maio 2023.

DIP - Desafio de Identificação de Personagens: objectivo, organização, recursos e resultados

DIP - Character Identification Challenge: goal, setup, resources and results

Diana Santos
Linguateca & ILOS, UiO

Cristina Mota
INESC-ID & Linguateca

Emanoel Pires
UEMA/UFPI

Marcia Langfeldt

Rebeca Schumacher Fuão

Roberto Willrich
Universidade Federal de Santa Catarina

Resumo

Este artigo apresenta o Desafio de Identificação de Personagens (DIP) em profundidade. Documenta a sua motivação, as escolhas feitas, o desenrolar do processo de organização, a avaliação conjunta, e os resultados que podemos mostrar, assim como os recursos compilados e que são públicos. Relatamos o que aprendemos com a organização do DIP e o que aprendemos sobre a literatura em português. Por exemplo, nas obras do DIP, (1) o número de personagens femininas é muito inferior ao das personagens masculinas, (2) existem sempre algumas personagens referidas com nomes diferentes na mesma obra, (3) a profissão mais mencionada é a de padre, (4) há mais referência a pais do que a mães, e (5) os diminutivos são bastante frequentes.

Palavras chave

avaliação conjunta, literatura lusófona, identificação de personagens

Abstract

This paper presents in-depth DIP, the character identification challenge in Portuguese. It aims to fully document its motivation, the choices taken, the organization process, the evaluation contest proper, and the results achieved. It also presents the public resources created by DIP. We report on what we have learned with DIP's organization, and what we learned about lusophone literature. For example, in the works analysed by DIP (1) the number of feminine characters is way less than masculine characters, (2) every work has some character with more than a name, (3) the most frequent profession is priest, (4) the works refer more to fathers than to mothers, and (5) diminutives are pretty frequent as character names.

Keywords

evaluation contest, lusophone literature, character identification

1. Introdução

1.1. Motivação

A ideia de organizar uma avaliação conjunta no âmbito da leitura distante surgiu na senda da organização do Primeiro Encontro de Leitura Distante em Português, que teve lugar em Oslo em outubro-novembro de 2019, relatado em Santos et al. (2020a). Nessa altura, estabeleceu-se uma relação informal entre o Núcleo de Pesquisas em Informática, Literatura e Linguística (NUPILL)[1] e a Linguateca[2], veja-se Santos (2022b), em que uma das vertentes de colaboração futura seria a organização dessa mesma avaliação conjunta, dada a experiência que a Linguateca tinha na organização de avaliações conjuntas para o português.

De facto, existia já uma longa história de avaliações conjuntas organizadas pela Linguateca, iniciadas no EPAV (Encontro Preparatório de Avaliação Conjunta em Processamento Computacional do Português) em 2002[3], e documentadas em Santos (2022a). O NUPILL, por seu lado, é um dos centros dedicados à literatura computacional lusófona mais antigos no mundo, com quase 30 anos de atividades na área da literatura e da computação.

[1] https://nupill.ufsc.br/
[2] https://www.linguateca.pt/
[3] https://www.linguateca.pt/aval_conjunta/Faro2002/

DOI: 10.21814/lm.15.1.399

Além dos corpos literários da Linguateca (Vercial[4] e OBras (Santos et al., 2018)), a participação da primeira autora na ação COST "Distant reading for European literature history"[5] levou à criação de mais recursos para a leitura distante em português, nomeadamente o corpo NOBRE[6] e a coleção ELTeC-por Schöch et al. (2021). Concomitantemente, houve um aumento significativo do número de obras brasileiras digitalizadas por ocasião da bolsa de pós-doutorado de Emanoel Pires na Universidade de Oslo no período 2020–2022.

Foi, portanto, considerado possível iniciar um trabalho na leitura distante em português, incitando grupos a desenvolver sistemas que pudessem contribuir para esse objetivo. A escolha recaiu sobre a identificação de personagens, por nos parecer, de todas as possíveis demandas em leitura distante, a mais simples de concretizar e também de tornar visível a um público não especializado. Além disso, pensávamos poder contar com a existência de vários sistemas de reconhecimento de entidades mencionadas que talvez pudessem ser adaptados a esta nova tarefa.

1.2. A noção de personagem

Houve um rápido consenso em relação à escolha da identificação e caracterização de personagens como a tarefa mais atraente e exequível. Aliás, já tinham sido manualmente anotadas personagens no projeto AC/DC em relação a alguns livros, para criar redes de personagens (Santos & Freitas, 2019) e sabíamos que seria muito prático se pudéssemos executar essa tarefa automaticamente para muitas obras.

Contudo, a noção de personagem veio a mostrar-se mais complexa do que imaginávamos à partida, por várias razões:

- Em primeiro lugar, embora os estudiosos de literatura concordem geralmente na atribuição das etiquetas personagem principal, secundária e figurantes em relação a obras estudadas, não conhecemos uma metodologia geral, operacionalizável, e consensual que dada uma obra qualquer, ainda não estudada, produza essas decisões sem ruído. Por isso, decidimos identificar todas as personagens, e deixar para depois, se necessário, fazer esse tipo de distinção.
- Em segundo lugar, não parece sequer haver um consenso sobre a diferença entre pessoas mencionadas numa obra, e personagens dessa obra. Isso é tanto mais complicado no caso de, por exemplo, romances históricos que romanceiam a vida e as ações de figuras históricas, ou que as mencionam de passagem para situar a época em que o romance se passa. Alguns teóricos da literatura consideram mesmo que qualquer menção a uma pessoa dentro de uma obra a torna personagem dessa obra.

Nós seguimos a seguinte definição no DIP: as personagens em que estamos interessados são fictícias, ou são pessoas históricas que participam/fazem avançar o enredo numa dada obra. Referências a outras pessoas fictícias através de intertextualidade, ou a pessoas históricas que não participam no enredo, não devem ser consideradas como personagens da obra em questão.

Adotamos a distinção entre personagens e referências a outras pessoas (fictícias ou não) por duas razões. Uma, por nos parecer ser concetualmente distinto o estatuto dos dois tipos de entidades, e nos interessar sobretudo as entidades próprias de uma obra, por contraposição àquelas mencionadas por muitas.[7] E a segunda razão foi para diferenciar esta tarefa do simples reconhecimento de pessoas em texto, que é um subconjunto do problema do reconhecimento de entidades mencionadas.[8]

Depois de fixarmos o que era uma personagem para o DIP e de explicarmos o interesse de as analisar para os estudos literários em Langfeldt et al. (2021), tivemos de definir o conjunto de características que pretendíamos que os sistemas identificassem sobre essas personagens.

Enquanto a questão da correferência sempre esteve decidida, visto que sabíamos que uma personagem pode e costuma ter vários nomes e/ou formas pela qual é tratada — e esta questão dos diferentes nomes já tem sido objeto de investigação em vários outros trabalhos, veja-se Santos & Freitas (2019) ou Krug et al. (2018) —, foi preciso tomar uma decisão em relação a que outras características de uma personagem nós gostaríamos que os sistemas nos facultassem.

Uma característica razoavelmente consensual (no sentido de ser estudada por muitos inves-

[4] https://www.linguateca.pt/acesso/corpus.php?corpus=VERCIAL

[5] https://www.distant-reading.net/

[6] https://www.linguateca.pt/acesso/corpus.php?corpus=NOBRE

[7] Vimos mais tarde que uma distinção semelhante, entre entidades *plot-internal* e *plot-external*, também tinha sido feita no projeto Namescape (de Does et al., 2017).

[8] Estamos a afirmar que a tarefa de reconhecimento de personagens é mais complexa do que a do reconhecimento de pessoas, porque além de ter de separar os casos de lugares e organizações ainda precisa de distinguir, e rejeitar, pessoas que não sejam personagens.

tigadores de literatura) é o género da personagem. Mas convém esclarecer que o género de uma personagem literária e o género ou o sexo de uma pessoa no mundo real são conceitos distintos. O DIP limitou-se a identificar a representação textual do género, ou seja, o modo como a personagem é representada numa obra literária, mas não os traços de personalidade, os comportamentos, as ações e os estereótipos associados ao género. De facto, o género de uma personagem literária é uma construção determinada pela cultura e o período histórico no qual a obra está inserida, bem como pela intenção do autor. No DIP, o género de uma personagem pode ser masculino, feminino, ambos, ou desconhecido, mas veja-se mais a este respeito em Pires et al. (2023).

Outra tarefa que nos pareceu interessante para estudos históricos e do romance, sugerida por um estudo preliminar feito no âmbito da ação COST já citada, Santos et al. (2020b), foi determinar quais as profissões e ocupações mencionadas nos livros. Mas à medida que fomos operacionalizando a tarefa, como está também descrito em Pires et al. (2023), a definição foi-se revelando mais complicada. Para algumas personagens é o seu título nobiliárquico que as define, como *conde*, para outras é uma ocupação não remunerada, como *dona de casa* ou mesmo forçada, como *escravo*. E em português não há um termo que represente estas três formas de descrever a posição social ou ocupacional de uma pessoa, por isso optámos por usar a expressão "profissão, ocupação ou estatuto social", abreviada por POES.

Ao contrário do género, que em geral se mantém constante ao longo da obra — embora tenhamos trabalhado com um modelo em que pode mudar — a POES pode ser múltipla, e variada, ou seja, uma pessoa pode ser ao mesmo tempo médico e duque, ou passar de pastora de cabras para professora, e — o caso claramente mais frequente — transformar-se de estudante em profissional de uma dada área.

Finalmente, pensámos que seria interessante detetar relações familiares entre as personagens, talvez inspirados pelo trabalho feito sobre relações familiares no Dicionário Histórico-Biográfico Brasileiro (DHBB) (Higuchi et al., 2019). Mas também a operacionalização desta escolha teve várias consequências, descritas em Mota & Santos (2023). Foi sobretudo muito discutido que relações entre as personagens faria sentido tentar identificar. Durante a fase inicial, algumas pessoas criticaram que nos dedicássemos simplesmente a relações "oficiais"[9] e não a outras como amigo, amante, concubina, namorado ou admirador. A principal razão da não incorporação destas (importantes) relações foi a de que não eram estáticas: muitas vezes o próprio enredo é dedicado a um namoro, os amigos podem deixar de o ser, assim como os admiradores.

Essa foi, aliás, também a razão por que decidimos, nesta primeira edição pelo menos, não pedir os lugares onde a ação decorria, nem a estrutura temporal da obra, ambos temas que alguns interessados no DIP queriam tentar obter.

A escolha do que pedimos aos sistemas para tentar identificar automaticamente nos textos literários foi uma tentativa de equilíbrio entre algo suficientemente interessante mas não demasiado difícil. Não é garantido que o tenhamos conseguido, mas essa foi uma preocupação que nos norteou.[10]

2. Organização do DIP

Para organizar uma avaliação conjunta é preciso, além de escolher inicialmente uma tarefa, divulgá-la pelo maior número de pessoas e grupos, para que seja uma avaliação verdadeiramente conjunta. Em seguida, é preciso obter algum consenso sobre o calendário e sobre os recursos a serem desenvolvidos, assim como qual o formato exato da avaliação. E é preciso documentar todas as escolhas e ter um lugar na rede que os interessados possam consultar sempre que precisem.[11]

Começamos por apresentar o calendário final, na Tabela 1. Os principais eventos são descritos nas seções que se seguem.

2.1. Especificação da tarefa

A primeira atividade da organização do DIP foi especificar a tarefa a ser realizada no desafio. Foi definido que seriam tornados públicos 100 textos em formato de texto, na codificação UTF-8, e 100 textos em pdf. Uma vez disponibilizadas as obras, os sistemas participantes do DIP teriam 48 horas, o período do desafio propriamente dito, para produzir toda a informação sobre cada obra.

Como parte da definição da tarefa, a organização definiu como as obras iriam ser dispo-

[9] No sentido de legais, verificáveis num cartório.

[10] Contudo, não podemos deixar de concordar com os comentários de Luísa Coheur, Alexandre Rademaker e Roberlei Alves Bertucci de que a justificação de que as relações familiares não de sangue não é fixa, e que aceitamos várias profissões — porque não aceitar lugares também? Ou seja, a justificação das nossas escolhas não é muito coerente.

[11] DIP criámos pois `https://www.linguateca.pt/DIP`.

Data	Atividade
10/2021	Início da organização
05/11/2021	Anúncio público
29/11/2021	Encontro virtual
29/11/2021 a 15/03/2022	Ensaio: participantes e interessados anotam dois novos textos
16/03/2022	Encontro virtual sobre o ensaio
15-17/09/2022	Desafio
01/10/2022	Resultados publicados
21/11/2022	Encontro do DIP

Tabela 1: Calendário do DIP

nibilizadas e qual a sintaxe de representação dos dados extraídos das obras. As obras foram distribuídas com o nome obra$_i$.txt ou obra$_j$.pdf, onde i e j são números inteiros, variando respetivamente de 0 a 99 e de 100 a 199. O resultado da análise da obra deveria ser representado na forma de dois ficheiros CSV: personagens.csv e relacoes.csv. O primeiro deveria indicar as personagens, uma por linha, seguindo a seguinte sintaxe: $\{i, k, correferencias, genero, POES\}$, onde i é o identificador da obra, k é o identificador da personagem na obra, $correferencias$ é a relação de menções no texto referenciando a personagem, $genero$ indica seu género (M, F ou A para ambos), e $POES$ indica o conjunto de profissão/ocupação/estatuto social da personagem. Dois exemplos deste formato encontram-se na Tabela 2.

O arquivo relacoes.csv deve incluir todas as relações familiares entre personagens utilizando a seguinte sintaxe: $\{i, s, relacao, o\}$, onde i é o identificador da obra, e s e o identificam as personagens (mesmos identificadores em personagens.csv) que têm a relação $relacao$. Exemplos deste formato encontram-se na Tabela 3.

Os participantes deveriam usar o sistema *EasyChair*[12] para enviar o resultado dos seus sistemas, num ficheiro comprimido `zip`. Para evitar problemas, deveriam testar todo este processo (formato e EasyChair) no ensaio.

2.2. Ensaio

Para familiarizar os participantes com o que lhes era pedido, e também com as várias escolhas que teriam de fazer, pedimos a todos os interessados para anotar manualmente, após leitura próxima, mais dois textos, e discutirmos os resultados em conjunto, o que deu origem a várias precisões e melhores diretivas, assim como a mais dois ficheiros de exemplo.

[12] https://easychair.org/

Como já mencionado, esses dois ficheiros também teriam de ser enviados pelo EasyChair para testar o envio.

O processo do ensaio e a discussão no encontro (remoto) foram também muito importantes para fixar a tarefa de construção da coleção dourada, que foi a atividade mais trabalhosa que a organização levou a cabo: ler mais trinta e oito obras de fio a pavio e produzir toda a informação exigida pelo DIP.

2.3. Avaliação

Outra incumbência da organização foi sugerir medidas de avaliação para a tarefa do DIP, e implementar e testar os programas que as calculassem, antes da própria avaliação conjunta, para todos os participantes saberem como iam ser avaliados.

O resultado deste trabalho, e as medidas a que chegámos, estão descritos pormenorizadamente em Willrich & Santos (2023).

2.4. A compilação da coleção do DIP

Finalmente, outra tarefa extremamente importante foi fixar quais as obras que seriam distribuídas no DIP (e, dentre estas, quais as que fariam parte da coleção dourada). Isso será o tema da próxima secção.

3. Recursos criados

Talvez o mais importante resultado de uma avaliação conjunta sejam os recursos criados no seu âmbito, além da especificação da tarefa e da medição do seu sucesso.

Além da descrição objetiva dos dados que pusemos à disposição da comunidade, parece-nos importante documentar a sua construção e as várias decisões que tivemos de tomar.

i,k,correferencias,genero,POES
021,0,Margarida|Guida|Guida dos Meadas|Margaridinha,F,cabreira|professora
021,1,Clara|Clarinha|Clarita|Clarita dos Meadas,F,
021,2,Daniel|Sr. Daniel|Danielzinho|Daniel do Dornas|Danielzinho do Dornas,M,estudante|médico
021,3,Francisca|Chica|Chica da Esquina,F,
021,4,Joana|Sra. Joana,F,criada

Tabela 2: Exemplo de como a informação deveria estar codificada no ficheiro personagens.csv

i,s,relacao,o
021,0,irmã,1
021,2,irmão,6
021,5,pai,2
021,9,marido,10
021,9,pai,3

Tabela 3: Exemplo de como a informação deveria estar codificada no ficheiro relacoes.csv

3.1. A que textos/obras podíamos recorrer

Em primeiro lugar, e como já descrito em várias outras ocasiões (Schöch et al., 2021), não existe infelizmente um manancial de obras em texto em português que possa ser imediatamente usado para o seu processamento, ao contrário de outras línguas. Essa foi, aliás, uma das razões que nos levou a pensar na vertente "tratamento do pdf" no DIP, visto que existem, ou imaginamos que existam, muito mais textos simplesmente digitalizados em PDF como imagem, como é o caso dos acessíveis através do Google Books[13] ou do Internet Archive.[14]

Para sermos mais explícitos: o reconhecimento ótico de carateres associado à maior parte das obras em domínio público digitalizadas em língua portuguesa, por exemplo as existentes no Internet Archive, produzidas por iniciativas de digitalização nos Estados Unidos, é de qualidade tão má que se pode considerar que a simples digitação manual do livro levaria o mesmo tempo que a revisão do que foi reconhecido automaticamente.[15] Digitalizações mais modernas, e/ou feitas por instituições com conhecimento (e ferramentas) mais adequadas para a língua portuguesa, como as das bibliotecas nacionais de Portugal e do Brasil, por exemplo, produzem objetos digitais muito mais fiáveis, mas mesmo assim (ainda) não perfeitos. Seja como for, se o objetivo último é fazer leitura distante sobre milhões de obras, não podemos esperar que estas sejam revistas e, por isso, era importante levar sistemas a trabalhar com PDF.

3.2. Critérios para a escolha da coleção do DIP

Para escolher os 100 textos que fariam parte da coleção em formato de texto, socorremo-nos da lista de romances e novelas acessíveis na Literateca (Santos, 2019), das quais escolhemos uma obra de cada autor. Havia exatamente 50 autores brasileiros, e um pouco mais portugueses, mas não fizemos grandes reflexões sobre o assunto, exceto que esgotámos as autoras (visto que sabíamos que havia poucas).

A maior parte dos 100 textos em formato PDF foram selecionados dentre os 460 arquivos com romances ou novelas em domínio público disponibilizados na Biblioteca Digital de Literaturas em Língua Portuguesa (BDLP[16]) do NUPILL. Neste caso havia muito mais obras brasileiras em PDF do que portuguesas — provavelmente porque o trabalho da BDLP, em andamento, é feito no Brasil —, e tivemos dificuldade em arranjar 50 obras em PDF de autores portugueses que ainda não constassem da coleção de texto. Para resolver esse problema, adicionámos obras diferentes de autores que já estavam na coleção de texto.

No caso dos autores brasileiros, como havia muito mais do que cinquenta novos autores (em relação aos incluídos na coleção de texto), além de escolher todas as autoras, utilizámos como critério o tentar maximizar a variação do estilo e da data, embora não conhecêssemos a maioria das obras em questão.

A lista de obras contidas na coleção DIP, em txt e em pdf, está no apêndice A. Na Figura 1 está a distribuição temporal destas obras. De referir que incluímos as obras usadas como exemplo e estudadas no ensaio.

[13] https://books.google.com/

[14] https://archive.org/

[15] Não temos uma demonstração desta afirmação, que corresponde simplesmente à nossa experiência com alguns textos.

[16] https://literaturabrasileira.ufsc.br/

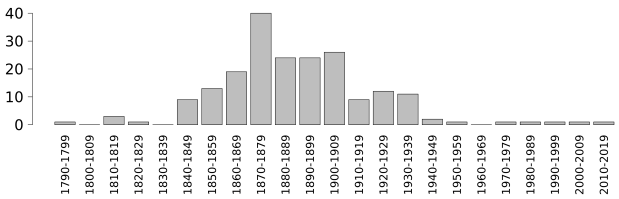

Figura 1: A distribuição das obras da coleção do DIP por década

Podemos constatar que existem 25 obras escritas por autoras e 178 por autores. O número de autores diferentes portugueses é de 88[17] e o número de autores diferentes brasileiros é de 96.[18]

3.3. Critérios para a escolha da coleção dourada

A coleção dourada é o subconjunto da coleção DIP para a qual compilámos os valores que pretendíamos que os sistemas obtivessem, 21 obras em texto e 17 obras em pdf, das 100 de cada.

A escolha destas obras não seguiu critérios pensados do início. Pelo contrário, no caso do texto aconteceu ao mesmo tempo que escolhemos as obras para a coleção do DIP e, no caso do pdf, foi eminentemente escolhida por questões práticas, nomeadamente já termos a obra em formato PDF em nosso poder, e escolhemos a maioria das outras obras em PDF mais tarde.

O único critério que tentámos seguir, para maximizar a variação da coleção, foi o de não termos mais do que uma obra por autor – critério esse que foi infelizmente desobedecido, por lapso, no caso de Carlos Pinto de Almeida, com duas obras na coleção dourada relativa aos pdf.[19]

Embora possa parecer que demos pouca importância à escolha da coleção dourada, é preciso salientar duas coisas: não fazíamos ideia da distribuição de personagens na literatura em geral e, portanto, não tínhamos muitas características sobre as quais diversificar; e esperávamos que, mais tarde ou mais cedo, obteríamos, a partir do resultado dos sistemas, informação para todas as 100 obras, por isso não era assim muito importante por quais começar.[20]

[17] Os autores portugueses com duas obras são Alberto Pimentel, Alice Pestana, Ana de Castro Osório, Ana Plácido, António Francisco Barata, Arnaldo Gama, Camilo Castelo Branco, Carlos Malheiro Dias, Carlos Pinto de Almeida, Eça de Queirós, Francisco Gomes de Amorim e Virgínia de Castro e Almeida

[18] Os autores brasileiros com duas obras são Aluísio Azevedo, Bruno Seabra, José da Rocha Leão e Júlio Ribeiro

[19] As obras são *A filha do emir* e *Os homens da cruz vermelha*.

[20] De facto, neste momento — junho de 2023 — encontra-se em curso a leitura próxima de mais obras da coleção de texto, alargando assim os recursos produzidos pelo DIP.

3.4. Decisões na anotação da coleção dourada

Muito mais trabalho e discussão (no seio da organização, e com os participantes no ensaio) levou a própria criação do conteúdo da coleção dourada, indicando que haveria muitas questões mais finas sobre as quais teríamos de chegar a um consenso de forma a poder documentar o que os sistemas deveriam fazer. Tal como no HAREM, em que tivemos de escrever páginas e páginas de diretivas (veja-se Cardoso & Santos (2007)), na questão das personagens houve muitas decisões que precisámos de tomar.

Embora não consigamos trazer aqui todas, a seleção que apresentamos dará uma ideia de que a operacionalização de uma tarefa computacional (ou, seja como for, exaustiva) requer a consideração de muitas questões, muitas delas não necessariamente intuitivas.

3.4.1. Narrador como personagem

Uma das primeiras coisas sobre as quais nos tivemos de pronunciar foi sobre personagens sem nome. Embora importantes para a análise literária, não nos pareceu possível arranjar uma forma natural de as incorporar no DIP.

Esse é o caso de narradores autodiegéticos ou homodiegéticos, na primeira pessoa, que não sejam nunca tratados pelo nome. Nesse caso, não aparecem nas listas das personagens do DIP.

3.4.2. Formas de indicar um parentesco

Há muitíssimas formas diferentes de indicar um parentesco, sobretudo no caso de um casamento (por exemplo: *mulher*, *esposa*, *a pessoa com quem casei*, *a minha patroa*, *a sua cara-metade*, ou *casaram-se*, *deram o nó*, *uniram os trapinhos*). Decidimos que o trabalho de as identificar e normalizar ficaria do lado dos sistemas participantes, e fixámos os nomes das relações de parentesco, nomeadamente: *mãe, pai, filho/a, neto/a, avó, avô, irmã/o, cunhado/a, primo/a, tio/a, sobrinho/a, bisavó, bisavô, bisneto/a, nora, genro, sogro/a, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado, afilhada.*

Alguns casos são o que poderíamos chamar de "parentesco social", e não laços de sangue. São os casos de *madrinha/padrinho*, *comadre/compadre* e *afilhado/afilhada*, e os casos de *madrasta/padrasto* e *enteado/enteada*. Por lapso, não colocámos estes últimos na lista.[21]

[21] Viemos a observar mais tarde que a relação de *madrasta* ou *padrasto* aparecia em sete das obras da coleção

Considerámos que as palavras *noivo* e *noiva* eram suficientemente próximas de formalização de um casamento para serem identificadas, ao contrário de *namorado*, *conversado*[22], etc.

Também considerámos que era relevante a "relação" de *viúvo* ou *viúva*, e que esta implicava uma situação marital diferente de *casado*.

Estabelecemos ainda que todas as relações que ocorressem entre as mesmas duas personagens durante a obra deviam ser encontradas, por isso `A noivo B` , `A marido B` e `A viúvo B` podiam ser a resposta certa, se a vida toda de A fosse contada na obra.

Mas é importante indicar que não requeríamos que uma relação e a sua inversa fossem indicadas, nem pelos sistemas participantes, nem na coleção dourada. O cálculo das relações inversas é feito durante a avaliação.

3.4.3. Que caracterização profissional escolher

Ao contrário das relações familiares, considerámos impossível normalizar e/ou prever de antemão tudo o que seria encontrado nas obras do DIP, e decidimos que a profissão, estatuto social e ocupação retornada deveria ser a encontrada na obra.

Mesmo assim, sugerimos que alguma reformulação teria de ser feita em casos como "despediram todos os cozinheiros, excepto a Maria". Num caso como esse, os sistemas deveriam colocar *cozinheira* na profissão da Maria.

Tal como em relação ao parentesco, todas as ocupações mencionadas na obra deveriam ser indicadas, por isso uma personagem poderia ter mais do que uma POES.

Se a personagem era descrita como ex-profissional, seria isso que constaria. Em princípio, e se descrito por ambas em épocas diferentes da obra, uma personagem podia ser por exemplo *professor* e *ex-professor*.

No caso de *aposentado* ou *reformado*, estipulámos que, se fosse antecedido pela profissão, ambas deviam ocorrer. Ou seja, *juiz aposentado*, *cozinheiro reformado*.

Quando a palavra *herdeiro* ou *herdeira* não aparecesse relacionada com outros de quem herda, e significasse uma pessoa rica porque herdou, também deveria ser considerada uma ocupação (ou falta dela).

3.4.4. Animais com nome

Se existissem animais com nome nas obras, eles deveriam ser considerados como personagens. No local do POES, deveria ser indicado o tipo de animal: cão, cavalo, etc.

3.4.5. Personagens que não chegam a existir

Este parece um caso estranho, mas não é tão raro como se poderia pensar. Por exemplo, considere-se a frase *imaginou que mais tarde teriam um filho a que chamariam Álvaro*. Nesse caso, determinámos que essa personagem deveria ser marcada.

3.4.6. Personagens provindas da loucura ou alucinação

Mais um caso que talvez não se imaginasse, sem ter lido de fio a pavio várias obras, é aquele em que as personagens deliram e pensam que eles e os outros são outras pessoas, como é o caso no *Quincas Borba* de Machado de Assis, em que a personagem principal enlouquece. Decidimos que, se tiverem nome, devem ser considerados como outros nomes (co-identificação) dessas mesmas personagens.

Também quando as profissões se referem a jogos de crianças, como *Laurita cozinheira* em *Amar, verbo intransitivo* de Mário de Andrade, consideramos que devem ser marcadas.

3.4.7. Nomes com partes em minúsculas

Em geral os nomes próprios em português são em maiúsculas, mas há um caso especial que é o das alcunhas (em português de Portugal) ou apelidos (em português do Brasil) em que só uma parte é em maiúscula, como em *João das pantorrinhas*. Nesse caso, e embora isso corresponda a uma exigência muito mais elevada, decidimos que seria necessário que o sistema identificasse o nome todo e não só a parte em maiúsculas.

3.4.8. Títulos de nobreza ou cargos

Finalmente, uma decisão muito importante e da qual mais tarde nos arrependemos, mas já não podíamos voltar atrás, sob pena de ter de refazer a leitura de várias obras já prontas, foi a de considerar que uma personagem mencionada apenas pelo seu título não era para identificar. E, da mesma forma, não era para identificar o título se fosse chamado por ele.

A justificação para esta decisão era de que poderia haver várias pessoas diferentes todas *conde*

dourada de texto, o que mostra que teria sido importante também identificar estes casos, já presentes, aliás, numa das obras de exemplo.

[22] Forma antiga de dizer namorado.

de Oeiras, ou *marquês da Palma*, e que o título não seria suficiente para as distinguir. Mas o que é certo é que verificámos mais tarde que, em muitas obras, sobretudo romances históricos, personagens extremamente importantes para o enredo eram assim descritas ao longo de toda a obra.

3.4.9. Outras microdecisões

Outras decisões referem-se a situações pontuais, mas que também não eram evidentes. Por exemplo, decidimos não marcar quando um nome próprio é chamado como um insulto, como no passo seguinte:

> O irrequieto arcebispo foi pôr cerco a Simancas; mas do alto das muralhas da velha cidade os sitiados escarneceram-n'o, chamando-lhe D. Opas; – o que significava compará-lo com o typo mais repugnante dos homens conhecidos por traidores (em *Pero da Covilhã*, de Zeferino Norberto Gonçalves Brandão)

Sobretudo em relação a profissões ou ocupações, muitas microdecisões tiveram de ser tomadas. Foi especialmente difícil decidir em relação a POES que têm um significado negativo ou usadas em contextos não tradicionais. Por exemplo, não considerámos *bohemio*, *fradalhão* ou *capataz de uma turma de vadios* como ocupações, mas marcámos *agiota* e *prostituta* em casos que poderiam ser interpretados como subjetivos.

Descrições como *dono de barco* ou *hospedeiro* (dono de hospedaria), em que não considerámos o primeiro como POES, mas sim o segundo, mostram como a fronteira é ténue entre aquilo que se pode ou não considerar uma atividade profissional, ocupação ou estatuto social. A explicação é que o dono de um barco — pelo menos no romance em questão — não necessita de estar associado às viagens marítimas, enquanto se pressupõe que o dono da hospedaria está lá a receber os hóspedes, e ocupa a maioria do seu tempo nessa atividade.

Finalmente, personagens no plural, tal como *os Pereiras*, ou *as manas Madureira*, não foram consideradas.

3.5. Caracterização da coleção dourada

Após a marcação das personagens em 43 obras, as 40 da coleção dourada e as 3 de exemplo, podemos dar uma primeira aproximação do assunto na literatura lusófona.[23]

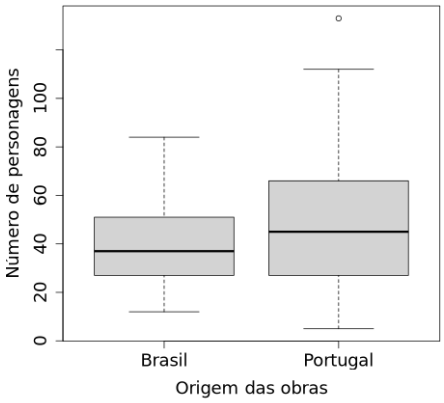

Figura 2: O número de personagens por obra, nas 43 obras brasileiras e portuguesas a que atribuímos uma solução

Na Figura 2, vemos o número de personagens nas 43 obras, por literatura. Este número variava entre 5 para a novela *A vinha* de Ana de Castro e Osório e 112 para o romance histórico *Pero da Covilhã* de Zeferino Norberto Gonçalves Brandão.

Observa-se que existe mais variação nas obras portuguesas, que também têm em média um número um pouco mais elevado de personagens.

Apenas para as 24 obras em texto para as quais temos a solução, averiguamos a relação entre o tamanho da obra e o número de personagens.

Na Figura 3 apresentamos o panorama relacionado com o tamanho em número de palavras das obras. Vemos que em geral existem mais personagens em obras mais longas, mas que existem algumas obras não muito longas com muitas personagens, e são todas romances históricos.

A Figura 4 apresenta a mesma questão de outra forma, indicando a densidade relativa de personagens nas 24 obras das quais temos o tamanho em palavras.

3.5.1. O género na coleção dourada

Se observarmos a panorâmica do género na coleção dourada, apresentada na Figura 5, observamos que em todas as obras — exceto uma, *A vinha*, uma novela de Ana de Castro Osório, com apenas 5 personagens, 3 mulheres — a maioria das personagens em cada romance são masculinas.

É preciso lembrar que estamos a medir a existência de todas as pessoas mencionadas na

[23] Convém referir, como discutiremos na Secção 4.2, que a marcação das personagens nas 21 obras da coleção de texto beneficiou do cotejo com os resultados do sistema participante.

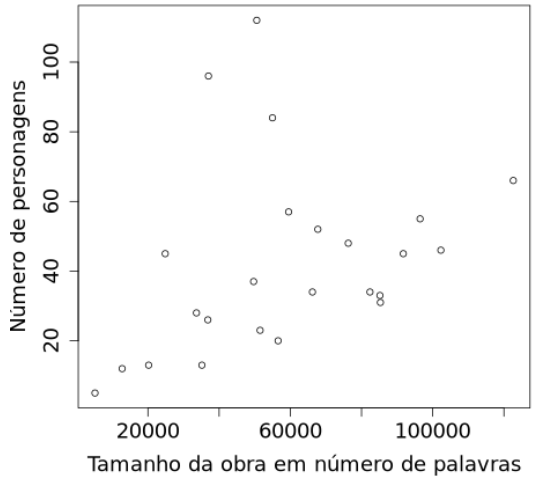

Figura 3: As obras em termos de tamanho e de número de personagens

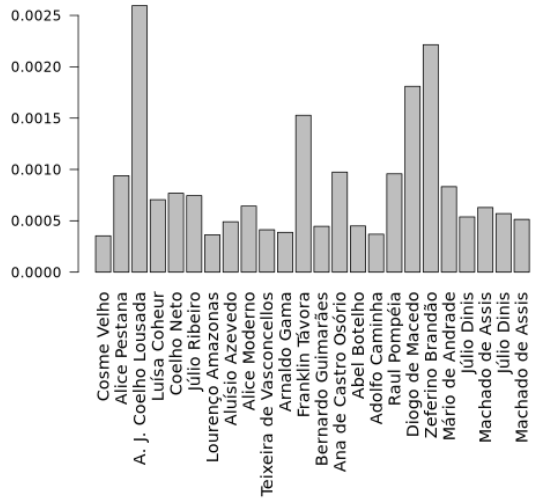

Figura 4: A densidade relativa do número de personagens por obra

obra por nome, não apenas as personagens principais. De facto, na novela que acabámos de mencionar, a personagem principal é um homem. Por isso a maioria de homens pode simplesmente significar que há mais homens na esfera pública na sociedade descrita nas obras, que as mulheres têm menos cargos, estão presentes sobretudo no âmbito privado.

Ao analisar o gênero dos personagens na literatura lusófona e sua profissão, é possível identificar desigualdades de género que refletem a realidade histórica e social das sociedades em que as obras foram escritas. Na maioria das obras, as personagens masculinas ocupam mais cargos públicos e têm mais visibilidade social do que as femininas, que muitas vezes são retratadas em papéis secundários e na esfera privada.

Esta conclusão é corroborada pela proporção maior de personagens femininas sem características de profissão, ocupação ou estatuto social em comparação com as personagens masculinas, como veremos a seguir.

No que se refere às profissões, estatuto social e ocupações, na coleção dourada total (incluindo os textos de exemplo), há 1944 personagens, das quais 942 (48,5%) não têm este tipo de caracterização.

No caso das personagens masculinas, 633 em 1504 não têm profissão, ocupação ou estatuto social, ou seja, 42,1%. No caso das personagens femininas, o mesmo ocorre para 309 em 440, ou seja, 70,2%.

Seja como for, para estudar a importância das mulheres nas obras em si, teríamos de primeiro identificar as personagens principais, e refazer as contas baseadas nestas.

3.5.2. O estatuto profissional na CD

Quanto às 409 diferentes POES identificadas (convém relembrar que muitas delas são apenas variações ortográficas da mesma, como *abadessa*, *abbadeça* ou morfológicas, *abade*), as mais frequentes foram *padre* (55), *general* (42), *escravo* (38) e *estudante* (37). Se adicionarmos a variação *escrava* (12) — *estudante* refere-se aos dois géneros, e não existe uma profissão equivalente a *padre* para mulheres[24] —, obteríamos 50 casos, portanto o segundo lugar.

Separando entre as duas literaturas, o panorama é surpreendentemente semelhante: Há 266 POES diferentes na literatura portuguesa, e 258 na literatura brasileira. E os casos mais frequentes são, nas obras brasileiras, os mesmos que em geral: *padre* (27), *escravo* (18), *general* (17) e *estudante* (16). Adicionando *escrava* (6), teríamos exatamente a mesma ordem da totalidade das obras. Nas obras portuguesas, a situação é a mesma: *padre* (28), *general* (25), *estudante* (21) e *escravo* (20), mas adicionando *escrava* (6) este estatuto social passa para segundo lugar.

A importância das profissões religiosas na literatura lusófona já tinha sido discutida por Santos (2022c) em relação ao número de vezes que as profissões eram mencionadas no texto literário, mas note-se que o estudo do DIP é diferente no sentido de contar as profissões das personagens — independentemente de serem mencionadas muitas vezes, serem personagens principais

[24] Não conhecemos casos de generais femininos nem sabemos qual a forma de mencionar, por isso podemos assumir, o que aliás se verificou nas obras lidas, que *general* se refere apenas a homens.

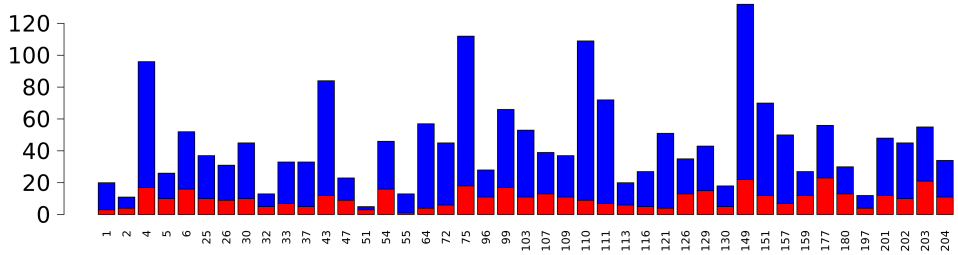

Figura 5: A distribuição de personagens por género na coleção dourada total: a vermelho, as personagens femininas; a azul, as masculinas

ou simples figurantes. Poder-se-ia imaginar que por exemplo padres seriam muitas vezes mencionados sem nome, e que portanto não necessariamente os dois estudos conduzissem aos mesmos resultados.

Por outro lado, quando uma personagem é padre e tem nome, será sempre descrita e mencionada por *padre X*, e geralmente tratada por *senhor padre X*, o que implica um grande acréscimo da palavra *padre* nos textos. Enquanto qualquer outra profissão, por exemplo médico, não seria usada em português para referir uma personagem que o fosse. (Seria tratada por *doutor Y*, não por *médico Y*).

Para uma análise mais detalhada das profissões no DIP, veja-se Pires et al. (2023).

3.5.3. As relações familiares na CD

Quanto às relações familiares nas 42 obras (visto que uma obra, como já mencionámos, não apresentava quaisquer relações entre as personagens), expandimos todas as relações passíveis de expansão, e obtivemos 810 relações familiares, apresentadas na Figura 6. Essas relações referiam-se a 777 personagens distintas.

Várias outras medidas e análises podem ser encontradas em Mota & Santos (2023), aqui apenas apontamos para a importância da relação pai, 106 casos (significativamente maior do que mãe, 64 casos) nas obras consideradas.[25]

3.5.4. Nomes diferentes para uma mesma personagem na CD

Quanto aos diferentes nomes pelos quais as personagens eram identificadas, que era um dos pressupostos do DIP, nomeadamente que seria um problema se não se identificasse a co-referência, as 43 obras que inspecionámos confirmaram indubitavelmente a necessidade de unir diferentes nomes. De facto, em todas as obras houve mais do que um nome para pelo menos uma personagem, como a Figura 7 ilustra.

Convém referir que a Figura 7 foi construída depois de termos juntado todas as possíveis formas de indicar a mesma coisa, apenas grafada diferentemente, ou seja, termos convertido por exemplo todos os casos de *sr.*, *snr.*, *Sr.*, e *Snr.* numa mesma forma, e corrigido o problema de acentos devido a reconhecimento ótico de carateres, como em *Álvaro*, *Alvaro* e *Àlvaro* referindo a mesma personagem numa dada obra. Se mantivéssemos as diferentes grafias da mesma forma de tratamento ao longo de uma obra, os números refletiriam ainda maior diversidade. Na Figura 8 mostramos a distribuição das personagens pelo número de formas diferentes por que são mencionadas, antes e depois da normalização.

3.5.5. As formas de tratamento na CD

Uma das características da língua portuguesa que também nos interessava explorar no DIP é a diversidade das formas de tratamento, tradicionalmente consideradas de grande complexidade na nossa língua. Aqui no DIP apenas poderíamos naturalmente identificar aquelas usadas em conjunção com um nome próprio, mas mesmo assim pudemos compilar alguns dados interessantes.

As formas de tratamento mais frequentes aparecem na Figura 9, em que as formas de senhor são claramente as mais usadas (com ou sem outras formas, como em *sr. dr.*). "redsenhor" corresponde a reduções de senhor, como *sôr*, ou *sinhô*, que não fazem parte da norma padrão mas que pretendem transmitir um certo dialeto ou socioleto. É interessante também a grande quantidade de *D.* (que corresponde a Dom ou Dona, por extenso). A palavra *sinhá* é apenas usada no Brasil, e não aparece frequentemente associada a um nome próprio nas obras que analisámos.

[25] Por outro lado pode-se também argumentar que, havendo mais personagens masculinas do que femininas, a relação de mãe é mais frequente para uma mulher nas obras (7,7%) do que a relação de pai para um homem (3,7%).

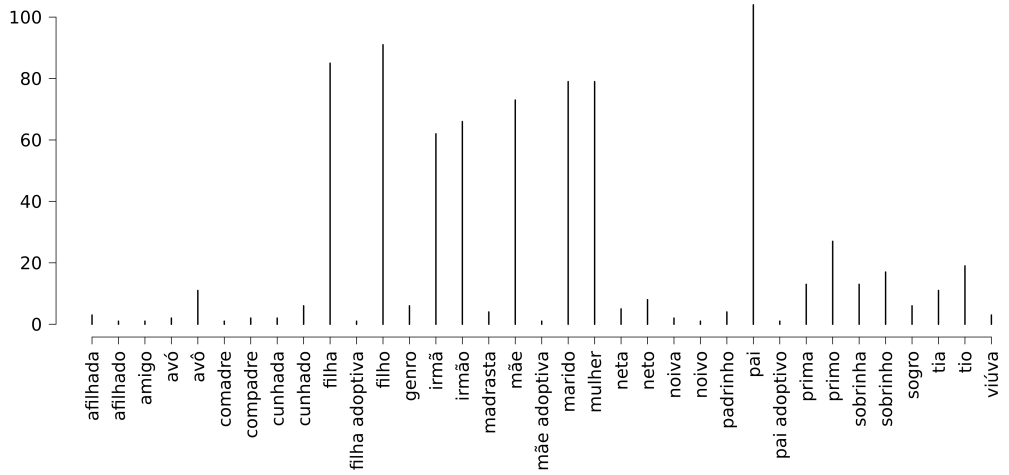

Figura 6: As relações familiares na coleção dourada total

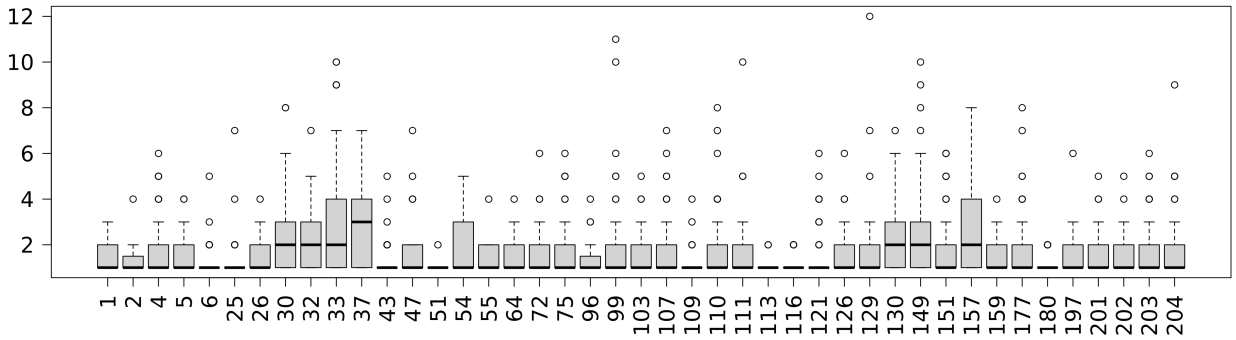

Figura 7: Número de nomes diferentes por personagem, depois da normalização, por obra

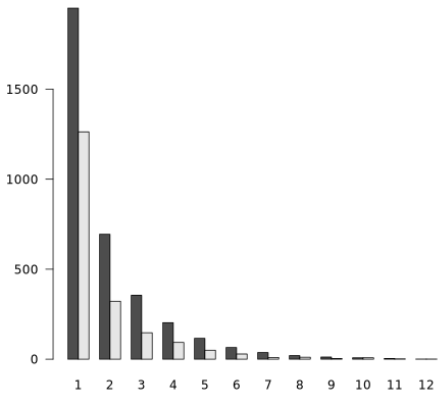

Figura 8: Número de nomes diferentes por personagem, antes e depois da normalização

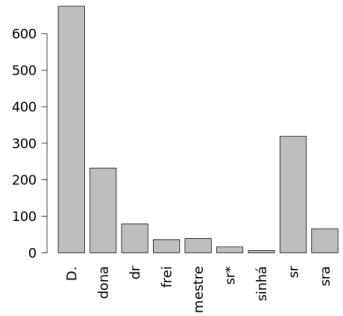

Figura 9: Formas de tratamento, depois da normalização. "sr*" significa formas não padrão do tratamento por senhor, como *seu*, *sinhô*, etc.

Outra vertente associada tanto a nomes diferentes como a formas de tratamento é o uso de diminutivos relativos a personagens. No artigo de apresentação do DIP à comunidade do processamento computacional do português (Santos et al., 2022), chegámos a apresentar a hipótese, baseada nas duas obras analisadas até aí, que o uso de diminutivos poderia diferir entre as duas literaturas, visto que em *As Pupilas do senhor Reitor* era usado para mostrar familiaridade e ternura para com as personagens, enquanto em *Dom Casmurro* era usado para distinguir mãe e filha com o mesmo nome.

Não nos debruçámos ainda sobre isso, por isso podemos apenas apresentar os valores quantitativos da existência de diminutivos nas obras da coleção dourada total, que apresenta 80 diminutivos, 44 masculinos e 36 femininos. 36 provêm da literatura de Portugal, e 44 da do Brasil. Ao todo são 57 diferentes.

Uma observação imediata é que em vários casos o diminutivo ocorre com outras formas de tratamento que em princípio indicariam mais distância e menos familiaridade, como *sr*, *sinhô*, *capitão*, *doutô*, etc. Em alguns casos, o diminutivo parece fazer parte da alcunha/apelido, como é o caso de *Miguel Mulatinho* ou *Mata Corcundinha*, ou ser aplicado ao apelido/sobrenome em vez de ao primeiro nome, como em *Mendonçazinho* ou *Pereirinha*. Parece-nos pois que a riqueza desta forma de tratamento merece ser mais explorada, para identificar o que está contido nestas formas de mencionar a personagem.

4. Resultados

4.1. Participação

Embora tivéssemos tido pelo menos quatro expressões de interesse, e tivéssemos mesmo adiado a data da avaliação conjunta para o semestre seguinte para ver se conseguíamos mais participantes, no fim apenas um participante enviou os resultados do seu sistema, o PALAVRAS-DIP, cuja descrição pode ser encontrada em Bick (2023).

O PALAVRAS-DIP participou apenas para as obras em formato de texto, o que resultou em que apenas pudéssemos calcular os resultados baseados em 21 obras da coleção dourada.

Pensamos que a falta de participação dos outros interessados, e da comunidade de estudos literários computacionais em geral, se pode dever a diferentes factores:

- Não haver suficientes grupos que trabalhem em estudos literários computacionais em língua portuguesa
- Não haver grande incentivo para esses grupos se dedicarem a uma tarefa que não estava nas prioridades deles
- A tarefa ser razoavelmente difícil, e geralmente feita por linguistas computacionais, ou investigadores de PLN (Processamento de Linguagem Natural): de facto, todos os interessados provinham desta área
- A tarefa, sendo um misto de extração de informação e de recolha de informação, também não se enquadrar naturalmente na maior parte das tarefas dos grupos de PLN
- Não podermos fornecer materiais de treino que permitissem usar técnicas de aprendizagem automática/aprendizado de máquina, que é o paradigma mais utilizado atualmente: o único sistema que concorreu é baseado em regras.

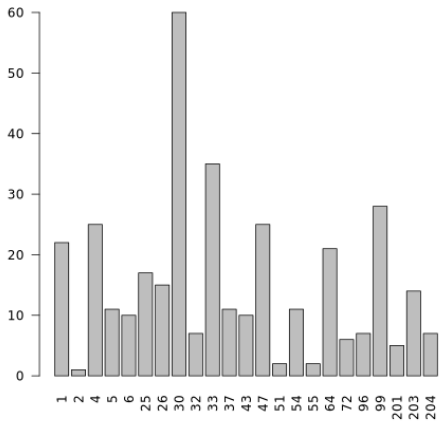

Figura 10: Melhoria das 24 obras da coleção dourada de texto e exemplos, graças ao cotejo com as respostas do PALAVRAS-DIP

4.2. A colaboração entre o resultado automático e a inspeção humana

Um dos resultados mais interessantes e inesperados que surgiu, quando principíamos a avaliação do PALAVRAS-DIP, foi o reconhecimento de que as soluções criadas por seres humanos tinham muitas faltas, que podiam ser facilmente colmatadas por um processo automático. Sobretudo em tarefas como identificar todas as grafias diferentes, os seres humanos não se comparam com uma máquina.

Por isso, fizemos uma nova ronda de "saneamento da coleção dourada", de forma a enriquecer a informação correta, e também para não penalizar a participação do sistema por ser avaliado com base em recursos deficientes.

Na Figura 10 mostramos a mais-valia contribuída pelo sistema participante, para a análise dos 21 textos da coleção dourada em texto e dos 3 textos exemplo em txt.

Em todas as obras foi possível melhorar o número de nomes das personagens, e em muitas delas também identificar mais casos de profissões, ocupações ou estatutos sociais.

4.3. Avaliação

Os resultados da participação do PALAVRAS-DIP foram tornados públicos, e são discutidos em pormenor em Willrich & Santos (2023).

Contudo, parece-nos que, mais do que os números obtidos, a existência de um sistema que conseguiu, embora não perfeitamente, fazer a tarefa que propusemos foi um dos resultados mais importantes do DIP.

Assim, temos uma forma de fazer leitura distante da literatura lusófona, mesmo que o sistema

não a faça perfeitamente. Vamos portanto apresentar de seguida o que aprendemos sobre as 100 obras usando o PALAVRAS-DIP como oráculo. E, depois, aquilo que podemos dizer sobre mais 213 obras, classificadas com uma nova versão do PALAVRAS-DIP em março de 2023, listadas no apêndice A. Essas 213 obras são todos os romances e novelas que faziam parte do corpo Literateca em março de 2013 e não tinham sido selecionadas para a coleção DIP de texto. Chamamos a esta coleção a "coleção extra".

4.4. O género no romance lusófono

Não houve diferenças significativas entre o que observámos com as 42 obras da coleção dourada, com as 100 da coleção de texto do DIP (que incluem 21 das primeiras), e as 213 obras que constituem a coleção extra.

Em praticamente todos os casos houve mais personagens masculinas. Nos dois casos de obras na coleção DIP em que se encontraram mais personagens femininas, eram escritos por mulheres e a personagem principal era um homem: *A vinha* de Ana de Castro Osório e *Jovens interessantes* de Paulina Filadélfia.

A Figura 11 mostra a distribuição de género na análise do PALAVRAS-DIP da coleção DIP, durante a avaliação conjunta.

Na coleção extra, houve onze obras em que o PALAVRAS-DIP identificou mais personagens femininas do que masculinas: *A feiticeira*, *Diário de uma criança* e *Sacrificada*, três novelas de Ana de Castro e Osório, *Um Homem de Brios* de Camilo Castelo Branco, *Os novelos da Tia Filomela*, uma novela de Júlio Dinis, *Herança de lágrimas*, de Ana Plácido, *Statira e Zoroastes*, de Lucas José de Alvarenga, *A Marquesa de Vale Negro*, de Maria O'Neill, *Astúcias de namorada*, de Manuel Pinheiro Chagas, e *Húmus* e *O pobre de pedir* de Raul Brandão.

A Figura 12 mostra a distribuição de género na análise do PALAVRAS-DIP da coleção extra em março de 2023, depois da melhoria do sistema baseada na avaliação do DIP.

4.5. Profissões

Nos resultados do PALAVRAS-DIP relativos à coleção DIP, das 6027 personagens, 4315 não têm profissão, ocupação ou estatuto social (71,6%). Se desagregarmos por género, 1275 mulheres em 1490 não têm este atributo, ou seja, 85,6%. Para os homens, 3040 em 4536 também não têm, ou seja, 67,0%.

Usando os resultados do PALAVRAS-DIP, obtemos exatamente 500 profissões distintas (99 femininas distintas e 428 masculinas distintas). As POES masculinas mais frequentes são *padre*, *general*, *escravo*, *estudante*, *rei* e *capitão*, enquanto que as femininas são *criada*, *rainha*, *escrava* e *princesa*.

No caso da coleção extra, o PALAVRAS-DIP identifica, em março de 2023, 896 profissões diferentes (com a ressalva de que muitas profissões são apenas variantes (orto)gráficas), e as mais frequentes são *padre* (365 casos!), *conde*, *rei* e *capitão*. Para mulheres, 180 profissões distintas foram identificadas pelo PALAVRAS-DIP (algumas erradamente, como por exemplo *abade*) e as profissões, ocupações ou estatutos sociais mais frequentes foram *criada* (72), *rainha*, *condessa* e *soror*.

Reconheceu além disso 22 personagens femininas que eram *escravas* e 3 *mucamas*, e 32 personagens masculinas que eram *escravos*.

Comparando superficialmente a literatura brasileira e a portuguesa através das POES mais frequentes, e embora ambas tenham como profissão mais frequente *padre*, as profissões que se seguem na literatura brasileira são *capitão*, *coronel*, *chefe* e *médico*, enquanto que na literatura portuguesa são *rei*, *conde*, *príncipe* e *mestre*, denunciando claramente o peso dos romances históricos. *Imperador*, pelo contrário, que aparentemente descreveria melhor a realidade brasileira, apenas aparece 13 vezes nesta, contra 12 de *rei*. (Para comparação, e relembrando que a coleção extra tem uma maioria de obras portuguesas, aparecem 14 *imperadores* e 110 *reis* na subcoleção portuguesa.)

4.6. Relações familiares

As relações familiares que o PALAVRAS-DIP identificou durante o DIP, e as relativas à coleção extra encontram-se nas Figuras 13 e 14.

É interessante ver que deixa de ser pai a relação mais frequente para ser filho em ambas as coleções, mas que pai continua a ser mais frequente que mãe.

4.7. Nomes e formas de tratamento

Apresentamos aqui os nomes mais comuns nas coleções analisadas, segundo o PALAVRAS-DIP, tendo sido manualmente removidos casos de erro.

A Figura 15 apresenta os nomes com frequência superior a 20 na coleção do DIP, e a Figura 16 os nomes com mais de 20 casos na coleção extra.

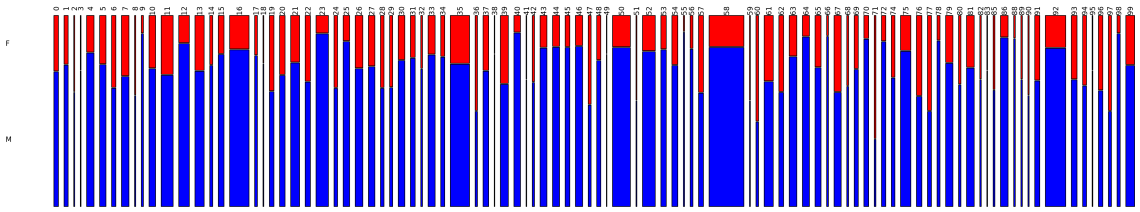

Figura 11: Género na coleção DIP de texto, de acordo com o PALAVRAS-DIP: cada coluna do mosaico representa uma obra, e a área vermelha marca as personagens femininas, e a azul as masculinas

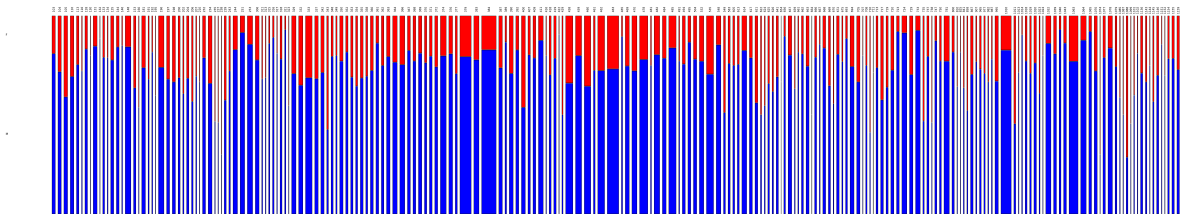

Figura 12: Género na coleção extra, de acordo com o PALAVRAS-DIP em março de 2023

Se quisermos apenas os nomes femininos, veja-se as Figuras 17 e 18.

Não é um assunto que seja provavelmente muito interessante do ponto de vista literário ou linguístico, mas é de notar que os nomes próprios mais frequentes não apresentam quase diferença nenhuma entre as literaturas portuguesa e brasileira, algo que muito provavelmente se deve à predominância de obras do século XIX.

Quanto às formas de tratamento, a situação também é semelhante nas várias coleções, veja-se as Figuras 19 e 20.

A maior diferença em relação à coleção dourada é o aparecimento da forma de tratamento *mestre*.

Debrucemo-nos agora sobre os diminutivos: Na coleção do DIP, o PALAVRAS-DIP identificou 299 diminutivos, 137 masculinos e 162 femininos. Destes, 157 eram portugueses, e 142 brasileiros.

Na coleção extra, o PALAVRAS-DIP identificou 497 diminutivos: 244 diminutivos masculinos e 253 femininos. Isto significa, visto que o PALAVRAS-DIP identificou 3281 personagens femininas e 10144 masculinas, que há muito mais diminutivos femininos: 7,7% contra 2,4%. Os 497 casos correspondem a 284 diminutivos diferentes.

4.8. Em resumo

Em resumo, embora certamente o PALAVRAS-DIP não consiga obter exatamente todas as personagens e só as personagens, os resultados acumulados nas duas coleções vão na mesma direção que a informação que tínhamos coligido manualmente na coleção dourada, o que nos dá esperança de que a visão — em leitura distante — da literatura lusófona que conseguimos obter, usando este sistema, seja relativamente correta.

5. Comentários finais

O DIP foi uma avaliação conjunta destinada a desenvolver sistemas que, dada uma obra literária, obtivessem as suas personagens e algumas características e relacionamentos destas. A ideia era conseguir olhar para a literatura lusófona como um todo e produzir algumas generalizações, assim como distinguir obras ou grupos de obras que se destacassem.

Oferecemos estes dados — que, aliás, se encontram públicos para permitir que outros investigadores os manipulem, corrijam e estudem — como um princípio para esse objetivo.

É importante salientar que o grosso da literatura lusófona ainda não se encontra em forma de texto de alguma qualidade, e que por isso todas as obras ainda não digitalizadas ou com um Reconhecimento óptico de Caracteres (ROC) muito deficiente não podem ainda ser tomadas em conta. Urge desenvolver sistemas de ROC fiáveis para a literatura não contemporânea em português.

Este artigo pretendeu antes do mais dar uma panorâmica geral sobre a organização do DIP e sobre os recursos compilados. Os outros artigos

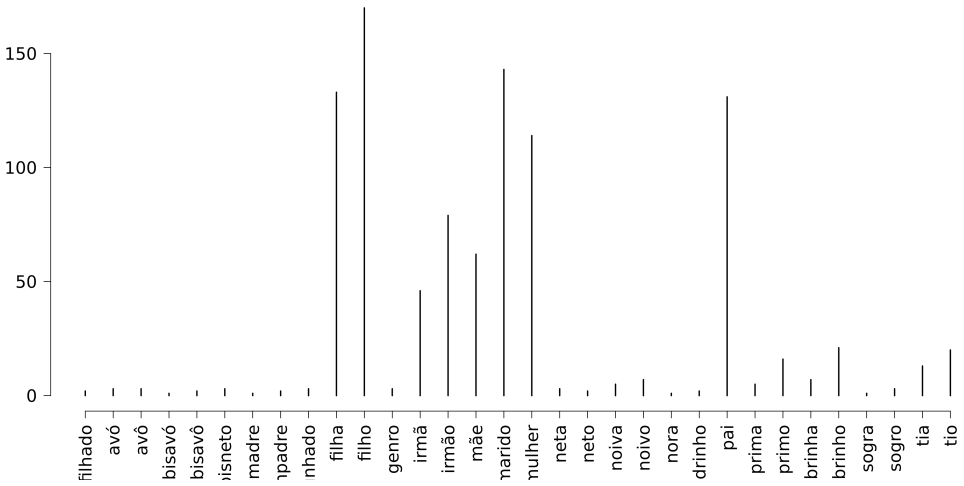

Figura 13: Relações familiares na coleção DIP de texto, de acordo com o PALAVRAS-DIP

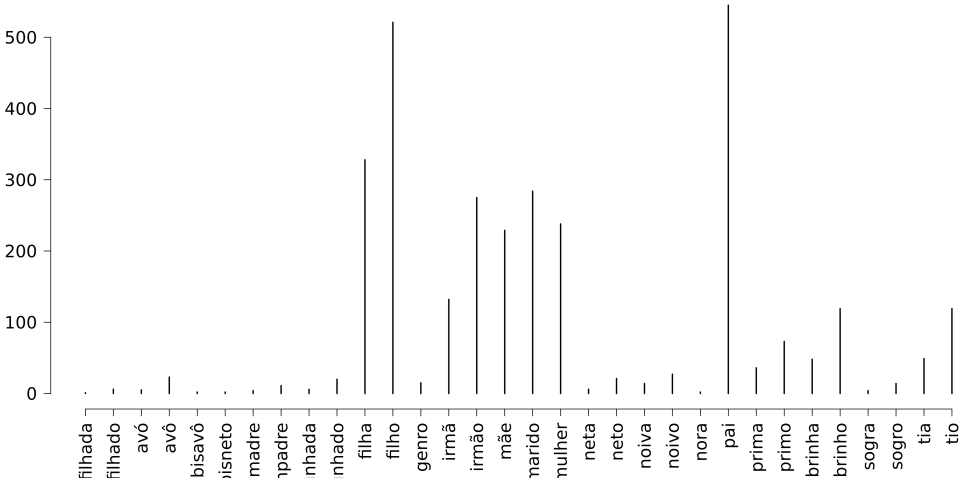

Figura 14: Relações familiares na coleção extra, de acordo com o PALAVRAS-DIP em março de 2023

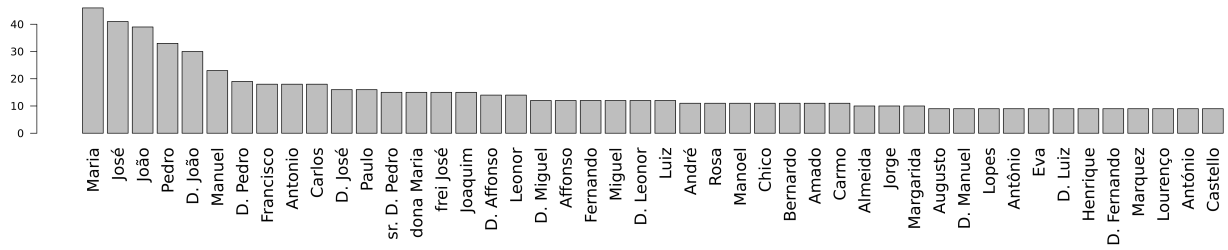

Figura 15: Nomes mais frequentes na coleção DIP de texto, de acordo com o PALAVRAS-DIP

deste volume descrevem com mais profundidade várias vertentes do DIP, como a caracterização do género e do estatuto profissional de um ponto de vista dos estudos literários (Pires et al., 2023), o estudo das relações familiares na literatura recorrendo a conceitos da teoria de redes (Mota & Santos, 2023), a forma de avaliação (Willrich & Santos, 2023) e, por último, mas provavelmente o mais importante, como o sistema participante resolveu a tarefa do DIP e como continua a evoluir (Bick, 2023).

Pensamos que, antes de organizar nova edição, é importante olharmos com muito cuidado para a informação sobre as obras que conseguimos obter, eventualmente melhorando-a e enriquecendo-a, de forma a propor outras maneiras de prosseguir na caracterização das personagens e das obras. É preciso que a comunidade se debruce sobre os dados já obtidos, os problemas encontrados, e os desejos expressos, para que todos possamos saber qual a contribuição que o DIP terá dado aos estudos literários lusófonos.

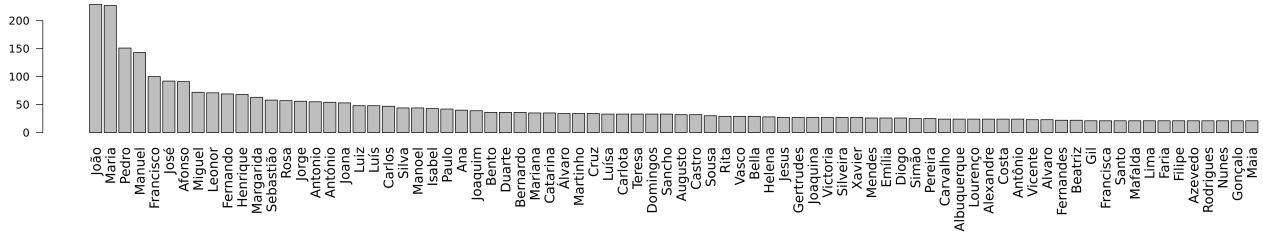

Figura 16: Nomes mais frequentes na coleção extra, de acordo com o PALAVRAS-DIP

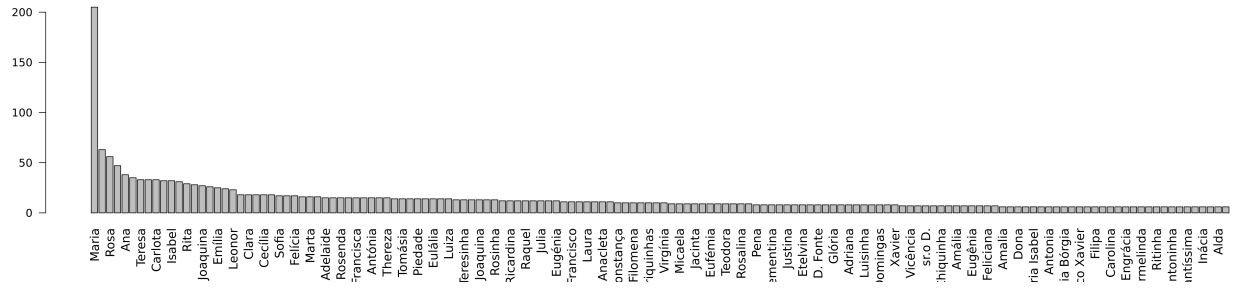

Figura 17: Nomes femininos mais frequentes na coleção DIP, de acordo com o PALAVRAS-DIP

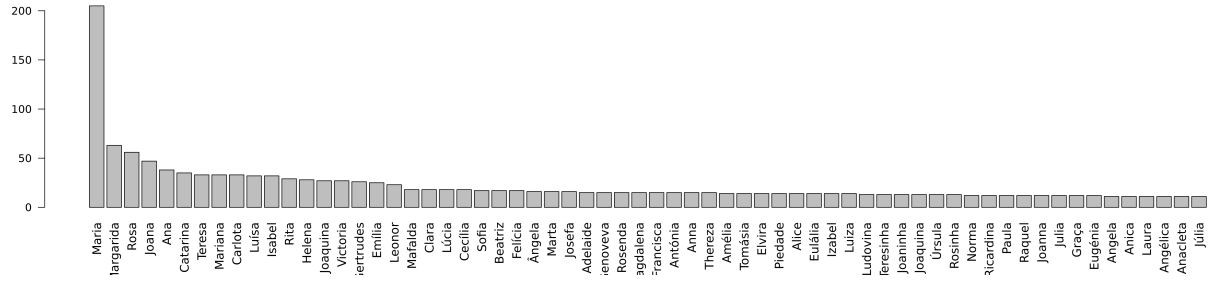

Figura 18: Nomes femininos mais frequentes na coleção extra, de acordo com o PALAVRAS-DIP

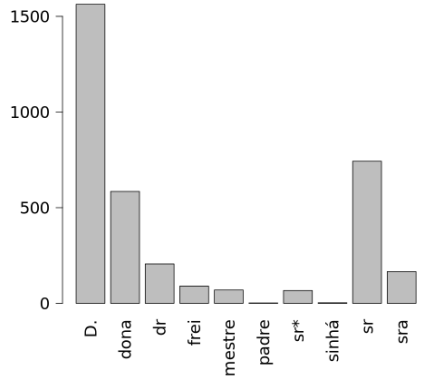

Figura 19: Formas de tratamento na coleção do DIP, de acordo com o PALAVRAS-DIP

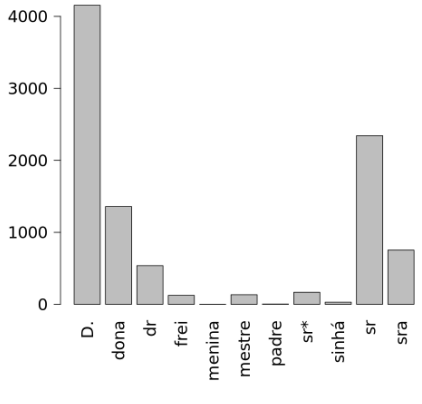

Figura 20: Formas de tratamento na coleção extra, de acordo com o PALAVRAS-DIP

Apresentamos pois o nosso trabalho apenas como um primeiro passo, cuja relevância depende de ter ou não estudiosos que se debrucem sobre os dados produzidos.

Agradecimentos

A organização do DIP tem de agradecer a muitas pessoas nele envolvidas. Em primeiro lugar, tivemos os compiladores da coleção dourada, que tiveram de ler as obras de fio a pavio e obter os resultados certos. Além da própria organização, agradecemos, por ordem alfabética, a Jonas Albuquerque, Luíla Lima, Luísa Lima, Marcus Vinicius Correa, Oriana Pereira, Patrícia Magalhães, Ruth Hoff e Sara Botelho.

Agradecemos calorosamente a Eckhard Bick, o único participante que conseguiu desenvolver um sistema para participar no DIP, sem o qual não poderíamos apresentar resultados.

Agradecemos aos participantes no Encontro do DIP, sobretudo aqueles que deram o seu contributo na hora da discussão, nomeadamente e mais uma vez por ordem alfabética, Cláudia Freitas, Luísa Coheur, Maria José Finatto, Raquel Amaro e Roberlei Alves Bertucci.

Agradecemos aos observadores do DIP pelo interesse e pelos comentários críticos, especialmente a Dionéia Motta, Alexandre Rademaker e Sílvia Araújo.

Agradecemos a Luisa Coheur, Alexandre Rademaker e Roberlei Alves Bertucci os muitos e pertinentes comentários a uma versão anterior deste texto, que contribuíram decisivamente para a sua melhoria.

Agradecemos o apoio da FAPEMA pelo financiamento de uma bolsa de pós-doutorado a Emanoel Pires.

Agradecemos também ao ILOS a contratação de dois assistentes de investigação para ajudar a organização do DIP, assim como o financiamento da viagem do participante a Oslo para o Encontro do DIP.

E finalmente, agradecemos à FCCN–Fundação para a Computação Científica Nacional (Portugal) o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências

Bick, Eckhard. 2023. Extraction of literary character information in Portuguese. *Linguamática* 15(1). 31–40. 10.21814/lm.15.1.397.

Cardoso, Nuno & Diana Santos. 2007. Directivas para a identificação e classificação semântica na colecção dourada do HAREM. Em Diana Santos & Nuno Cardoso (eds.), *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 211–238. https://www.linguateca.pt/aval_conjunta/LivroHAREM/Cap16-SantosCardoso2007-CardosoSantos.pdf.

de Does, Jesse, Katrien Depuydta, Karina van Dalen-Oskamb & Maarten Marx. 2017. Namescape: Named entity recognition from a literary perspective. Em Jan Odijk & Arjan van Hessen (eds.), *CLARIN in the Low Countries*, 361–370.

Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. 2019. Distant reading Brazilian politics. Em *4th Conference of The Association Digital Humanities in the Nordic Countries*, 190–200.

Krug, Markus, Lukas Weimer, Isabella Reger, Luisa Macharowsky, Stephan Feldhaus, Frank Puppe & Fotis Jannidis. 2018. Description of a corpus of character references in german novels: DROC. Relatório técnico. Georg-August-Universität. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2018-27.pdf.

Langfeldt, Marcia Caetano, Emanoel Pires, Rebeca Schumacher Fuão & Ricardo Gaiotto. 2021. Considerações sobre a personagem literária. https://www.linguateca.pt/aval_conjunta/dip/personagem.html.

Mota, Cristina & Diana Santos. 2023. Pais, filhos, e outras relações familiares no DIP. *Linguamática* 15(1). 41–53. 10.21814/lm.15.1.402.

Pires, Emanoel, Marcia Caetano Langfeldt & Rebeca Schumacher Fuão. 2023. Desafios e vantagens do processo de identificação automática do gênero e das profissões das personagens no dip. *Linguamática* 15(1). 55–67. 10.21814/lm.15.1.401.

Santos, Diana. 2019. Literature studies in literateca: between digital humanities and corpus linguistics. Em Martin Doerr, Øyvind Eide, Oddrun Grønvik & Bjørghild Kjelsvik (eds.), *Humanists and the digital toolbox: In honour of Christian-Emil Smith Ore*, 89–109. Novus Forlag.

Santos, Diana. 2022a. Evaluation contests in Portuguese: Linguateca. *submitted* https://www.linguateca.pt/Diana/download/AvalConjLRE.pdf.

Santos, Diana. 2022b. Futuro risonho: prolegómenos para uma colaboração entre a Linguateca e o NuPILL. Em Isabela Melim Borges & Paulo Henrique Pergher (eds.), *Literatura e seus híbridos III: 25 anos do NuPiLL*, 285–308. UFSC.

Santos, Diana. 2022c. A gramateca e a literateca como macroscópios linguísticos. *Domínios de Lingu@gem* 16(4). 1242–1265. 10.14393/DL52-v16n4a2022-2.

Santos, Diana, Daniel Alves, Raquel Amaro, Isabel Araújo Branco, Olivia Fialho, Cláudia Freitas, Suemi Higuchi, Marcia Langfeldt, João Marques Lopes, Alckmar Luiz dos Santos, Emanoel Pires, Barbara Ramos, Danielle Sanches, Rebeca Schumacher Fuão, Paulo Silva Pereira & Paula Terra. 2020a. Leitura distante em Português: resumo do primeiro encontro. *Materialidades da Literatura* 8(1). 279–298. 10.14195/2182-8830_8-1_16.

Santos, Diana, Eckhard Bick & Marcin Wlodek. 2020b. Avaliando entidades mencionadas na coleção ELTeC-por. *Linguamática* 12(2). 29–49. 10.21814/lm.12.2.336.

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

Santos, Diana, Cláudia Freitas & Eckhard Bick. 2018. OBras: a fully annotated and partially human-revised corpus of Brazilian literary works in public domain. Em *CorLex*, s.p.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational processing of the Portuguese language (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Schöch, Christof, Tomaz Erjavec, Roxana Patras & Diana Santos. 2021. Creating the European Literary Text Collection (ELTeC): Challenges and perspectives. *Modern Languages Open* 1. 1–19. 10.3828/mlo.v0i0.364.

Willrich, Roberto & Diana Santos. 2023. Avaliação no desafio de identificação de personagens. *Linguamática* 15(1). 69–87. 10.21814/lm.15.1.398.

A. A lista das obras da coleção do DIP

Id	Autor	Título	Ano
0	Miguel Vale de Almeida	*Euronovela*	1998
1	Cosme Velho	*Miss Kate*	1909
2	Alice Pestana	*A vida por um prejuízo*	1908
3	José da Rocha Leão	*A cruz de fogo*	1862
4	António José Coelho Lousada	*Os tripeiros*	1857
5	Luísa Marques da Silva	*mISTério@Tagus*	2021
6	Coelho Neto	*Turbilhão*	1904
7	Lindolfo Rocha	*Maria Dusá*	1910
8	José Joaquim Rodrigues de Bastos	*O médico do deserto*	1864
9	Apolinário Porto-Alegre	*O vaqueano*	1872
10	Inglês de Sousa	*O cacaolista: scenas da vida do Amazonas*	1899
11	Manuel de Oliveira Paiva	*Dona Guidinha do Poço*	1891
12	Artur Lobo de Ávila	*Os Caramurús: romance histórico da descoberta e independência do Brasil*	1900
13	Luís Guimarães Júnior	*Histórias para gente alegre: A família Agulha*	1870
14	Antônio Gonçalves Teixeira e Souza	*Maria ou a menina roubada*	1852
15	Conde de Ficalho	*Uma Eleição Perdida*	1888
16	António Campos Junior	*A Ala dos Namorados*	1905
17	António Pedro Lopes de Mendonça	*Memorias d'um doido: romance contemporâneo*	1849
18	Guiomar Torrezão	*Severina*	1890
19	Almiro Caldeira da Andrade	*Arca açoriana*	1984
20	Faustino da Fonseca	*Os bravos do Mindelo: Romance histórico*	1906
21	Alfredo de Mesquita	*A rua do oiro*	1905
22	Domingos Olímpio	*Luzia Homem*	1903
23	Luís Augusto Rebelo da Silva	*A Casa dos Fantasmas*	1908
24	Maria O'Neill	*Por bom caminho*	1914
25	Júlio Ribeiro	*A carne*	1888
26	Lourenço Amazonas	*Simá*	1899
27	Antonio Lobo	*A carteira de um neurastênico*	1903
28	Pedro Ivo	*O selo da roda*	1876
29	Manuel Antonio de Almeida	*Memórias de um sargento de milícias*	1852
30	Aluísio Azevedo	*Casa de pensão*	1884
31	Visconde de Taunay	*No declínio*	1889
32	Alice Moderno	*O Dr. Luís Sandoval*	1892
33	Antonio Augusto Teixeira de Vasconcellos	*A ermida de Castromino*	1870
34	Rodolpho Theophilo	*O paroara: scenas da vida cearense e amazonica*	1899
35	João da Câmara	*O Conde de Castel Melhor*	1903
36	Manuel Maria Rodrigues	*A Rosa do Adro*	1870
37	Arnaldo Gama	*El-rei dinheiro*	1876
38	Alfredo de Morais Pinto	*Aventuras do sr. Criptogamo : romance humorístico*	1899
39	João do Rio	*A Profissão de Jacques Pedreira*	1910
40	Teófilo Braga	*Viriato*	1904
41	Maria Amália Vaz de Carvalho	*Alice*	1877

(Continua)

Id	Autor	Título	Ano
42	Raul de Azevedo	*Tríplice Aliança*	1907
43	Franklin Távora	*O cabeleira: História pernambucana*	1876
44	Luiz Gonzaga Duque Estrada	*Mocidade morta*	1899
45	Urbano Loureiro	*A infâmia de Frei Quintino*	1878
46	António Francisco Barata	*Um duelo nas sombras, ou D. Francisco Manuel de Melo (1630)*	1875
47	Bernardo Guimarães	*A escrava Isaura*	1875
48	Francisco Gomes de Amorim	*Os selvagens*	1875
49	Lima Barreto	*Triste Fim de Policarpo Quaresma*	1911
50	Oliveira Mascarenhas	*O trovador da infanta*	1906
51	Ana de Castro Osório	*A vinha*	1908
52	Alberto Pimentel	*A guerrilha de Frei Simão*	1895
53	Ramalho Ortigão e Eça de Queirós	*O Mistério da Estrada de Sintra*	1870
54	Abel Botelho	*O Barão de Lavos*	1891
55	Adolfo Caminha	*O Bom-Crioulo*	1895
56	Souza Lima	*O Tupinambá*	1931
57	José Augusto Vieira	*A divorciada*	1881
58	Rocha Martins	*Gomes Freire*	1900
59	Viriato Bandeira Duarte	*Ida*	1865
60	Júlia Lopes de Almeida	*A viúva Simões*	1895
61	Teixeira de Queirós	*O Salústio Nogueira*	1909
62	Ana Plácido	*Adelina*	1863
63	Almeida Garrett	*Viagens na minha terra*	1846
64	Raul Pompéia	*O Ateneu*	1888
65	Graça Aranha	*A viagem maravilhosa*	1929
66	Machado de Assis	*O alienista*	1882
67	Carlos Malheiro Dias	*Filho das ervas*	1900
68	Pardal Mallet	*Hóspede*	1887
69	Aluísio Azevedo	*O coruja*	1889
70	Alexandre Herculano	*O bobo*	1843
71	Paulina Filadélfia	*Jovens interessantes*	1865
72	Diogo de Macedo	*O Cristão novo*	1876
73	José do Patrocínio	*Os retirantes*	1879
74	Francisco Luís Gomes	*Os Brâmanes*	1866
75	Zeferino Norberto Gonçalves Brandão	*Pero da Covilhã: Episódio Romântico do Século XV*	1897
76	M.M.S.A. e Vasconcelos	*O Cura de São Lourenço*	1855
77	Augusto Loureiro	*A bruxa: cenas açorianas*	1901
78	Amadeu Amaral	*Memorial de um passageiro de bonde*	1938
79	Valentim Magalhães	*Flor de sangue*	1897
80	Maria Peregrina de Sousa	*Henriqueta*	1867
81	Eça de Queirós	*A ilustre Casa de Ramires*	1900
82	Bruno Seabra	*Paulo*	1861
83	José de Alencar	*A viuvinha*	1857
84	Antônio de Alcântara Machado	*Brás, Bexiga e Barra Funda*	1927
85	Maria Firmina dos Reis	*Úrsula*	1859
86	Paulo Setúbal	*O sonho das esmeraldas*	1935
87	Antônio de Alcântara Machado	*Mana Maria*	1936
88	Manuel Pinheiro Chagas	*Um melodrama em Santo Tirso*	1873
89	Caetano Alves de Sousa Filgueiras	*Adelaide de Sargans*	1870

(Continua)

Id	Autor	Título	Ano
90	Raúl Brandão	*A Morte do Palhaço*	1926
91	Emília Bandeira de Melo	*A luta*	1911
92	Camilo Castelo Branco	*A Brasileira de Prazins*	1882
93	Virgílio Várzea	*George Marcial*	1901
94	Virgínia de Castro e Almeida	*Aventuras de Dona Redonda*	1943
95	Casimiro de Abreu	*Carolina*	1856
96	Mário de Andrade	*Amar, verbo intransitivo*	1927
97	Joaquim Manoel de Macedo	*A Moreninha*	1844
98	Moreira de Azevedo	*Os franceses no Rio de Janeiro*	1870
99	Júlio Dinis	*Uma família inglesa*	1867
100	Pinheiro Chagas	*O terremoto de Lisboa*	1874
101	António Inocêncio Barbuda	*O português generoso, ou Aventuras de J... e S... e seu ditoso fim: História verdadeira*	1820
102	A. Augusto de Pinho	*Remédio para matar paixões*	1879
103	Afonso Arinos de Melo Franco	*Os jagunços*	1897
104	Camilo Castelo Branco	*O judeu*	1866
105	Luiz Caetano de Campos	*Viagens de Altina, nas cidades mais cultas da Europa e nas principais povoações dos Balinos, povos desconhecidos de todo o mundo*	1790
106	Eduardo de Borja Reis	*O crime do beato Antônio: Romance original português*	1887
107	Carneiro Vilela	*Noêmia*	1894
108	Claudia de Campos	*Último amor*	1894
109	Germano Hasslocher	*A espelunca: Romance de atualidade*	1889
110	Pinheiro Chagas	*Os guerrilheiros da Morte*	1872
111	Domingos Jaguaribe	*Os herdeiros do caramuru*	1880
112	Érico Veríssimo	*Caminhos Cruzados*	1935
113	Araripe Júnior	*Luizinha: Romance de costumes cearenses*	1878
114	Alice Pestana	*A filha do João do Outeiro*	1933
115	Flávio de Carvalho	*Os ossos do mundo*	1936
116	Felício dos Santos	*Acayaca*	1866
117	Lúcio de Mendonça	*O marido da adúltera: Crônica fluminense*	1881
118	Visconti Coaracy	*Amor que mata*	1873
119	António Ernesto Tavares de Andrade	*Eugénio, ou o livre pensador*	1871
120	Bezerra de Menezes	*A casa mal-assombrada: Romance de costumes sertanejos*	1888
121	António Joaquim da Rosa	*A cruz de cedro*	1854
122	Teixeira e Sousa	*O filho do pescador: Romance brasileiro original*	1843
123	José Agostinho de Macedo	*O arrependimento premiado: História verdadeira*	1818
124	Francisco Coelho Duarte Badaró	*Fantina (Cenas da escravidão)*	1881
125	António Francisco Barata	*Os jesuítas na corte*	1877
126	Francisca Senhorinha da Motta Dinis	*A Judia Raquel*	1886
127	Menotti Del Picchia	*Laís*	1921
128	José Lins do Rego	*Doidinho*	1934
129	Carlos Pinto de Almeida	*A filha do emir*	1875
130	Alfredo Hogan	*A pedinte de Lisboa*	1859
131	Maria ONeill	*Lucta de sentimentos*	1912

(Continua)

Id	Autor	Título	Ano
132	Pedro Américo	*O foragido*	1899
133	Pedro Ribeiro Viana	*Elzira, a morta virgem*	1883
134	Tomás de Mello	*O Conde de S. Luiz*	1874
135	Reinaldo Ferreira	*O Presidente da República*	1923
136	Júlio Ribeiro	*Padre Belchior de Pontes*	1874
137	Cornélio Pena	*A menina morta*	1954
138	Júlio César Machado	*Cláudio. Romance*	1852
139	Alberto Pimentel	*As netas do Padre Eterno*	1895
140	Vicente Temudo Lessa	*O velho manuscrito*	1888
141	Luísa F. de Camargo Pacheco	*Alice*	1903
142	Lourenço Caiola	*Conversão*	1923
143	Ana Luísa de Azevedo Castro	*D. Narcisa de Vilar: Legenda do tempo colonial pela Indígena do Ipiranga*	1858
144	desc.	*O sapateiro de Azeitão*	1865
145	Elias António da Fonseca	*Doroteia, ou A lisbonense infeliz*	1816
146	Batista Cepelos	*O vil metal*	1910
147	Rosendo Moniz	*Favos e travos*	1872
148	José da Fonseca	*Aventuras de Telêmaco, filho de Ulisses, seguidas das de Aristonoo e de Ulisses: Compendiadas para uso dos meninos*	1854
149	Carlos Pinto de Almeida	*Os homens da cruz vermelha*	1879-1880
150	Augusto Emílio Zaluar	*O doutor Benignus*	1874
151	Teixeira de Vasconcellos	*O prato de arroz doce*	1862
152	Fortunato Correia Pinto	*O agitador*	1906
153	Júlio Lourenço Pinto	*O Bastardo: Cenas da vida contemporânea.*	1889
154	Leonel de Alencar, Barão de Alencar	*A sonâmbula da Itapuca*	1861
155	Bruno Seabra	*Memórias de um pobre diabo*	1868
156	Antônio Deodoro de Pascual	*Um episódio da história pátria: As quatro derradeiras noites dos inconfidentes (1792)*	1868
157	L.A. Rebello da Silva	*Ódio velho nao cança*	1848
158	Conde Afonso Celso	*Lupe*	1894
159	Virgínia de Castro e Almeida	*Trabalho bem-dito*	1908
160	José da Rocha Leão	*Os subterrâneos do Morro do Castelo: Seus mistérios e tradições*	1878
161	Eça de Queirós	*O mandarim*	1880
162	Dionísia Gonçalves Pinto	*Dedicação de uma amiga*	1850
163	Joaquim Norberto	*O martírio do Tiradentes, ou Frei José do Desterro. Lenda Brasileira*	1882
164	Antônio Manuel Policarpo da Silva	*Cadelinha*	1816
165	Antero de Figueiredo	*Leonor Teles: Flor de altura*	1916
166	Teixeira de Queiroz	*Morte de D. Agostinho*	1895
167	Emília Freitas	*A rainha do ignoto: Romance psicológico*	1899
168	António Joaquim de Mesquita e Melo	*D. Sancho II, quarto rei de Portugal*	1869
169	Medeiros e Albuquerque, Afrânio Peixoto, Coelho Neto, Viriato Correia	*O mistério*	1920
170	Lúcio Bruno	*A mão negra e a polícia: Sensacional romance dos crimes célebres, praticados pelo Dioguinho, o terror dos sertões paulistas*	1923
171	Mário de Sá Carneiro	*A confissão de Lúcio*	1913
172	João de Andrade Corvo	*O sentimentalismo*	1871

(Continua)

Id	Autor	Título	Ano
173	Soeiro Pereira Gomes	*Esteiros*	1941
174	D. Bruno da Silva	*A beata de Évora: Romance histórico 1764-1828*	1890
175	J. M. Pereira da Silva	*Jerônimo Corte Real*	1840
176	Carlos Malheiro Dias	*A mulata*	1975
177	Francisco Gomes de Amorim	*As duas fiandeiras*	1881
178	Carolina Michaëlis de Vasconcelos e Afonso Lopes Vieira	*O romance de Amadis: Composto sobre o Amadis de Gaula, de Lobeira*	1922
179	Luís da Silva Alves de Azambuja Suzano	*O Capitão Silvestre e Frei Veloso, ou A plantação de Café no Rio de Janeiro*	1847
180	Rachel de Queiroz	*O quinze*	1930
181	Luís Ratozi	*Amores pagãos*	1934
182	Cônego Ulisses de Penaforte	*Mandu (o eremícola): Romance indo-brasileno neontológico e nativista*	1901
183	Francisco Soares Franco Júnior	*Memórias da mocidade:*	1867
184	Ana de Castro Osório	*Mundo novo*	1927
185	Alfredo Campos	*A filha do cabinda*	1873
186	Xavier Marques	*O feiticeiro*	1922
187	Júlio César Leal	*Casamento e mortalha no céu se talha*	1876
188	Barão de Teffé	*A corveta Diana: Romance marítimo. Original brasileiro.*	1873
189	Ana Plácido	*Herança de lágrimas*	1871
190	Luís Ramos Figueira	*Dalmo ou Mistérios da noite*	1863
191	Salvador de Mendonça	*Marabá*	1875
192	José Antônio do Vale Caldre Fião	*A divina pastora: Novela rio-grandense*	1847
193	Alberto Osório de Castro	*Dramas da côrte*	1905
194	Alberto Braga	*Os confidentes*	1887
195	Dona Maria Benedita Câmara de Bormann	*Estátua de neve*	1890
196	Ana Maria Ribeiro de Sá	*Matilde*	1874
197	João Salomé Queiroga	*Maricota e o Padre Chico*	1871
198	Arnaldo Gama	*O satanás de Coura: Memórias do século XVII*	2002
199	Almada Negreiros	*Nome de guerra*	1938

B: A lista das obras da coleção extra

Id	Autor	Título	Ano
103	Abel Botelho	*Amanhã*	1901
104	Abel Botelho	*Amor crioulo*	1919
105	António da Costa Couto Sá de Albergaria	*Os filhos do padre Anselmo*	1904
110	Adolfo Caminha	*A Normalista*	1893
113	Adolfo Caminha	*Tentação*	1896
128	Alberto Osório de Castro	*Dramas da corte*	1905
130	Alberto Pimentel	*Cristo não volta*	1873
131	Alberto Pimentel	*O Anel Misterioso: Cenas da Guerra Peninsular*	1873
132	Alberto Pimentel	*A última ceia do Doutor Fausto*	1876
133	Alberto Pimentel	*As noites do asceta*	1876

(Continua)

Id	Autor	Título	Ano
134	Alberto Pimentel	*O Romance da Rainha Mercedes*	1879
135	Alberto Pimentel	*Noites de Sintra*	1892
144	Alexandre Herculano	*Eurico o Presbítero*	1844
146	Alexandre Herculano	*O Galego*	1846
148	Alexandre Herculano	*O Monge de Cister I*	1848
153	Alexandre Herculano	*O Pároco de Aldeia*	1851
160	Alfredo Campos	*A filha do Cabinda*	1873
181	J. B. da Silva L. de Almeida Garrett	*O Arco de Santana*	1845
191	J. B. da Silva L. de Almeida Garrett	*Helena*	1854
192	José de Almada Negreiros	*A engomadeira: novela vulgar lisboeta*	1917
195	Aluísio Azevedo	*Uma lágrima de mulher*	1879
196	Aluísio Azevedo	*O Mulato*	1881
197	Aluísio Azevedo	*A Condessa Vésper ou Memórias de um Condenado*	1882
198	Aluísio Azevedo	*Girândola de Amores ou Mistério da Tijuca*	1882
200	Aluísio Azevedo	*Filomena Borges*	1884
202	Aluísio Azevedo	*O Homem*	1887
204	Aluísio Azevedo	*O Cortiço*	1890
206	Aluísio Azevedo	*A Mortalha de Alzira*	1894
207	Aluísio Azevedo	*O Livro de uma Sogra*	1895
210	Álvaro do Carvalhal	*Os Canibais*	1868
232	A.M. da Cunha e Sá	*Da parte d'el-rei*	1873
234	Ana de Castro Osório	*Ambições*	1903
235	Ana de Castro Osório	*A feiticeira*	1908
237	Ana de Castro Osório	*Diário de uma criança*	1908
238	Ana de Castro Osório	*Sacrificada*	1908
239	Ana Plácido	*Adelina*	1863
243	Anna Maria Ribeiro de Sá	*Matilde*	1874
244	António de Albuquerque	*O Marquês da Bacalhoa*	1908
251	António Francisco Barata	*O Manuelinho de Évora*	1873
253	António Francisco Barata	*O último cartuxo da Scala Caeli de Évora: Romance histórico (1808-1865)*	1891
306	Augusto Sarmento	*Providência*	1863
311	Bernardo Guimarães	*O ermitão do Muquém*	1868
313	Bernardo Guimarães	*O seminarista*	1872
315	Bernardo Guimarães	*Maurício*	1877
316	Bernardo Guimarães	*O bandido do Rio das Mortes*	1905
317	Bernardo Guimarães	*O garimpeiro*	1972
318	Bernardino Pereira Pinheiro	*Arzila: Romance do Século XV*	1862
321	Brito Camacho	*Ao de leve*	1913
323	Bulhão Pato	*A pálida estrela*	1864
329	Camilo Castelo Branco	*Anátema*	1851
332	Camilo Castelo Branco	*A Filha do Arcedíago*	1854
333	Camilo Castelo Branco	*Mistérios de Lisboa*	1854
337	Camilo Castelo Branco	*Livro Negro de Padre Dinis I*	1855
342	Camilo Castelo Branco	*Onde Esta a Felicidade*	1856
343	Camilo Castelo Branco	*Um Homem de Brios*	1856
348	Camilo Castelo Branco	*A Vingança*	1858
349	Camilo Castelo Branco	*O Que Fazem Mulheres*	1858
350	Camilo Castelo Branco	*Cenas da Foz*	1860
352	Camilo Castelo Branco	*Romance dum Homem Rico*	1861
353	Camilo Castelo Branco	*Amor de Perdição*	1862

(Continua)

Id	Autor	Título	Ano
354	Camilo Castelo Branco	*Coisas Espantosas*	1862
355	Camilo Castelo Branco	*Coração Cabeça e Estômago*	1862
358	Camilo Castelo Branco	*Aventuras de Basílio Fernandes Enxertado*	1863
360	Camilo Castelo Branco	*O Bem e o Mal*	1863
361	Camilo Castelo Branco	*A Filha do Doutor Negro*	1864
362	Camilo Castelo Branco	*Amor de Salvação*	1864
363	Camilo Castelo Branco	*No Bom Jesus do Monte*	1864
364	Camilo Castelo Branco	*Vinte Horas de Liteira*	1864
366	Camilo Castelo Branco	*A Queda dum Anjo*	1866
367	Camilo Castelo Branco	*O olho de vidro: romance histórico*	1866
368	Camilo Castelo Branco	*A Doida do Candal*	1867
369	Camilo Castelo Branco	*O Retrato de Ricardina*	1868
370	Camilo Castelo Branco	*Os Brilhantes do Brasileiro*	1869
371	Camilo Castelo Branco	*A Infanta Capelista*	1872
372	Camilo Castelo Branco	*Livro de Consolação*	1872
374	Camilo Castelo Branco	*O Carrasco de Vitor Hugo*	1872
376	Camilo Castelo Branco	*A Filha do Regicida*	1875
377	Camilo Castelo Branco	*A Freira no Subterraneo*	1875
379	Camilo Castelo Branco	*A Caveira da Mártir*	1876
383	Camilo Castelo Branco	*Eusébio Macário*	1879
384	Camilo Castelo Branco	*A Corja*	1880
387	Camilo Castelo Branco	*Vulcões de Lama*	1886
389	Cândido de Figueiredo	*Lisboa no Ano Três Mil*	1892
390	Carlos Malheiro Dias	*A Mulata*	1896
392	Carlos Pinto de Almeida	*A conquista de Lisboa*	1866
400	Claudia de Campos	*Ele*	1899
407	Coelho Neto	*Miragem*	1895
409	Coelho Neto	*O morto*	1898
411	Coelho Neto	*A conquista*	1899
416	Coelho Neto	*Esfinge*	1908
418	Coelho Neto	*Rei negro*	1914
419	Coelho Neto	*A capital federal*	1915
422	Coelho Neto	*Mano*	1924
429	Conde de Ficalho	*Mais Uma*	1888
458	José Maria Eça de Queirós	*O Crime do Padre Amaro*	1875
459	José Maria Eça de Queirós	*A Tragédia da Rua das Flores*	1878
460	José Maria Eça de Queirós	*O Primo Basílio*	1878
461	José Maria Eça de Queirós	*O Mandarim*	1880
462	José Maria Eça de Queirós	*A Relíquia*	1887
463	José Maria Eça de Queirós	*Os Maias*	1888
465	José Maria Eça de Queirós	*As Minas de Salomão*	1891
469	José Maria Eça de Queirós	*Fradique Mendes*	1900
470	José Maria Eça de Queirós	*A Cidade e as Serras*	1901
478	José Maria Eça de Queirós	*A Capital*	1925
481	José Maria Eça de Queirós	*Alves e Companhia*	1925
482	José Maria Eça de Queirós	*O Conde d Abranhos*	1925
484	José Maria Eça de Queirós	*Cartas Inéditas de Fradique Mendes*	1929
485	Eduardo de Noronha	*O agonizar de uma dinastia*	1908
491	Francisco d'Athayde Machado de Faria e Maia	*Vencido*	1914
494	António Feliciano de Castilho	*A chave do enigma*	1861
495	José Maria Ferreira de Castro	*A selva*	1930
504	Faustino da Fonseca e Joaquim Leitão	*Os filhos de Inês de Castro*	1905

(Continua)

Id	Autor	Título	Ano
532	Francisco Barros Lobo	*O Tio João Gil*	1906
535	Francisco da Fonseca Benevides	*No tempo dos franceses*	1908
548	Franklin Távora	*O Matuto*	1878
549	Franklin Távora	*O Sacrifício*	1879
583	Graça Aranha	*Canaã*	1902
605	Harry Laus	*Os papéis do coronel*	1995
613	Inácio Pizarro de Morais Sarmento	*O Engeitado*	1846
614	Inglês de Sousa	*O missionário*	1891
617	Jayme de Magalhães Lima	*Transviado*	1899
621	Joaquim Manuel de Macedo	*O moço louro*	1845
622	Joaquim Manuel de Macedo	*Os Dois Amores*	1848
624	Joaquim Manuel de Macedo	*A luneta mágica*	1869
625	Joaquim Manuel de Macedo	*As Vítimas-Algozes*	1869
626	Joaquim Manuel de Macedo	*As Mulheres de Mantilha*	1870
641	João José Grave	*A morte vence*	1916
654	José de Alencar	*Cinco Minutos*	1856
656	José de Alencar	*O Guarani*	1857
657	José de Alencar	*As minas de prata*	1862
659	José de Alencar	*Diva*	1864
661	José de Alencar	*A pata da gazela*	1870
662	José de Alencar	*O gaúcho*	1870
663	José de Alencar	*Sonhos dOuro*	1872
664	José de Alencar	*A alma de Lázaro*	1873
666	José de Alencar	*O Garatuja*	1873
667	José de Alencar	*Ubijarara*	1874
668	José de Alencar	*O sertanejo*	1875
669	José de Alencar	*Senhora*	1875
670	José de Alencar	*Encarnação*	1877
672	José do Patrocínio	*Mota Coqueiro*	1877
675	José Régio	*O príncipe com orelhas de burro*	1942
693	José da Silva Mendes Leal	*Infaustas Aventuras de Mestre Marçal Estouro: Vítima duma paixão*	1862
694	Júlio César Machado	*A vida em Lisboa*	1858
705	Júlio Dinis	*A Morgadinha dos Canaviais*	1868
707	Júlio Dinis	*As apreensões de uma mãe*	1870
708	Júlio Dinis	*Justiça de Sua Majestade*	1870
710	Júlio Dinis	*Os novelos da tia Filomela*	1870
711	Júlio Dinis	*Uma flor de entre o gelo*	1870
713	Júlio Dinis	*Os Fidalgos da Casa Mourisca*	1871
717	Julia Lopes de Almeida	*A falência*	1901
719	Julia Lopes de Almeida	*A Intrusa*	1905
720	Júlio Lourenço Pinto	*Margarida*	1879
723	Lima Barreto	*O subterrâneo do morro do castelo*	1905
724	Lima Barreto	*Recordações do escrivão Isaías Caminha*	1909
729	Lima Barreto	*Clara dos anjos*	1948
733	Artur Lobo de Ávila	*A descoberta e conquista da Índia pelos portugueses: romance historico*	1898
735	Lopo de Sousa	*Herança de lágrimas*	1871
737	Luciano Cordeiro	*A senhora duquesa*	1889
738	Lucas José de Alvarenga	*Statira, e Zoroastes*	1826
747	Luís Filipe Silva	*O futuro à janela*	1991
750	Luís Magalhães	*O Brasileiro Soares*	1886

(Continua)

Id	Autor	Título	Ano
781	Joaquim Maria Machado de Assis	*Os trabalhadores do mar*	1866
800	Joaquim Maria Machado de Assis	*Oliver Twist*	1870
810	Joaquim Maria Machado de Assis	*Ressurreição*	1872
825	Joaquim Maria Machado de Assis	*A Mão e a Luva*	1874
841	Joaquim Maria Machado de Assis	*Helena*	1876
858	Joaquim Maria Machado de Assis	*Iaiá Garcia*	1878
867	Joaquim Maria Machado de Assis	*Memórias póstumas de Brás Cubas*	1881
907	Joaquim Maria Machado de Assis	*Casa velha*	1885
965	Joaquim Maria Machado de Assis	*Esaú e Jacó*	1904
977	Joaquim Maria Machado de Assis	*Memorial de Aires*	1908
982	S. de Magalhães Lima	*A senhora viscondessa*	1875
987	Manoel da Cruz Pereira Coutinho	*Elvenda, ou Conquista de Coimbra por Fernando Magno*	1858
995	Manuel de Oliveira Paiva	*A afilhada*	1899
1010	Marcelino Mesquita	*Os quatro reis impostores*	1908
1011	Maria O'Neill	*A Marquesa de Vale Negro*	1914
1013	Mário de Sá-Carneiro	*Loucura...*	1912
1014	Mário de Sá-Carneiro	*A Confissão de Lúcio*	1913
1018	Matilde Isabel de Santana e Vasconcelos Moniz Bettencourt	*O soldado de Aljubarrota*	1857
1019	João Baptista de Mattos Moreira	*Tempestades do Coração*	1867
1020	Maurícia C. de Figueiredo	*O exilado*	1900
1023	Maria Benedicta Mousinho de Albuquerque Pinho	*Marina: romance passional*	1912
1024	Melo de Matos	*Lisboa no ano 2000*	1906
1032	Miguel J. T. Mascarenhas	*Um conto português: episódio da guerra civil: a Maria da Fonte*	1873
1039	J.P. Oliveira Martins	*Febo Moniz*	1867
1040	Oliveira Mascarenhas	*O frade arrábido: romance histórico do século XVIII*	1881
1043	Othon Gama d'Eça	*Vindita braba*	1923
1063	Paulo Setúbal	*A Marquesa de Santos*	1925
1064	Paulo Setúbal	*O Príncipe de Nassau*	1925
1065	Paulo Setúbal	*Os irmãos Leme*	1933
1070	A.J. Pereira Varela	*Os miseráveis da aristocracia*	1864
1074	Manuel Pinheiro Chagas	*Astúcias de namorada*	1873
1077	Manuel Pinheiro Chagas	*A Lenda da Meia-Noite*	1906
1078	Pedro José Supico de Morais	*O mundo no ano 3000*	1895
1079	Policarpo da Silva	*O piolho viajante: Viagens em mil e uma carapuças*	1802
1085	Raúl Brandão	*A Farsa*	1903
1087	Raúl Brandão	*Os Pobres*	1906

(Continua)

Id	Autor	Título	Ano
1088	Raúl Brandão	*Húmus*	1919
1099	Raúl Brandão	*O Pobre de Pedir*	1931
1101	Raul Pompeia	*As jóias da Coroa*	1882
1102	Raul Pompeia	*Uma tragédia no Amazonas*	1882
1136	Tomaz de Melo	*O Conde de S. Luís*	1874
1143	Virgínia de Castro e Almeida	*Decameron*	1916
1144	Virgínia de Castro e Almeida	*Inocente*	1916
1145	Virgínia de Castro e Almeida	*O Solar dos Pavões*	1916
1146	Virgínia de Castro e Almeida	*A história de Dona Redonda e da sua gente*	1941
1151	Virgílio Várzea	*Rose-Castle*	1893
1152	Virgílio Várzea	*A noiva do paladino*	1901
1154	Virgílio Várzea	*O brigue flibusteiro*	1904
1155	Visconde de Taunay	*Inocência*	1872
1159	Visconde de Villa-Moura	*Nova Sapho: Tragedia Extranha*	1911

,

Extração de Informação sobre Personagens Literários em Português

Extraction of Literary Character Information in Portuguese

Eckhard Bick
University of Southern Denmark

Abstract

Este capítulo descreve o PALAVRAS-DIP, um sistema para a identificação automática de personagens e dos seus perfis sociais na literatura portuguesa e brasileira. O sistema foi concebido como um módulo adicional para um analisador morfossintáctico e semântico. Etiquetamos as entidades nomeadas (NE) humanas para profissão e posição social, e usamos as etiquetas relacionais do formalismo Constraint Grammar (Gramática de Restrições, CG) para estabelecer co-referências (por exemplo, anáfora de pronomes, verbos com sujeito zero) assim como relações familiares entre as personagens. A anotação de base resultante permite a extração de redes de personagens. O programa de extração reconhece e agrupa as variantes de nomes de personagens e distingue entre nomes que têm função narrativa e nomes contextuais de referência cultural. O desenvolvimento do sistema foi motivado pelo DIP, uma avaliação conjunta sobre 100 romances históricos, evento em que uma versão protótipo do sistema obteve medidas F razoáveis para as tarefas de identificação de personagens (63,4%) e de unificação/co-identificação de nomes (68,1%), mas teve problemas com as relações familiares (15,5%).

Keywords

leitura distante, extração de informação, reconhecimento de entidades nomeadas, constraint grammar, resolução de anáforas

Abstract

This chapter describes PALAVRAS-DIP, a system for the automatic identification of characters and their social profiles in Portuguese and Brazilian literature. The system has been designed as an add-on module for a morphosyntactic and semantic parser. We tag human named entities (NE) for profession and social position, and use Constraint Grammar (CG relational tags to keep track of co-reference (e.g. pronoun anaphora, zero-subject verbs) and family reations between the characters. The resulting base annotation allows the extraction of character networks. The extraction program recognizes and bundles character name variants and distinguishes between names with a narrative function and simple cultural references. System development was motivated by DIP, a shared-task evaluation on 100 historical novels, where a prototype version achieved reasonable F-scores for character identification (63.4%) and alias resolution (68.1%), but underperformed for family relations (15.5%).

Keywords

Distant reading, IE, NER, Constraint Grammar, anaphora resolution

1. Introduction

At first glance, the task of automatically extracting characters and their social relations from literature seems like an extension of named entity recognition (NER). However, while classical NER does identify candidate tokens for characters, the method needs to be adapted to the literary genre (e.g. Bornet & Kaplan (2017), for French), and as characters may be identified by many different variants of their name (e.g. first or second name, with or without a honorific, title, middle name, nickname etc.), name instances need to be unified. Also, the basic NER tag of "person" does not make the distinction between narrative, "functional" names and cultural reference names referring to gods, poets and historically important people. Finally, characters form social networks that go beyond simple recognition. In order to extract these networks, characters' social attributes and their mutual relations must be extracted too — information that may change chronologically throughout a book. In a wider context early literary analysis, other narrative character information may added, such their actions, plot event participation and affect states (Goyal et al., 2010). Much of the previous work in the field has been done on English (e.g. Labatut & Bost (2019); Valls-Vargas et al. (2014), often using classical, older texts. The work described here has a similar focus on classical literature, but addresses Portuguese, an under-represented language where previous research had targeted the

DOI: 10.21814/lm.15.1.397

less complex topic of children's stories Mamede & Chaleira (2004). Our research was carried out in the context of the DIP shared task (Desafio de identificação de personagens), a Portuguese-language character identification challenge organized by Linguateca[1], NuPILL, UEMA and UiO (Santos et al., 2022), and described in detail in this volume. The system uses a morphosyntactic and semantic parser, PALAVRAS (Bick, 2014) to provide a grammatical base annotation and named entity (NE) mark-up. The new DIP extension unifies name instances, verifies name gender at the text level and adds tags for title, profession or social standing for those names that it deems characters rather than cultural references. It also adds a new type of relational tags for family relations and extends PALAVRAS' experimental co-reference annotation using longer spans, text variables and explicit referent tags. The DIP extension is a rule-based system based on the same formalism as PALAVRAS itself, Constraint Grammar (CG3). Specifically, we use the CG3 variant (Bick & Didriksen, 2014), which supports the use of long-distance relational tags, as well as the capture, use and unification of both tag-level and text-level variables.

2. Grammatical base annotation

PALAVRAS' annotation scheme comprises information from various linguistic levels, including lemma, morphology, syntactic function and dependency structure. At the semantic level, in addition to the afore-mentioned NER, the parser provides a (disambiguated) noun ontology, as well as framenet structures and semantic roles (Bick, 2022). This linguistically high level of pre-annotation is an important prerequisite for the extraction of character networks, as pointed out by Chaturvedi et al. (2017), who use linguistic information such as frame semantics to complement the simpler bag-of-words approach in their feature vectors when tracking character relationships.[2] Both PALAVRAS itself and the add-on DIP module are rule-based systems. This makes for great transparency and allows fairly straightforward error correction and genre adaptation, but as a rule-based set-up cannot simply copy its tokenization and category distinction from a body of training data, some adaptations have to be made to meet a given annotation standard — in this case the one dictated by the DIP conventions. Of course, the problem is limited, as it does not concern internal tagging, but only what is visible in the final output. Specifically, changes had to be made regarding the inclusion (or non-inclusion) of honorifics, titles, family terms and title-like profession terms in names. As a rule of thumb, title-like words were included in character names, if they can co-occur with a name in the vocative.

- part-of-name: *Com(p)adre, Dama, Dom, Don(a), Doutor, Dr., Frau, Fräulein, Frei, Herr, Lady, Lord, Madame, Mademoiselle, Maestro, Mano, Miss, Mister, Monsenhor, Monseor, Monsior, Monsenhor, Monsieur, Mlle, Nhá, Nhô, Padre, Prima, Primo, Prof(a)., Senhor(a), Senhorita, Sô, S(n)r., S(n)r.ª, Tia, Tio*

- not part-of-name: *Conde(sa), Duque(sa), Imperador(a), Príncipe, Rei, Rainha, Vinsconde(sa), coronel, juiz, mãe, pai, neto, avó etc.*

Another necessary adaption concerned the textual input itself, as most DIP texts were historical in nature and contained lexical and orthographial variation not seen in modern texts, often compounded by what might be photo reproduction, "de-pdf'ing" and OCR posing problems. The resulting unrecognizable wordforms pose a problem to both the base parser and later tasks — inparticular, name unification. As PALAVRAS has been used on both historical data, transcribed speech and social network input, it contains some non-standard lexical extensions, and it does perform a certain amount of normalization itself, including some automatic spell checking useful for the task at hand. In the output, both the original and the normalized wordforms are provided, but the lemma as well as morphosyntactic and semantic annotation will be based on the normalized form. The parser does not normalize names, though, as this will produce many false positive "corrections." Therefore, the DIP module implements its own, "relaxed" name unification method, where a Levenshtein (spelling) distance[3] of 1 is tolerated for names between 4 and 9 letters (e.g. *Luíza, Luiza, Luísa, Luisa* or *Hamlet, Hamleto*), and a Levenshtein distance of 1 or 2 for longer words (e.g. *Christovam, Chrystovam, Christovão*), provided there also is a gender match (i.e. not *Francisco, Francisca*).

[1] https://www.linguateca.pt/DIP/

[2] In this work, relationships are seen as (evolving) latent states. The family relations addressed in DIP can be seen as a more stable subset of overall relations.

[3] meaning 1 exchanged, added or removed letter

3. Co-reference resolution

Though a name or its variant may occur many times in a given literary text, crucial information about the character, such as profession, marital status and descent (parents' names), may be provided explicitly only once, and henceforth assumed to be known to the reader. The information may also be implicit-only, for instance providing a work place or typical tool instead of a profession, or hinting at family relations through the form of address in dialogue. In any case, all (or as many as possible) mentions of a given character need to be kept track of in order to make such connections where and when they occur. The importance of this task is illustrated by the fact that most character occurrences are not names, but pronouns[4]. For our Portuguese data, zero-subject finite verbs with pronoun ellipsis (50.3% in the first 100 books of the DIP collection) are more frequent than finite verbs with personal pronoun subjects (9.7%) and need to be lumped with the latter. Together, pronoun references and zero-subject verbs account for over half (52.6%) of all established character references in the data when excluding reflexives, and 57.5% including np (noun phrase) mentions. The percentage of indirect mentions rises to 64.8% when also including +HUM personal pronouns without an established character reference — the metrics used for English literature by Bamman et al. (2014), who also report a high prevalence (74%) of pronominal mentions. Finally, we might add participle and infinitive clauses, which usually make do with an implicit subject, linked to a preceding finite verb. In any case, given the high prevalence of pronominal and zero-subject character references, it is of prime importance to resolve anaphora relations.

To this end, we expanded a set of existing, experimental anaphora rules in PALAVRAS' Constraint Grammar pipe with additional CG rules, improving its coverage, scope and accuracy and adapting it to the task at hand. Here, we use the RELATIONS[5] operator to establish referent links between pronouns and underspecified noun phrases and a target referent, optimally a named entity (NE) of the PERSON category. The equivalent solution for subject-less verbs establishes links to a preceding surface subject, or — as a fallback — another subject-elliptic verb. In both cases, name targets will also be mapped, as `<REF:name>` tags, on the anaphorical element itself. This is useful for "promoting" the antecedent information, if link targets are themselves anaphorical (e.g. chains of pronouns or subject-elliptical verbs), in which case the ultimate name referent may be outside the rolling CG focus window. For syntax, this window would be just one sentence at a time, but for co-reference resolution, we opted for a ±6 sentences as a compromise between reach and recall on the one hand, and precision on the other. The basic co-reference rule algorithm is based on recency and similarity of antecedents, as suggested by Elson et al. (2010)), weighting features such as +HUM, top-level subject-hood, definiteness and topic or focus function[6]. As most personal pronouns in Portuguese are marked for gender and number, and verbs are inflected for (subject) person and number, this morphological information can be used to impose tag conditions on the name or np antecedent. This is true not only of clause-level pronouns, but also of possesives, where a reference link will allow us correctly assing family relations mentioned in the possessive's head (e.g. his father/mother/son/daughter). The relative pronoun que presents a special case, as it is underspecified with regard to gender and number. However, dependency syntax makes it relatively easy to recover an antecedent that can then be associated with information provided in the dependent relative clause (e.g. Pedro, que se casou com Júlia em tenra idade [Peter, who had married Julia at an early age] or seu amigo, que trabalhava como porteiro no turno da noite [his friend, who worked night shifts as a porter]).

For pronouns in a +HUM semantic frame slot, or for subject-less verbs with a human verb frame, the co-reference antecedent should be a human name. But if none can be found, without interfering blocking material (e.g. a non-human top-level subject), in the context window, or if there is more than one candidate, it may be necessary (or safest) to settle for an intermediate antecedents, even if it is a pronoun or zero-subject verb itself. We call such underspecified references "stepping stones". If the stepping stone itself can be assigned a reference link to a name antecedent, this link can later be recovered by a meta rule, and raised to the originnal pronoun that has been "stranded" with a stepping-stone reference. In real-life narrative text, there will

[4] Sometimes, +HUM noun phrases, e.g. o vigário [the vicar], are also used to refer to names, but with a much lower frequency, according to PALAVRAS' anaphora annotation.

[5] The RELATIONS operator allows the assignment of bi-directionally named, non-unique relations between tokens.

[6] Bamman et al. (2014) used gender and linear word distance. The equivalent to the latter, in our CG rules, are so-called barriers (e.g. non-matching top-level subjects or paragraph breaks), blocking further search left.

often be multiple stepping stones, for instance in a chain of action statements with pronominal subjects all referring to the same human agent introduced in the beginning of a paragraph. In the example, a subject complement (@SC, id-6), médico (physician), has been assigned an attribute relation (R:n-attr:5) to a pronoun subject (id-5). This pronoun is itself linked to a top-level subject antecedent, a character named XXX (id-1), four sentences to the left, through three stepping stones — first a subject-less finite verb (id-4), then two pronouns (id-2 and id-3). The verb carries an elliptic-subject relation (R:e-subj:3) to the closest of the pronouns (id-3). All pronouns carry referent relations (R:ref:id) linking them to either a stepping stone (R:ref:4) or the first, full subject (R:ref:1). Ultimately, this links the profession information ('doctor' id-6) to the full subject (id-1), even across text (. . .) spanning multiple sentences. Secondary annotation rules can propagate these links (e.g. R:ref:3 on id-5 or R:n-attr:1 on id-6) and assign explicit attribute tags to names, here the profession tag NA:Hprof/médico (on *XXX*, id-1).

- "XXX" main referent: top level @subject (PROP, +HUM, np-def)
  → R:be:6
  → <NA:Hprof/médico>
  . . .
- subject pronoun <REF:XXX> **R:ref:1**
  . . .
- subject pronoun <REF:XXX> **R:ref1** R:subj:4
  . . .
- subject-less VFIN R:e-subj:3
  . . .
- subject pronoun **R:ref:4** R:be:6
  → R:ref:3 → R:ref:1
  → <REF:XXX>

- "médico" <Hprof> @SC §ATR <R:n-attr:5>
  → R:n-attr:1

4. Quoted speech

In many literary works, an important part of character-related information is to be found in quoted speech. The density of quotes is quite text dependent. Thus, for English, Elson et al. (2010) found a spread of 19-71% for text included in quotes. For the DIP data (100 non-pdf texts), the average was 30.5% for direct speech. This

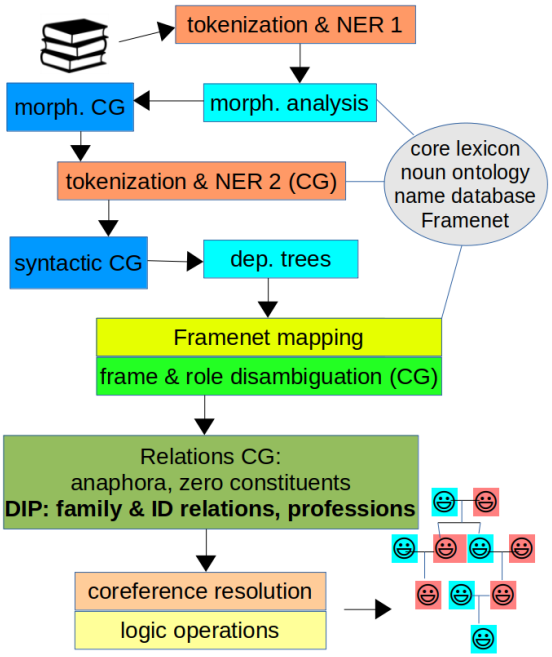

Figure 1: Relation chains

fact is relevant for the parser, as direct speech often manifests as an independent utterance without an externally linked dependency head (which would be typical of indirect speech), making it more difficult to keep track of who is who. To tackle this problem, and to facilitate the identification of speaker and addressee in dialogue turn-taking, we perform a quote mark-up as an early annotation step. The relevant tags are <quote-edge> (quote opener), <quote-end> and <quote-ana> (quote continuation after a quoting verb). In the literary data, dashes were used rather than quotation marks, leaving quote closures unmarked, unless they were followed by a quoting verb. To compensate for the missing dependency relations, we also mark the syntactic top node(s) in a quoted speech (<quote>), as well as the quoting verbs (<v-quote>), even if occurring in a separate sentence. In-quote name mentions are distinguished from body text name mentions by tagging the forme <quo>, and the latter <nquo>. In its newest version, our grammar also keeps track of turn-taking, marking alternate turns as <turn-1> and <turn-2>, and setting a speaker variable for each turn and <quote> mark.

This annotation not only makes it possible to extract and correctly name-link information from inside direct speech, but also helps establishing what is a character and what is not. Thus, narrative characters are more likely to occur in direct speech, or to produce it. A direct syntactic clue are vocatives (surface addressees), which almost always refer to characters. Co-reference rules will

link np vocatives (typically family terms or titles) to the speaker (quoting subject) in an adjacent turn, or — conversely — link named vocatives to noun speakers (or the antecedents of speaker pronouns). In both cases, the link can be used to infer attributive information from an np to a name. Even without extracting further information, speaker-discourse links can be used to establish relations between turn-taking characters based on the quantity of one-on-one dialogue. Thus, Elson et al. (2010) define and analyze character relationships as networks of social conversations. Similarly, dialogue relations could be used alongside family relations and cooccurrence-strength to the extraction and visualisation of social networks (Section 6). Linking discourse turns to literary characters is not a big discipline in NLP, but there is prior work on Portuguese, who present a rule-based system with trained decision trees for children's stories, reporting a 89% success rate for discourse separation, while recall and precision for speaker character identification were 10.6% and 65.7%, respectively. Our own speech annotation method was experimental and incomplete at the time of the shared task, but now recognizes 98.1% of all direct speech utterances, with an overall F-score for speaker identification of 92.0% (Bick, 2023).

5. Character annotation

The second layer of our CG annotation rules exploits existing relational links of the base annotation and the co-reference module, making relevant implicit information explicit on name tokens that refer to characters, or — as an intermediate step — on +HUM nouns or pronouns that are referent-linked to such a character token. First of all, this means mapping explicit name referent tags (red in Figure 1), e.g. `<REF:Pedro=da=Silva>`, but it also involves tags for the characters' social attributes and relations.

5.1. Social attributes

Social attributes are harvested from profession and title nouns (green annotation in Figure 1), and tagged on name tokens with an 'NA' (noun attribute) prefix, e.g. `<NA:Hprof:ministro>`. The process can make use of existing framenet information in PALAVRAS' base annotation, such as R:attr relational tags, the syntactic functions of subject and object complement, apposition or noun predicate, as well as the semantic roles of §ATR (attribute), with a name head, and §ID (identity, with a name dependent. These syntactic or semantic links are exploited to relate names to profession nouns (e.g. 'carpenter') and titles (e.g. 'chairman'), or to family nouns (e.g. 'father' or 'daughter'). For the latter, additional rules are needed to identify the argument of a given family noun, which may be "hidden" in a postnominal pp (e.g. 'X, mother of Y). Of course, as discussed in the co-reference Section, the immediate framenet or dependency links may not lead to a name, but rather to a noun or pronoun, in which case the link needs to be propagated to a name antecedent, following anaphora links and possibly bypassing one or more of the aforementioned "stepping stones". In addition to existing attributive relations, rules can also make use of a variety of specific clues and semantic reasoning, exploiting, for instance, profession-specific verbs (such as teaching for teachers), or nouns denoting profession-related tasks, products or institutions.

5.2. Family relations

The annotation of family relations is more complex than that of social attributes, as it involves two targets rather than one. We want to know not just that somebody is a daughter, but also whose daughter. Family relations are marked at both ends of a relation and may either be symmetrical (siblings, spouses) or asymmetrical (parent–child). A CG rule establishing such a relation, will add both tags at the same time, e.g. R:parent:id at one end and R:child:id at the other. Especially if one of the two names is "out of reach" (outside the window of analysis, or not mentioned as a proper noun), the name in question may also be recovered from a co-reference link or tag (including stepping stones), and the information tagged on the other name with an RI prefix, e.g. `<RI:parent_of:Maria>`. With few exceptions (heuristic proper nouns without morphological clues or attributes), gender is already tagged in PALAVRAS input. Therefore, family relation tags can be kept gender-neutral, reducing tag set size and grammar complexity. All in all, the grammar covers eight basic family relations. Four of these are symmetrical (parent, sibling, spouse, cousin, "gbfriend" [girl friend or boy friend]), four are paired/asymmetrical. Apart from the parent–child pair, the latter group includes the portmanteau categories of "auncle" [aunt or uncle] and "nephie" [nephew or niece]). Where relevant, the set of relations can be expanded with prefixes for 'great-' (g-), 'great-great-' (gg-), 'in-law' (i-) and 'god' (god-), e.g 'gparent', 'isibling' or 'godchild'. Finally, there is a non-directional relation 'widow' and one non-family relation, 'friend'. The various combinations correspond to about 40 Portuguese words.

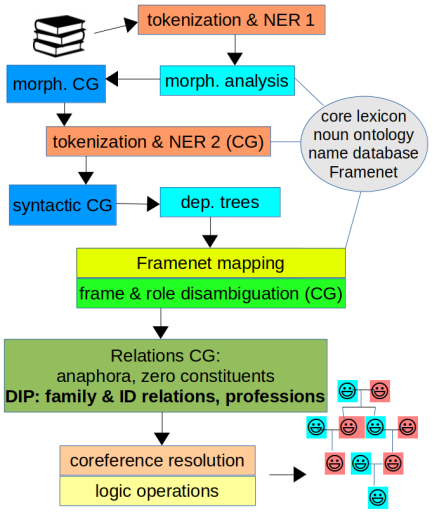

Figure 2: System architecture

5.3. Non-characters

Because PALAVRAS does not itself distinguish between characters and "cultural" name references (saints, poets, emperors etc.), because neither can be covered by closed lists, and because some names may function as either, we also need rules to help making this distinction. These rules exploit morphological, semantic and contextual clues[7] to assign a `<cult>` (cultural) or `<noncult>` (fictional) tag. For instance, names with an affection suffix (-inh[oa], -zinh[oa]) or a family relation are deemed to be characters, as are names that are part of speaking or movement frames. Conversely, surname-only names[8], or names in a royal or religious context will be tagged as `<cult>`. The final cast extraction also employs text-level statistics as a further means of distinction.

[7]With only a "proper noun" tag as input, i.e. from a parser without the NE category of PERSON, the distinction between character and non-character would have to subsume the ±PERSON distinction. Paul & Das (2017), for instance, also using grammatical context features, describes a neural network-version of such a classifier for an English version of Mahabharata, achieving F-scores of 76% and 88% for proper nouns and proper noun-containing NP's, respectively.

[8]This surprisingly safe heuristics was originally motivated by inspection of DIP's first example novels, but can be corroborated on the rest of the collection: typically, only famous people (scientists, poets etc.) are referred to by surname alone. Character names exhibit more variation and are normally introduced at least once with a more complete name. Also, character surnames are usually used with honorifics. First names, on the other hand, do occur on their own, especially for children and servants.

6. Cast extraction

The third module in our system pipeline, after primary (general) and secondary (task-driven annotation), extracts and structures the (now) explicit character information encoded in the annotation (Figure 2). The extractor first builds a data structure for all name IDs and name relations, storing, for each person name token:

- gender labels
- social attribute labels (NA tags)
- relative_of:name tags (RI tags)

As a fall back, in the absence of an explicit RI tag, the extractor will follow family relation links and retrieve the target name, either directly (for name lemmas), or from a `<REF:name>` tag, or by following stepping stone links.[9] If this process leads to circular ID references or self-relations,[10] the extractor will ignore the information in question.

6.1. Character name unification

For each name in the data structure, the extractor loops through all other names and decides which are aliases ("synonyms") of the same name, creating named synsets based on:

- maximum number of shared core name elements (first names and surnames)
- unification of attributes for gender and social "role" (profession or family role)
- ratio of occurrence inside/outside quotes

For this process, titles/honorifics (e.g. Sra — 'Mrs') and morphological variation (-inho — 'dear') are ignored, and titles in isolation are not regarded as names. Synset names should be unambiguous, and will therefore typically consist of a multi-part version of a given name. Theoretically, isolated instances of ambiguous first names can then be attributed to different synsets based on social role.

In conjunction with coreference resolution, the cast extractor weeds out `<cult>`-marked names,[11] unless they can be assigned to an ex-

[9]Since the extractor has built a data structure for the whole text, it is not limited by the ±6 sentences analysis window.

[10]This is rare, but can theoretically happen due do errors in the CG annotation, in particular dependency errors or frame link errors.

[11]In the case of conflict, i.e. if there are both `<cult>` and `<noncult>` tags for a given name, `<noncult>` wins with a simple majority, while `<cult>` is valid only if it is tagged on more than half of the occurrences of that name.

isting name synset. 1-part names that are rare in text in both absolute and relative terms, and that are not part of a synset, are also discarded — unless they have been specifically tagged as <noncult> (cp. Section 3). The same goes for multi-part names if they consist only of honorifics and/or start with an article (e.g. o=Santo=Padre — 'the Holy Father'). In unclear cases, occurrence in direct speech is regarded as an indicator of characterhood. A first-person narrator is regarded as a special case character. Thus, if there are 1st-person verbs or pronouns, outside of quotes, these are flagged and synset-linked to possible 3rd-person, named mentions of the narrator. In other words, a 1st-person narrator will be regarded as part of the cast.

6.2. Logic operations

In a second stage, the cast extractor expands the family tree using logic:

- Propagation: If X is a child of Y, and Z is a parent of Y, then X is a grandchild of Z;
- Symmetry: If X is a sibling of Y, then Y is a sibling of X

Also, professions (or other, in-context relatively unique, noun attributes) may be used for unification: If X is a doctor, and Y the child of a doctor, then Y is a child of X. Because human NPs, just like names, may be given relation-tags, this method, albeit a bit risky, may even be applied where the profession-providing NP has not been name-resolved: If X is the spouse of a (nameless) doctor, and Y the child of (said nameless) doctor, then Y is a child of X.

6.3. Output formats

The native output of the cast extractor is an alphabetical list of name synsets with their members and gender, followed by profession attributes and a list of family relations. A condensed format with numbered name synsets (nns) is available for evaluation, consisting of two .csv files for each text, one for character synsets, gender and profession, one for family relations.

- characters.csv:
  `nns,syn-1|syn-2|...,gender,prof-1|prof-2|...`
- relation.csv:
  `nns-1,relation,nns-2`

Task	Average F-score	F-score spread
character identif.	63.4	40–80
name unification	68.1	(20–) 40–90
gender	89.5	60-100
prof./occupation	24.6	(0–) 10–55
family relations	15.5	0–60

Table 1: System performance (shared task)

7. Evaluation

As already explained elsewhere in this volume, the DIP shared task addressed Portuguese and Brazilian literary works, mostly historical novels from the 19th and early 20th century, and comprised five subtasks: (a) character identification, (b) co-reference resolution, (c) character gender, (d) profession/occupation or other position in society, (e) family relations. Two novels were provided as examples by the organizers, with manually extracted character information. The test run had to be performed in 48 hours on a collection of 100 books (mostly older novels), 20 of which had manually extracted gold casts for the evaluation. The literary period constraint, motivated by public domain availability, caused some annotation problems for the base parser, as texts contained a certain amount of orthographical variation (e.g. *cavallo/cavalo* — 'horse', *sabbados/sabados* - 'Saturdays') not found in modern Portuguese, as well as errors introduced by OCR scanning, or combinations of both (e.g. *offerecimento* — 'offering'). Our cast extractor was the only participating system that solved the task for the provided data (historical literature) and within the given time frame. It achieved reasonable F-scores for character identification (63.4%), co-reference resolution (68.1%) and gender assignment (89.5%), but did not perform well for professions (F=24.6%) and family relations (F=15.5%).

Results differed a great deal between works, not least for the difficult subtasks (d) and (e). Thus, the best books had F-scores of 80-90% for the identification subtasks (a) and (b), 100% for gender, and 50-60% for the social information subtasks (d) and (e). On the flipside, several books scored 0 for relations, as did one for professions. Disregarding one outlier, identification was more robust, with the lowest-scoring books at around 40% for (a) and (b). The very pronounced text dependence of the task was also noted by Dekker et al. (2019), who evaluated the performance of different (English) NER tools in the co-occurrence-based extraction of social net-

works. Here, systems worked well for e.g. *Huckleberry Finn* and *Game of Thrones*, with the best ones achieving F-scores of 80-90% for the isolated NER task of person recognition, while many had single-digit results for *Brave New World*. Also, all systems performed worse for classical novels and better for modern novels. For the full task of character detection, Vala et al. (2015), when evaluating their 8-stage system on three different English literary data sets, also report a substantial spread in accuracy (F-scores of 44.8, 54.0 and 75.8), similar to our own spread found for Portuguese (F=40-80). When interpreting our results, it should be born in mind that character identification is more than named entity recognition. Thus, errors were caused mainly by the (sometimes unclear) distinction between "real" characters and cultural background names, not by the underlying NER, which was much more robust, as it conflates the two categories into "person". Second, the unification of names with each other (b) and with pronouns and non-name np's (both here only evaluated indirectly) makes for additional complexity compared to the underlying NER task. Also, when developing the system, the distinction between titles and occupation or social position was not always entirely clear from the examples. Therefore, some errors in (d) resulted from fuzzy definitions and not the rules or algorithm of the software. Finally, historical spelling variation and non-standard uppercasing created false name candidates caused by POS errors.

8. Network analyses

Automatic cast extraction provides a quantitative-comparative[12] angle to literary analysis difficult to achieve by other means. Thus, it is possible to compare the works of different authors or across different periods og genres, focusing on features such as cast complexity, gender distribution, and societal representativeness. The same data can also be used for the vizualisation of character networks, where quantitative network parameters allow e.g. the distinction between central and peripheral characters. The example in figure 3 shows a close-up of one such network, generated with Cytoscape[13], for the Brazilian novel "Quincas Borba", by Machado de Assis (published 1891).

The necessary .csv tables were created by exporting standardized, unique character id names (u-names) as network nodes and as-is "surface names" (s-names) as their attributes. Family relations were exported as directed arcs ("edges") between source (SN) and target (TN) node (e.g. SN child_of TN). In order to further populate the network, an un-named relation was assigned to a given pair of character tokens, if their in-text difference was less than 3 "semantic" tokens (defined as carrying a semantic role or frame tag). Relation frequencies were then used as a strength attribute for the relation in question. We end up with 6 columns in the *Cytoscape* .csv table: SN u-name, SN s-name, relation, relation strength, TN u-name, TN s-name. The table allows the computation of various network parameters, such as edge count, stress, closeness and clustering, which can be used to evaluate and describe the narrative importance and connectedness of the characters. Figure 3 shows a close-up of the central portion of one possible vizualisation of the character network, using edge-weighted spring-embedded layout, with family relation edges in red, and unnamed relations in gray. Though based on a mathematical model, the graphic is immediately interpretable by a human reader, singling out a handful of main characters and providing an overview of their mutual connections.

9. Conclusion

We have discussed the implementation of a Constraint Grammar-based system for the extraction of character information from Portuguese text. The method harnesses existing mark-up from a morphosyntactic and semantic parser to assign relational tagging for name co-reference and family relations, as well as social attributes. Character names are unified and distinguished from non-characters using clues like dialogue participation and network centrality. In the DIP shared task, a first version of our system achieved reasonable results for character identification and unification. However, information about family relations and social position proved to be much more difficult to extract, as it is often provided through indirect clues, or through attribution in direct speech subject to long turn-taking sequences. Future versions of the cast extractor should address these problems, for instance by developing a full-fledged speaker- and adressee attribution module and by the use of text-level coreference variables.

[12] Quantitative analysis is particularly robust in the face of a certain error rate, since distributional patterns are likely to be visible despite errors in individual characters.

[13] https://cytoscape.org/

Figure 3: Character network (Cytoscape)

Acknowledgments

We are grateful to the DIP team at Linguateca, NuPILL, UEMA and UiO for preparing and organizing the shared task, and appreciate the work that has gone into the manual compilation of gold-standard evaluation data.

References

Bamman, David, Ted Underwood & Noah A. Smith. 2014. A bayesian mixed effects model of literary character. In *52nd Annual Meeting of the Association for Computational Linguistics*, 370–379. doi 10.3115/v1/p14-1035.

Bick, Eckhard. 2014. PALAVRAS, a constraint grammar-based parsing system for Portuguese. In *Working with Portuguese Corpora*, 279–302. Bloomsburry Academic.

Bick, Eckhard. 2022. PFN-PT: A framenet annotator for Portuguese. *Domínios de Lingu@gem* 16(4). 1401–1435. doi 10.14393/dl52-v16n4a2022-7.

Bick, Eckhard. 2023. Attribution of quoted speech in Portuguese text. In *Contraint Grammar: Methods, Tools and Applications (NoDaLiDa 2023 Workshop)*, forthcoming.

Bick, Eckhard & Tino Didriksen. 2014. CG-3 — beyond classical constraint grammar. In *Nordic Conference of Computational Linguistics (NoDaLiDa)*, 31–39.

Bornet, Cyril & Frédéric Kaplan. 2017. A simple set of rules for characters and place recognition in french novels. *Frontiers in Digital Humanities* 4. n/p. doi 10.3389/fdigh.2017.00006.

Chaturvedi, Snigdha, Mohit Iyyer & Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. In *AAAI Conference on Artificial Intelligence*, 3159–3165. doi 10.1609/aaai.v31i1.10982.

Dekker, Niels, Tobias Kuhn & Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5. e189. doi 10.7717/peerj-cs.189.

Elson, David, Nicholas Dames & Kathleen McKeown. 2010. Extracting social networks from literary fiction. In *48th Annual Meeting of the Association for Computational Linguistics*, 138–147.

Goyal, Amit, Ellen Riloff & Hal Daumé III. 2010. Automatically producing plot unit representations for narrative text. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 77–86.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Mamede, Nuno & Pedro Chaleira. 2004. Character identification in children stories. In *Advances in Natural Language Processing*, 82–90. 10.1007/978-3-540-30228-5_8.

Paul, Apurba & Dipankar Das. 2017. A deep dive into identification of characters from Mahabharata. In *14th International Conference on Natural Language Processing (ICON)*, 447–455.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese. In *15th International Conference on Computational Processing of Portuguese (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Vala, 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. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 769–774. 10.18653/v1/D15-1088.

Valls-Vargas, Josep, Santiago Ontañón & Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. In *7th Intelligent Narrative Technologies Workshop*, 38–44.

Proposta recebida em Maio 2023 e aceite para publicação em Junho 2023.

Pais, filhos e outras relações familiares no DIP

Fathers, sons and other family relations in DIP

Cristina Mota
INESC-ID & Linguateca

Diana Santos
Linguateca & ILOS, UiO

Resumo

Neste artigo é descrita em pormenor a tarefa de identificação de relações familiares no Desafio de Identificação de Personagens (DIP), uma avaliação conjunta para identificar personagens em textos literários em português. Explicamos a motivação para esta subtarefa, e quais as dificuldades em criar uma coleção dourada com os valores corretos. Depois de referir em abstrato como se processa a avaliação desta sub-tarefa, relatamos os resultados do sistema participante, o PALAVRAS-DIP, e comentamos alguns problemas na sua avaliação. Além disso, descrevemos aquilo que aprendemos sobre a literatura lusófona com esta tarefa, assim como sugerimos outras pesquisas possíveis com este material.

Palavras chave

relações familiares, literatura lusófona, redes de personagens, identificação automática de relações

Abstract

In this paper we detail the relation identification extraction task of DIP, the character identification challenge in Portuguese. We describe the task and the choices made, explain how to evaluate it, and evaluate the results of the only participant, PALAVRAS-DIP. We also describe what we learned about lusophone literature with this task. Finally, we discuss further research with the compiled material.

Keywords

family relations, lusophone literature, character networks, automatic relation identification

1. Introdução

O desafio de identificação de personagens (DIP) foi uma avaliação conjunta organizada pela Linguateca, NuPILL-UFSC e Universidades do Maranhão e Oslo para estimular o desenvolvimento de sistemas que, dada uma obra completa, obtivessem as personagens, as formas de as identificar, características como o seu género e as suas profissões, ocupações e estatutos sociais, e — o que nos interessa no presente artigo — as relações familiares entre elas.

A própria avaliação conjunta foi descrita inicialmente em Santos et al. (2022) e mais especificamente em Santos et al. (2023). Aqui concentramo-nos na identificação de relações e nas várias formas de as caracterizar.

De qualquer maneira, é importante deixar claro que os sistemas teriam de entregar dois ficheiros: um que indicasse as personagens, a sua forma de serem chamadas, o seu género e a sua profissão, para cada obra. Nesse ficheiro uma identificação numérica era atribuída à personagem. E noutro ficheiro teriam de indicar as possíveis relações familiares entre essas personagens, usando os identificadores definidos antes.

A avaliação dos resultados dos sistemas seria feita comparando-os com uma coleção dourada (CD), ou seja, com as respostas certas previamente criadas pela organização, num subconjunto dos textos postos à disposição dos participantes. A organização do DIP criou uma coleção dourada para 40 obras, mais quatro usadas como exemplo, que é o que chamamos a coleção dourada total. Contudo, o PALAVRAS-DIP (Bick, 2023), o único sistema participante, apenas tratou as 100 obras em texto (e não as 100 obras em pdf), o que significa que apenas pôde ser avaliado sobre 21 obras, o que chamamos a CD de texto.

A motivação para incluir esta informação foi a nossa convicção de que as relações familiares eram importantes para o enredo, e frequentemente mencionadas, na literatura, e que por isso valeria a pena vê-las em conjunto, através de uma leitura distante.

2. Que relações?

As relações familiares que os sistemas participantes deveriam extrair de uma dada obra literária são as que existissem entre personagens com nome (identificadas, portanto, numa fase anterior), entre as seguintes: *mãe, pai, filho, filha,*

DOI: 10.21814/lm.15.1.402

neto, neta, avó, avô, irmã, irmão, cunhado, cunhada, primo, prima, tio, tia, sobrinho, sobrinha, bisavó, bisavô, bisneto, bisneta, nora, genro, sogro, sogra, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado, afilhada.

Devido a variadas grafias (*mãi* e *mãe*, *pae* e *pai*, etc.) nós normalizámos o nome das relações, não só graficamente, mas também lexicalmente, visto que existem muitas e variadas formas (sobretudo) de identificar um casal, como mencionado em Santos et al. (2023).

A escolha de adotar um vocabulário controlado das relações familiares teve o objetivo de facilitar o processo de avaliação dos sistemas participantes no DIP. Outra alternativa seria solicitar a identificação da forma exata pela qual estas relações são expressas na obra, mas isso exigiria muito provavelmente ma nova etapa de processamento (possivelmente manual), que mapeasse as relações identificadas pelos sistemas com as disponíveis na coleção dourada.

Como já referido em Santos et al. (2023), o vocabulário controlado de relações familiares adotado pelo DIP não cobriu todas as relações familiares existentes, faltando, por exemplo, relações como *padastro*, *madastra*, *enteado* e *enteada*. Além disso, e embora não estivessem na lista inicial, foram contempladas, e discutidas no ensaio, as relações *noivo* e *noiva* e *viúvo* e *viúva*.

3. Dificuldades na construção da coleção dourada

Embora a identificação das relações escolhidas pareça uma tarefa relativamente fácil, surgiram várias questões ao criar a coleção dourada, que gostaríamos de documentar aqui.

Em primeiro lugar, deveriam os anotadores também explicitar outras relações (biologicamente indiscutíveis, mas não expressas), tal como `X filho de Y` e `Y filho de Z` implica necessariamente `X neto de Z`?

Em segundo lugar, e essa questão já menos biologicamente determinada, mas muito mais frequente, se `X é marido de Y`, e `Y é mãe de Z`, deveriam os anotadores assumir, se nada fosse dito em contrário, que `X é pai de Z`?

A nossa resposta foi negativa em ambos os casos, mas é claramente uma opção discutível, à qual voltaremos mais tarde.

Em terceiro lugar, ao observar como as palavras do campo da família eram usadas nas obras, descobrimos que as formas de tratamento nem sempre são uma prova indiscutível duma relação familiar. É possível, por exemplo, tratar uma madrasta, e mesmo uma sogra, por *mãe*, assim como muitas personagens tratam pessoas mais novas sem qualquer vínculo familiar por *filhos* ou *filhas*.

Além disso, em alguns casos há informação incorreta (ou inconsistente) nos livros, ou por lapso do autor ou para indicar que a personagem em questão está equivocada — ou é ignorante. Veja-se o seguinte exemplo, em *Pero da Covilhã: Episódio Romântico do Século XV* de Zeferino Norberto Gonçalves Brandão: Catarina de Áustria exprime a seguinte queixa a seu cunhado D. Luiz, falando em relação ao seu marido, D. João III:

> Amor do povo e da patria como o nutriam em seus heroicos seios seu pai e avô Dom Manoel e Dom João!

Ora o pai deles era de facto Dom Manuel, mas D. João II, o rei anterior, era primo e não pai de Dom Manuel, e portanto não era avô de D. João III e de D. Luiz.[1]

O criador da CD tem de decidir (e eventualmente corrigir) a informação dada pela rainha, o que não é necessariamente fácil, até porque não seria de esperar erros destes num romance histórico.

Este é um tema que convém realçar: não é tão fácil quanto seria de esperar determinar as relações familiares entre as personagens de uma obra literária.

Finalmente, considerámos que não valia a pena escrever em duplicado relações que apareçam duas vezes, como em `X mulher de Y` e `Y marido de X`[2] ou `X irmão de Y` e `Y irmão de X`,[3] deixando ao sistema de avaliação a expansão automática de todos estes casos. Mas uma suposição, não tornada explícita na altura, era a de que o anotador da CD colocaria a primeira (muitas vezes única) vez que a relação era mencionada na obra. É aliás por isso que apresentaremos a panorâmica das relações na CD antes e depois da expansão.

[1] Este é um caso de romance histórico, em que portanto se podem confirmar as relações familiares corretas das personagens históricas, mas essas relações também se encontram corretamente especificadas noutras partes da mesma obra. Se o lapso foi propositado ou acidental, o que é certo é que não prejudica o enredo, mas prejudica certamente a extração automática das relações.

[2] O que significa que `mulher` e `marido` são relações inversas.

[3] Estas relações são simétricas, ou seja, a relação é igual à sua inversa — no caso das personagens terem o mesmo género, claro.

4. Relações identificadas no DIP

Depois destas explicações sobre a forma como encarámos a anotação das relações familiares, vejamos agora o panorama das relações encontradas através do DIP.

4.1. Encontradas na coleção dourada

O primeiro, e trivial, resultado, é que de facto nas 44 obras lidas atentamente, apenas uma não apresentava quaisquer relações familiares entre as personagens, nomeadamente *O Bom-Crioulo*, de Adolfo Caminha, o que valida a nossa intuição inicial de que este era um campo semântico recorrente na literatura.

Vemos na Figura 1 que as formas mais frequentes de descrever parentesco nas obras da coleção dourada eram *filho* e *filha*. Ao expandir (Figura 2), a relação familiar mais frequente tornou-se *pai*, o que deu origem ao título do presente artigo.

Duas conclusões podem ser tiradas: por um lado, é inegável a importância que a paternidade e a estrutura patriarcal têm (ou tinham) na literatura lusófona.[4] Por outro lado, dado que há muito mais personagens masculinas do que femininas, a posição de *mãe* é afinal mais significativa do que *pai*. Ou seja, a percentagem das mulheres que são mães é maior do que a dos homens que são pais.[5] Conforme comentado por Marcia Langfeldt, esta presença da mãe pode ser interpretada como mais uma prova da estrutura patriarcal, dado que na época da maior parte dos romances tratados no DIP ser mãe era uma profissão, senão a profissão da mulher.

Observando mais uma vez as relações familiares mais frequentes antes de expandir, pode também observar-se ser mais comum apresentar alguém (mulher) como mulher de outra pessoa, do que alguém (homem) como marido de outra.[6] Naturalmente, após a expansão, o número de maridos e de mulheres é o mesmo.

Uma coisa que, contudo, não podemos ainda concluir é qual a percentagem de personagens que está relacionada familiarmente com outras. Na próxima subsecção tratamos disso.

[4]Com o DIP, só nos podemos pronunciar sobre a literatura lusófona, mas é bem provável que isto seja uma constante da literatura ocidental.

[5]Visto que há 106 pais e 64 mães em 2813 homens e 830 mulheres, na coleção dourada total.

[6]Embora não possamos ter certeza absoluta de que não foram os próprios anotadores que introduziram este viés.

4.2. Uma visão de género através das relações familiares

Se observarmos todas as personagens identificadas na coleção dourada, 1075, e as classificarmos pelo número de relações familiares que apresentam, depois da expansão, vemos que existe uma diferença significativa entre personagens femininas e masculinas.

Nas Tabelas 1 e 2 vemos quantas relações por personagem foram encontradas, por personagem feminina e masculina, respetivamente.

Enquanto 80,3% dos homens não têm qualquer relação familiar com outras personagens, apenas 61,9% das mulheres estão na mesma posição "independente".

Além disso, o número médio de relações familiares de um homem, 0,29, é muito menor do que o de uma mulher, 0,73.

Num. de relações	casos
0	155
1	50
2	24
3	9
4	10
5	4

Tabela 1: Número de relações de personagens femininas

As personagens femininas com 5 relações familiares são Rosalina, de *O Cabeleira*, que tem uma irmã, um marido, e três enteadas;[7] D. Laura de *Amar, verbo intransitivo*, que tem marido e quatro filhos; Etelvina de *O Doutor Luís de Sandoval* com dois maridos, um pai, uma mãe e um filho, e finalmente Dona Glória, do *Dom Casmurro*, que tem um filho, um marido do qual fica viúva,[8] um irmão e um primo.

As personagens masculinas com 5 relações familiares são Gabriel, também de *O Cabeleira*, que tem mulher e irmão, dois filhos e um genro, e D. Affonso V, em *Pero da Covilhã: Episódio Romântico do Século XV*, que tem duas mulheres, uma irmã, uma sobrinha e um filho.

[7]Embora não estivesse na lista das relações a marcar, o/a anotador/a marcou na CD a relação de *madrasta*. Donde se conclui que deveríamos também ter feito um programa que verificava a lista de relações na CD, e a lista de relações na resposta dos participantes, para garantir que apenas as relações "oficiais" seriam avaliadas. Diga-se de passagem que o PALAVRAS-DIP também marca *amigo* e *amiga*, que retirámos automaticamente.

[8]E que portanto foi marcado tanto como mulher dele, como viúva dele.

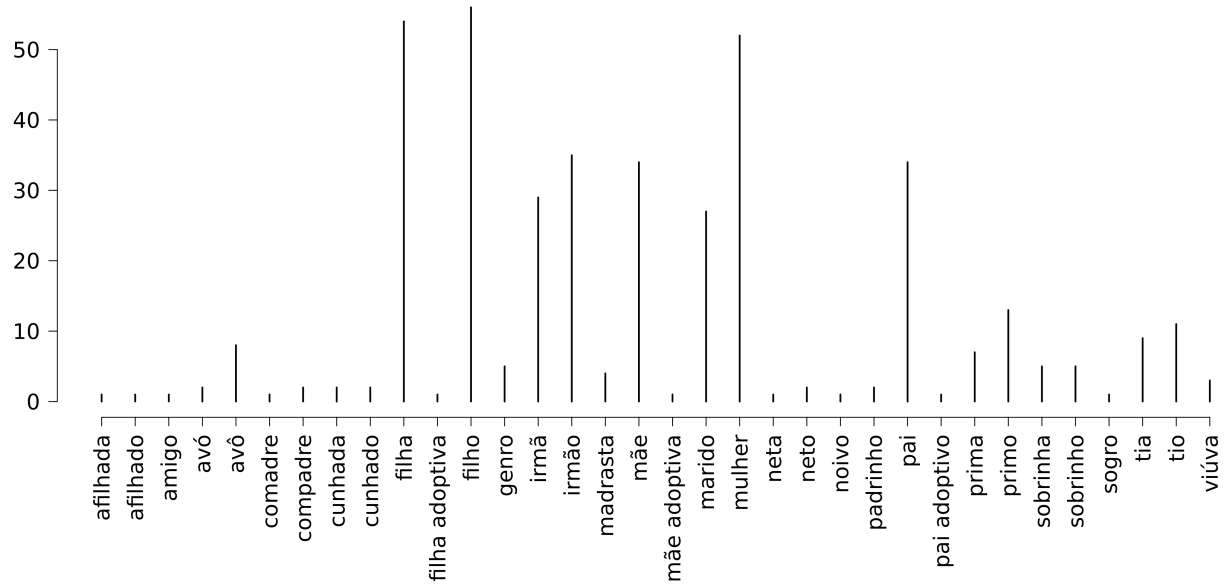

Figura 1: Distribuição das 414 relações em 43 obras

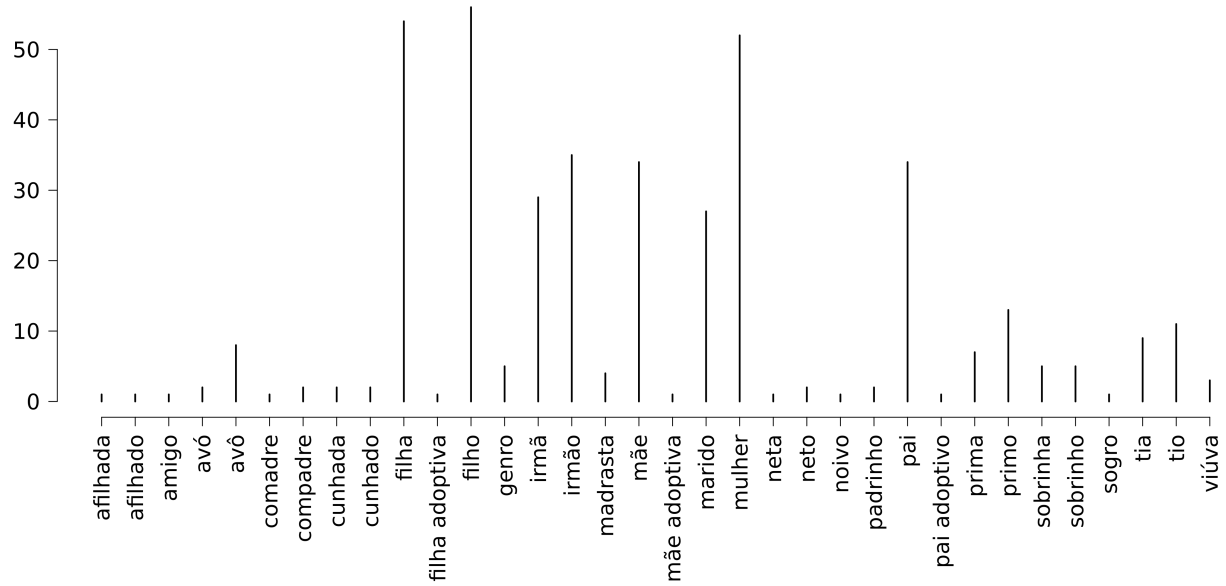

Figura 2: Distribuição das 811 relações após expansão

Num. de relações	casos
0	661
1	111
2	32
3	15
4	2
5	2

Tabela 2: Número de relações de personagens masculinas

4.3. Personagens principais e relações familiares

Vimos que é mais frequente que as personagens femininas sejam relacionadas familiarmente com outras personagens. Mas e se considerarmos simplesmente as personagens principais?

Para fazermos este estudo, e não tendo — como explicado em Santos et al. (2023) — feito diferenciação entre tipos de personagens, tivemos de usar uma operacionalização simples: A personagem com maior número de menções é a (ou uma das) personagens principais. Para isso

usámos a CD associada à coleção de texto (mais os três textos de exemplo), 24 textos, portanto, que pudemos investigar através do AC/DC (Santos, 2014).

Uma observação superficial permitiu confirmar que esta operacionalização parecia produzir resultados consonantes com a nossa impressão subjetiva das obras. No apêndice A apresentamos a personagem principal de cada obra obtida pelo método anterior, em que marcamos os poucos casos em que nos parece incorreta.

Obtivemos 5 obras com personagens principais femininas, e 19 com personagens principais masculinas. O número de relações familiares é muito maior nestes casos. De facto, apenas 4 personagens principais masculinas não tinham relações familiares com outras.

Em média, uma personagem principal feminina tem 2,20 relações com outras personagens, contra 1,35 relações de uma personagem principal masculina.

4.4. Relações extraídas pelo PALAVRAS-DIP

O panorama das relações extraídas das obras pelo PALAVRAS-DIP, apresentado na Figura 3, também foi submetido a um processo de expansão, embora o programa calcule automaticamente algumas relações inversas.

A situação ainda é mais flagrante em relação a *pai*, que é duas vezes mais frequente do que *mãe*. *Filho* é a segunda relação mais frequente identificada.

4.5. Obtidas pelo PALAVRAS-DIP na coleção extra

Como explicado em Santos et al. (2023), pedimos ao PALAVRAS-DIP para identificar as personagens noutra coleção maior, chamada coleção extra e descrita no texto referido, para podermos ter uma visão mais global da literatura lusófona. Os resultados apresentam-se na Figura 4.

Embora a avaliação tenha sido feita alguns meses após o DIP, e portanto com uma versão diferente do PALAVRAS-DIP, os resultados não são significativamente diferentes: *pai* continua a ser a relação mais frequente, *filho* continua a ser mais frequente do que *filha*, e *irmão* mais do que *irmã*, o que não admira sabendo que há mais personagens masculinas que femininas na coleção.

Seja como for, e visto que — ao contrário de outras características — os resultados do PALAVRAS-DIP não foram especialmente bons em relação à identificação das relações, não tiramos muitas conclusões destes dados.

096,4,mulher,1	Dona Laura mulher de Sousa Costa
096,5,irmão,6	Carlos irmão de Maria Luísa
096,4,mãe,5	D. Laura mãe de Carlos
096,4,mãe,6	**D. Laura mãe de Maria Luísa**
096,7,irmã,5	**Aldinha irmã de Carlos**
096,7,irmã,6	Aldinha irmã de Maria Luísa
096,7,filha,4	Aldinha filha de D. Laura
096,8,filha,4	**Laurita filha de D. Laura**
096,8,irmã,7	Laurita irmã de Aldinha
096,8,irmã,6	Laurita irmã de Maria Luisa
096,8,irmã,5	Laurita irmã de Carlos

Tabela 3: O conteúdo da CD: a negrito as relações que o sistema identificou, as outras estão em falta

096,1,filha,9	**Aldinha filha de D. Laura**
096,1,filha,7	Aldinha filha de Carlos
096,7,pai,1	Carlos pai de Aldinha (=ant)
096,1,irmã,7	**Aldinha irmã de Carlos**
096,7,irmão,1	Carlos irmão de ALdinha (=ant)
096,4,pai,9	não há 4
096,9,filha,4	não há 4 (=ant)
096,7,marido,9	Carlos marido de Dona Laura
096,24,filha,7	Laurita filha de Carlos
096,24,filha,9	**Laurita filha de D. Laura**

Tabela 4: O resultado de um sistema: a negrito as relações que o sistema identificou, as outras são espúrias

5. Avaliação da identificação das relações

Como explicado no artigo Willrich & Santos (2023), usámos como medida de avaliação a medida F, contando as relações certas, as em falta e as espúrias.

Contudo, a simplicidade da medida esconde vários pormenores complicados e não necessariamente satisfatórios — ou cabalmente resolvidos –, que tentaremos ilustrar aqui, com este exemplo concreto, relativo ao romance *Amar, verbo intransitivo*, de Mário de Andrade.

As relações anotadas na CD para esta obra são apresentadas na Tabela 3. A primeira coluna da tabela mostra o conteúdo da CD enquanto a segunda é a nossa tradução, tendo identificado as personagens a que se referem os identificadores.

Imaginemos agora que um sistema tinha produzdo o seguinte resultado, na Tabela 4.

Em ambas as tabelas colocámos já o resultado da avaliação. Na Tabela 3, vemos que há

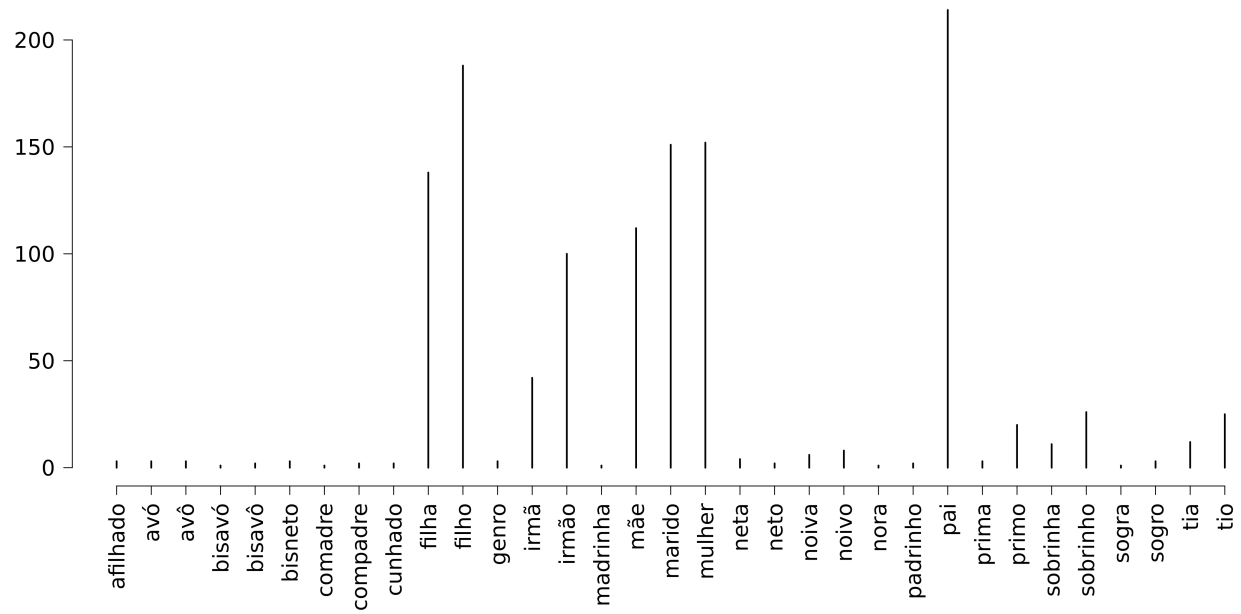

Figura 3: Distribuição das 1245 relações encontradas pelo PALAVRAS-DIP (expandidas) em 100 obras processadas

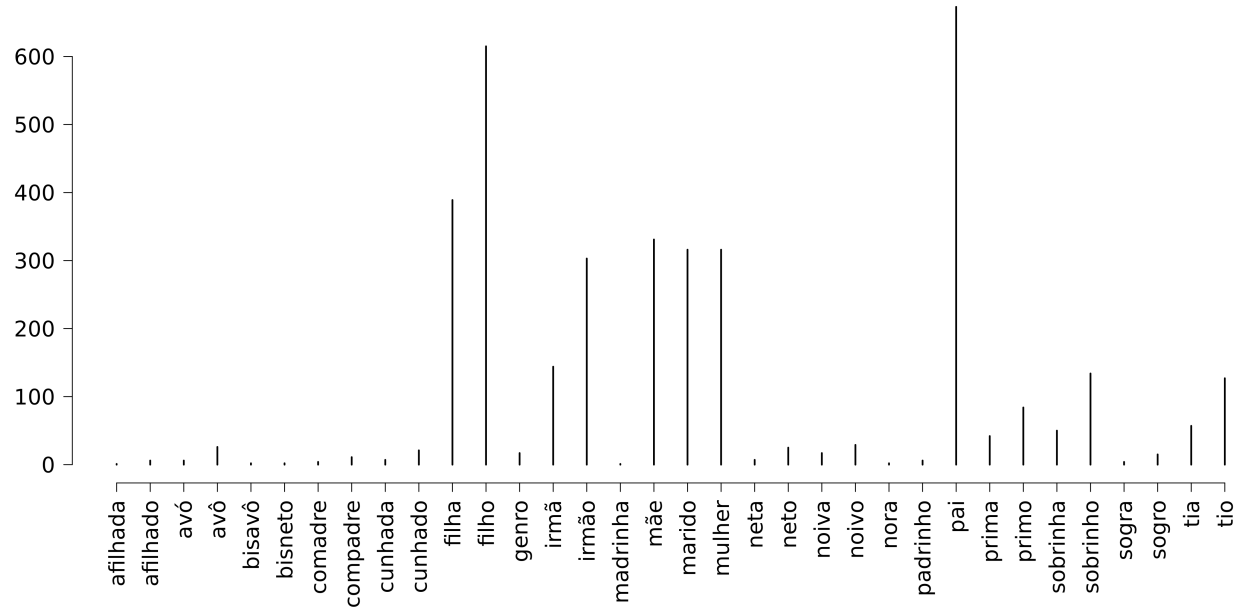

Figura 4: Distribuição das 3791 relações encontradas pelo PALAVRAS-DIP (expandidas) em 213 obras processadas

11 relações. O sistema identificou 3, e há portanto 8 relações que faltam. Ao expandir, ficamos com 16 que faltam e 6 certas, donde a abrangência é de 6/22=0,2727.

Na Tabela 4 — em que, note-se só contámos uma vez quando as relações e as suas inversas foram apresentadas pelo sistema –, há 7 relações: 3 certas (identificadas corretamente pelo sistema) e 4 espúrias, donde a precisão, depois de expandir, é de 6/14=0,4286.

A medida-F é, então, de 0,333.

O problema desta análise, que não é visível olhando apenas para estas tabelas, é o facto de o sistema ter amalgamado numa mesma personagem duas que além disso estavam relacionadas familiarmente, nomeadamente Carlos e o seu pai, como se pode ver na Figura seguinte, relativa às personagens:

```
096,7,Carlos Alberto Sousa Costa|Carlos|
Senhor Costa|Senhor Sousa Costa|Sousa
Costa|senhor Sousa Costa,M,palhaço|
filósofo|artista
```

Na CD, há uma personagem, Carlos, com os nomes `Carlos Alberto Sousa Costa|Carlos` e outra (o pai), com os nomes `Senhor Costa|Senhor Sousa Costa|Sousa Costa|senhor Sousa Costa`.

Para avaliar, os programas de avaliação tentam alinhar as personagens "reais" com as propostas pelo sistema, mas têm de escolher uma interpretação. E por isso, embora o sistema tivesse identificado corretamente que Sousa Costa era pai de Aldinha, e que Carlos era irmão de Aldinha, não pudemos considerar ambas as relações corretas, visto que Carlos e o pai tinham sido considerados a mesma personagem. Pior ainda, temos de considerar algumas certas e outras erradas, penalizando duramente o sistema.

Outra questão que este exemplo suscita é a arbitrariedade do conteúdo da coleção dourada. A mesma família (ou melhor, os mesmos graus de parentesco) podiam ter sido exprimidos mais completamente. Afinal se `A é irmã de B` e `B é irmão de C`, A e C também são irmãos (repare-se que não considerámos no DIP os meios-irmãos, que complicariam certamente estas deduções).

Mas se tivéssemos colocado mais uma relação de irmãos, e o sistema não, estaríamos a piorar a sua abrangência. Pior ainda, se o sistema tivesse colocado essa relação (certíssima) e não estivesse na CD, seria penalizado com a suposição de que a relação era espúria.

Isso poderia levar a crer que deveríamos sempre ter calculado a expansão total (o fecho transitivo) das relações, tanto na coleção dourada como no resultado do sistema a avaliar, para evitar estes problemas.

Contudo, estaríamos a penalizar muitíssimo um sistema que não tivesse encontrado, por exemplo, uma relação de irmão, numa família de cinco irmãos. Em vez de faltar essa (e a sua inversa) faltariam 8 relações, diminuindo radicalmente a abrangência. E a medida de avaliação seria mais ou menos rigorosa dependendo da estrutura familiar do romance, algo que também não parece muito justificado.

Por isso mesmo, acabámos por deixar que a falta de identificação de uma dada relação apenas contasse 2, embora conscientes de que poderíamos em alguns casos penalizar o sistema.

A nossa conclusão é que a única maneira de produzir uma avaliação justa obriga a realizar o processo de forma semi-automática, em que os casos espúrios são analisados por uma pessoa, a fim de, se for caso disso, adicionar mais relações à CD.

6. A identificação de famílias

Até agora temos apenas falado e refletido sobre relações entre duas personagens (e a sua relação inversa). Mas um dos interesses óbvios da extração de relações familiares é estudar a estrutura das famílias, que é o que tentaremos fazer nesta secção, juntando a informação das relações obtidas.

Nas Figuras 5 a 8, apresentamos as ligações familiares no mar de personagens de cada obra, para quatro obras com características diferentes.

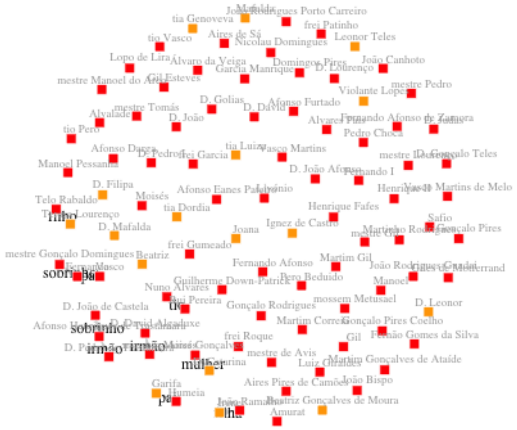

Figura 5: Relações entre as personagens de *Os tripeiros*

Os tripeiros, de António José Coelho Lousada, como romance histórico que é, tem muitas personagens, mas poucas relacionadas por parentesco, veja-se a Figura 5.

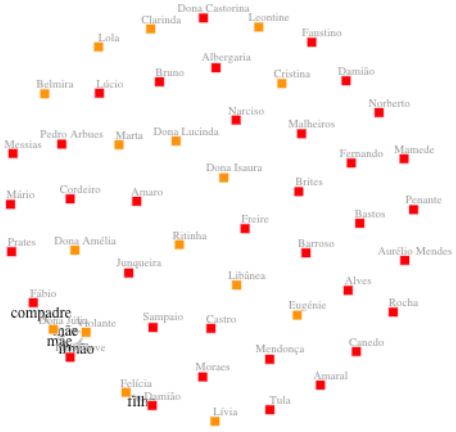

Figura 6: Relações entre as personagens de *Turbilhão*

Turbilhão, de Coelho Neto, por outro lado, tem um menor número de personagens, mas também poucas têm relações familiares entre si, cf. a Figura 6.

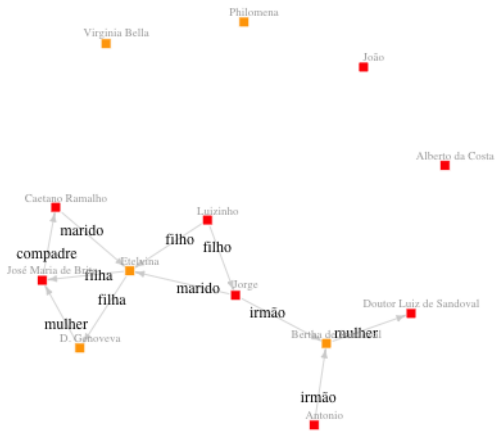

Figura 7: Relações entre as personagens de *O Dr. Luis Sandoval*

O Dr. Luis Sandoval, de Alice Pestana, pelo contrário, é um caso de pouquíssimas personagens, praticamente todas relacionadas entre si, como a Figura 7 ilustra.

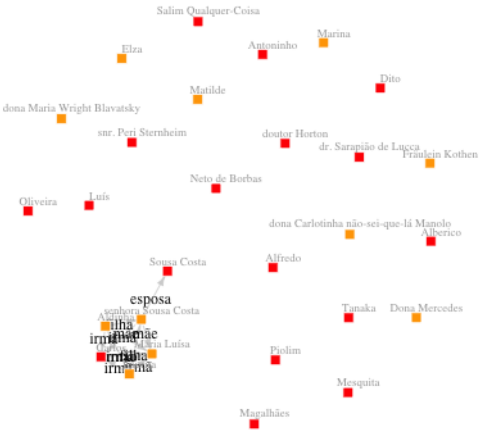

Figura 8: Relações entre as personagens de *Amar, verbo intransitivo*

Finalmente, *Amor, verbo intransitivo*, de Mário de Andrade, que é passado praticamente dentro de uma família apenas, é interessante notar na Figura 8 que existe um núcleo muito denso de relações, mas que se assemelha mais aos dois romances iniciais.

Se, pelo contrário, fizermos uma visualização em que apenas apresentamos as personagens relacionadas, podemos apreciar melhor a questão das diferentes famílias, e o tamanho destas. Vejam-se os seguintes casos:

Em *A ermida de Castromino*, de Teixeira de Vasconcelos, identificam-se três famílias, veja-se a Figura 9.

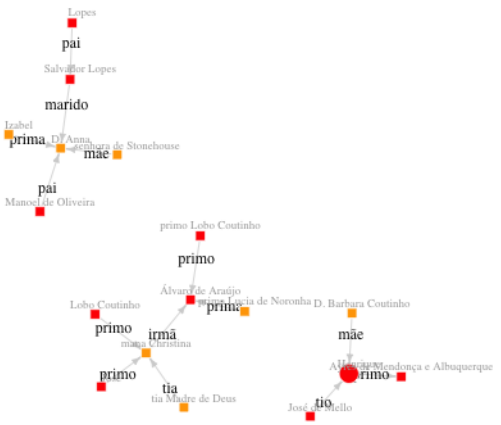

Figura 9: Relações entre as personagens de *A ermida de Castromino*

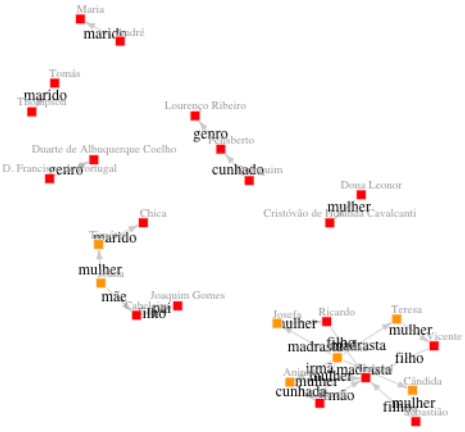

Figura 10: Relações entre as personagens de *Cabeleira*

Enquanto em *O Cabeleira*, de Franklin Távora, ilustrado na Figura 10, se descortinam não menos do que sete famílias diferentes, uma delas com dez elementos.

Em contraponto, apresenta-se novamente a única família de *Amar, verbo intransitivo*, na Figura 11, com seis elementos.

E para finalizar, na Figura 12 apresentamos as quatro famílias presentes em *Uma família inglesa*, de Júlio Dinis. Conhecido como é o enredo ser a relação entre as duas famílias Whitestone e Quintino, observa-se que outros dois relacionamentos, pouco relevantes para a obra, aparecem na figura como igualmente válidos, o que leva a concluir que seria definitivamente interessante refazer as figuras com alguma informação sobre a importância (e/ou frequência) que as diferentes personagens têm nas obras em questão.

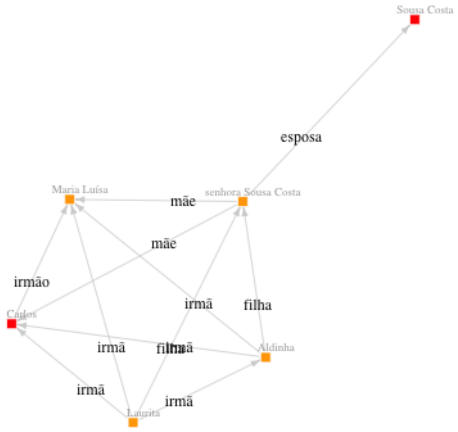

Figura 11: Relações entre as personagens de *Amar, verbo intransitivo*

Figura 12: Relações entre as personagens de *Uma família inglesa*

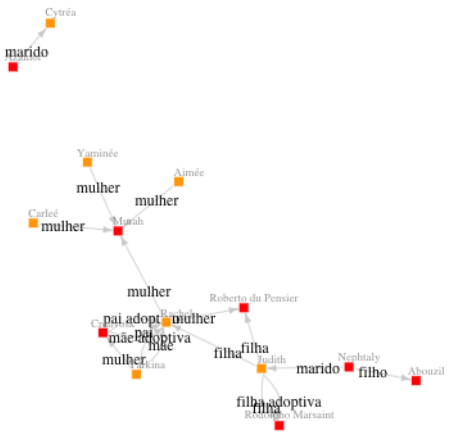

Figura 13: Relações entre as personagens de *A Judia Raquel*

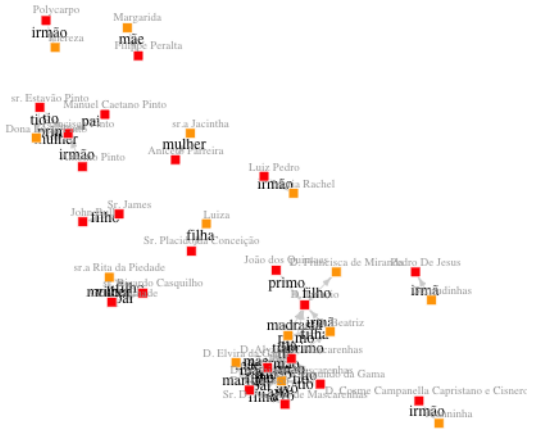

Figura 14: Relações entre as personagens de *Os cavaleiros da cruz vermelha*

Uma última distinção pode ser ilustrada com as famílias presentes em *A Judia Raquel*, de Francisca Senhorinha de Motta Dinis: uma família extremamente complexa e outra muito simples é o que podemos apreciar na Figura 13.

Compare-se com as 11 famílias dos 4 volumes de *Os homens da cruz vermelha*, de Carlos Pinto de Almeida, em que apenas uma é significativamente maior ou mais complexa do que as outras, na Figura 14.

Para mais figuras veja-se também a apresentação no Encontro do DIP (Mota, 2022).

7. Observações finais

Do aqui apresentado, podemos concluir o seguinte:

- A maioria das personagens não tem relações de parentesco com outras personagens, mas isso já não se verifica quando são personagens principais.
- A relação pai (e consequentemente a sua inversa, filho ou filha) é a que ocorre mais frequentemente entre personagens, sobrepondo-se à de mãe-filho/a ou mulher/marido.
- O número de famílias varia entre 1 a 11 (de acordo com a coleção dourada)
- A solução encontrada para avaliar as relações na primeira edição do DIP não é perfeita, e poderia ser preferível ter uma avaliação humana de casos que deveriam ser adicionados à coleção dourada das relações.

Um assunto extremamente importante não foi, contudo, ainda aqui abordado, nomeadamente: As relações familiares podem variar ao longo do tempo (em obras diferentes) ou no espaço temporal da obra — por exemplo, levando à união de duas famílias pelo casamento.

O que nos leva à seguinte autocrítica: No DIP quisemos apenas aceitar informação relativa à obra e não ao desenrolar do enredo, mas falhámos claramente nisso quando sugerimos identificar relações matrimoniais, visto que numa grande maioria dos casos as ditas fazem parte do enredo. Muitas vezes os protagonistas são solteiros e acabam por casar, ou casam no princípio e depois têm filhos, etc.

Seja como for, o que queremos sublinhar aqui é que deveríamos ter tentado distinguir esse estabelecimento de relações, parte integrante do enredo, dos casos em que essas relações permanecem as mesmas durante a obra toda.

Idealmente, deveríamos fazer uma nova revisão da coleção dourada para encontrar esses casos e marcá-los separadamente, por exemplo `X mulherENREDO Y` se o casamento fizer parte da história que se desenrola na obra.

O mesmo poderá fazer sentido nos casos — embora consideravelmente mais raros — em que há uma descoberta de relações familiares como parte da intriga,[9] que seria então marcado, por exemplo, `X filhoENREDO Y`.

8. Próximos passos

Nesta secção damos uma panorâmica de estudos ou projetos interessantes que poderiam ser feitos como continuação do trabalho descrito aqui.

Em primeiro lugar, a continuação (ou melhoria) do cruzamento da identificação das personagens principais com as relações familiares.[10] Fizemos isso apenas para 24 obras, mas poderíamos tentar fazê-lo para as 213+80 analisadas pelo PALAVRAS-DIP.

Por outro lado, até agora, nas redes que representámos, não existe qualquer informação sobre a importância das personagens, e assim não podemos saber realmente o que significa que a relação *pai* é a mais frequente: é porque são os pais os protagonistas, ou porque os protagonistas filhos são definidos/apresentados através da sua filiação?

A construção de redes que representassem não só as ligações familiares entre personagens mas também a sua importância relativa (medida através do número de vezes que eram mencionadas na obra) permitiria uma maior compreensão da importância ou não das relações. Isto é aliás prática corrente na construção de redes, como feito por exemplo em Santos & Freitas (2019).

De facto, outro trabalho (relacionado) que traria mais algum conhecimento sobre as relações familiares seria saber quantas vezes os familiares estão em presença uns dos outros — ou seja, fazer redes interacionais como as apresentadas em Santos & Freitas (2019) ou em Bick (2023).

Investigar as formas de tratamento entre personagens seria algo também extremamente interessante para compreender códigos culturais, identificando como é que as personagens se dirigem aos pais, aos filhos, e aos esposos.

Para isso, no entanto, seria preciso identificar os trechos em discurso direto, algo que também permite caracterizar melhor as obras. Por exemplo, obras com muito discurso direto também permitiriam investigar o protagonismo das diferentes personagens. Quem fala, e quem cala. Quem é abordado/mandado, e quem manda. Estas são chamadas redes conversacionais, e são das primeiras usadas na leitura distante em análise literária, propostas em Moretti (2011).

Finalmente, algo que pretendíamos fazer mas que ficou para trabalho futuro foi o uso de conceitos de redes (centralidade, conetividade, etc.) para caraterizar os diferentes romances e novelas.

9. Outros trabalhos de identificação de relações familiares

Terminamos este artigo com uma breve panorâmica de estudos relacionados. Não em relação a todas as possíveis redes de personagens possíveis de extrair de obras literárias, mas apenas daqueles (poucos) trabalhos que se dedicaram aos laços de parentesco. Veja-se Willrich & Santos (2023) para uma panorâmica de formas de avaliar a identificação de relações entre personagens e Abreu et al. (2013) para uma revisão da identificação de relações entre entidades em geral com um foco específico na aplicação dessa tarefa ao português.

No ReReLEM (Freitas et al., 2009), a tarefa de identificação de relações entre entidades mencionadas na avaliação do Segundo HAREM [11], algumas das relações identificadas eram de família, mas não foram tratadas de forma especial, sendo apenas um dos tipos de relação entre entidades, não sendo pedido para marcar especificamente o tipo de relação familiar.[12] Dos 10 sistemas par-

[9] Por exemplo, descobre-se o verdadeiro pai de uma protagonista.

[10] De facto, esse cruzamento seria interessante para todas as facetas estudadas no DIP, mas aqui limitamo-nos a considerar as relações de parentesco.

[11] https://www.linguateca.pt/HAREM/

[12] De facto, durante a avaliação, a relação familiar estava englobada na categoria OUTRA e só após a avaliação foi marcada especificamente juntamente com outras subcategorias num total de 22. Dessas, a relação familiar é a mais frequente a seguir à do vínculo institucional.

ticipantes no Segundo HAREM, apenas 3 participaram na pista do ReRelEM.

Em Higuchi et al. (2019) foi incluída a anotação das relações familiares no AC/DC para estudá-las no contexto da política brasileira, mas não limitámos a identificação dos laços de parentesco a políticos com nome, nem estudámos — até agora — a conetividade das redes familiares.

Em 2010, Santos et al. (2010) propuseram um sistema baseado em regras para identificar relações familiares não apenas entre entidade mencionadas mas também entre outras entidades mesmo que não referidas pelo seu nome próprio. Este sistema foi avaliado em dois corpos: um contendo as biografias de todos os reis portugueses encontradas na Wikipedia e outro composto por frases extraídas do CETEMPúblico (Rocha & Santos, 2000) que incluem um nome de relação. Numa primeira avaliação, a relação só era considerada correcta se os argumentos fossem exactamente iguais, mas numa segunda avaliação, os argumentos podiam ser parcialmente coincidentes. Este dois corpos foram anotados manualmente depois do sistema ter sido desenvolvido. Não é claro quais as relações familiares tratadas, mas foram identificadas pelos anotadores 105 relações no primeiro corpo e 21 relações explícitas em 110 frases no segundo corpo. Neste último caso, 89 frases incluem um nome de relação mas a relação familiar não está definida explicitamente entre duas personagens e como tal foram excluídas da avaliação.

Para o inglês, Azab et al. (2019) apresentam um novo modelo de palavras pulverizadas ("word embeddings") para representar personagens de filmes e as suas interações em diálogos, entrando em conta com os interlocutores. Esse modelo é utilizado para avaliar duas tarefas: identificação do grau de relação entre personagens ("character relatedness") e classificação da relação entre personagens ("character relation") — as relações familiares são classificadas de forma fina (e.g., pai/filho/tio/inimigo), grosseira (e.g., familar/social/profissional) e também quanto a sentimento (positivo/negativo/neutro). Este modelo foi avaliado em 31 obras de Shakespeare que fazem parte de um corpo literário com 109 peças anotado manualmente com recurso ao Mechanical Turk da Amazon (Massey et al., 2015) e que inclui 18 classes finas de relações, 4 classes grosseiras de relações e 3 classes de sentimento. Neste corpo foram anotadas 2170 relações das quais mais de 800 são relações de parentesco.

He et al. (2013) também identificam relações familiares e sociais (como de amizade) com o objectivo de construir redes de relações entre personagens cujos arcos entre personagens representam relações mútuas existentes entre essas personagens. Este modelo foi treinado na obra *Pride and Prejudice* de Jane Austen e testado adicionalmente nas obras *Emma* da mesma autora and *The Steppe* de Chekov (em inglês).

Agradecimentos

Agradecemos a Eckhard Bick muitas perguntas e críticas pertinentes, a Roberto Willrich várias referências relevantes e muitos comentários de melhoria ao próprio texto, e ao resto da organização do DIP pelo trabalho de equipa.

Agradecemos a Marcia Langfeldt vários comentários e sugestões para tornar os resultados mais relevantes para um público literário.

E finalmente, agradecemos à FCCN – Fundação para a Computação Científica Nacional (Portugal) o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências

Abreu, Sandra Collovini de, Tiago Luis Bonamigo & Renata Vieira. 2013. A review on relation extraction with an eye on Portuguese. *Journal of the Brazilian Computer Society* 19. 553–571. 10.1007/s13173-013-0116-8.

Azab, Mahmoud, Noriyuki Kojima, Jia Deng & Rada Mihalcea. 2019. Representing movie characters in dialogues. Em *23rd Conference on Computational Natural Language Learning (CoNLL)*, 99–109. 10.18653/v1/K19-1010.

Bick, Eckhard. 2023. Extraction of Literary Character Information in Portuguese. *Linguamática* 15(1). 31–40. 10.21814/lm.15.1.397.

Freitas, Cláudia, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira & Paula Carvalho. 2009. Detection of relations between named entities: report of a shared task. Em *NAACL-HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, 129–137.

He, Hua, Denilson Barbosa & Grzegorz Kondrak. 2013. Identification of speakers in novels. Em *51st Annual Meeting of the Association for Computational Linguistics*, 1312–1320.

Higuchi, Suemi, Diana Santos, Cláudia Freitas & Alexandre Rademaker. 2019. Distant reading

Brazilian politics. Em *4th Conference of The Association Digital Humanities in the Nordic Countries*, 190–200.

Massey, Philip, Patrick Xia, David Bamman & Noah A. Smith. 2015. Annotating character relationships in literary texts. *CoRR* abs/1512.00728. http://arxiv.org/abs/1512.00728.

Moretti, Franco. 2011. Network theory, plot analysis. *New Left review* 68. 80–102.

Mota, Cristina. 2022. Pais, filhos e outras relações no Desafio de Identificação de Personagens (DIP). Apresentação. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/Encontro_DIP_relacoes_Dec2022.pdf.

Rocha, Paulo Alexandre & Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. Em *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR)*, 131–140.

Santos, Daniel, Nuno Mamede & Jorge Baptista. 2010. Extraction of family relations between entities. Em *II Simpósio de Informática (INForum)*, 549–560.

Santos, Diana. 2014. Corpora at Linguateca: Vision and roads taken. Em Tony Berber Sardinha & Telma de Lurdes São Bento Ferreira (eds.), *Working with Portuguese Corpora*, 219–236. Bloomsbury.

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

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP - desafio de identificação de personagens: objectivo, organização, recursos e resultados. *Linguamática* 15(1). 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational Processing of the Portuguese Language (PROPOR)*, 413–419.

Willrich, Roberto & Diana Santos. 2023. Avaliação no desafio de identificação de personagens. *Linguamática* 15(1). 69–87. 10.21814/lm.15.1.398.

A. Personagens principais na CD de texto

Só há três casos, sinalizados por ponto de interrogação, em que a nossa interpretação não concorda com o resultado quantitativo, e num deles, o caso da obra *O Ateneu*, a personagem principal é o narrador.

001-1	472	M	0	Agrippino Simões
002-1	91	M	2	António
004-32	148	M	0	João Bispo?
005-2	345	M	1	Diogo
006-2	340	M	2	Paulo
025-3	295	F	1	Helena
026-1	208	F	2	Simá
030-2	735	M	2	Amâncio
032-3	220	F	4	Etelvina
033-1	1134	M	3	Henrique
037-21	379	M	0	Fernão?
043-1	335	M	2	Cabeleira
047-3	395	F	2	Isaura
051-1	23	M	1	Luis
054-15	193	M	1	Eugénio
055-4	232	M	0	Amaro
064-1	150	M	4	Aristarco?
072-5	78	M	3	Dom Luiz
075-8	300	M	1	Pero da Covilhã
096-5	326	M	4	Carlos
099-3	1044	M	1	Carlos
201-1	698	M	1	Rubião
203-2	598	M	3	Daniel
204-13	344	F	2	Capitu

B. Famílias na CD total

ID	num. de famílias	tam. das famílias
001	1	2
002	1	3
004	8	3 3 2 2 2 2 2
005	3	3 2 2
006	2	4 2
025	2	2 2
026	4	3 3 3 2
030	3	3 3 3
032	1	9
033	3	7 6 4
037	2	7 3
043	7	10 5 3 2 2 2 2
047	2	4 3
051	1	2
054	6	5 3 3 2 2 2
055	0	
064	1	5
072	1	8
075	8	13 4 4 2 2 2 2 2
096	1	6
099	4	3 3 2 2
103	5	4 2 2 2 2
107	5	5 3 3 2 2
109	1	5
110	6	6 3 3 2 2 2
111	4	5 3 2 2
113	3	3 3 2
116	1	3
121	2	2 2
126	2	12 2
129	4	4 3 3 2
130	2	5 2
149	11	12 5 3 2 2 2 2 2 2 2 2 2
151	4	8 7 2 2
157	5	3 3 3 2 2
159	4	5 4 2 2
177	2	10 3
180	2	4 3
197	2	2 2
201	4	8 5 4 3
203	3	4 3 2
204	2	9 4

Proposta recebida em Maio 2023 e aceite para publicação em Junho 2023.

Desafios e vantagens do processo de identificação automática do gênero e das profissões das personagens no DIP

Challenges and advantages of the automatic identification of character gender and professions in DIP

Emanoel Pires
Universidade Estadual do Maranhão

Marcia Caetano Langfeldt

Rebeca Schumacher Fuão

Resumo

O desenvolvimento de sistemas para identificação automática de personagens e de algumas de suas características é o objetivo central do projeto Desafio de Identificação de Personagens (DIP) desenvolvido junto à Linguateca. Dentre essas características, trataremos neste artigo da identificação do gênero e das profissões das personagens. Primeiramente, justificaremos a nossa escolha em trabalhar com esses dois dados, apresentando os diferentes caminhos que trilhamos para estabelecer diretrizes para a identificação dos mesmos. A identificação manual do gênero e da profissão é exaustiva e passível de falhas, sendo cada vez mais comum o uso de sistemas computacionais para essa tarefa. A análise das profissões permitiria refletir sobre questões como a definição de profissão, sua frequência em obras brasileiras e portuguesas, e possíveis relações com os gêneros literários. Em seguida, apresentaremos alguns resultados provenientes da leitura distante e da leitura próxima de um grupo de obras. Contrastaremos esses resultados e comentaremos os desafios e as vantagens que encontramos ao longo dessa tarefa e que parecem reforçar a nossa hipótese de preferência por um esforço combinado de sistemas automáticos e interpretação humana na identificação de personagens.

Palavras chave

leitura distante, identificação de personagens, gênero, profissão

Abstract

The development of systems for automatic identification of characters and some of their characteristics is the central objective of the Character Identification Challenge (DIP) project developed in conjunction with Linguateca. Among these characteristics, 2 this article will focus on the identification of gender and professions of the characters. Firstly, we will justify our choice to work with these two data sets, presenting the different paths we have taken to establish guidelines for their identification. Manual identification of gender and profession is exhaustive and susceptible to errors, making the use of computer systems increasingly common for this task. The analysis of professions would allow reflection on issues such as the definition of a profession, its frequency in Brazilian and Portuguese works, and possible relationships with literary genres. We present some results from distant and close reading of a group of works, contrast these results and comment on the challenges and advantages we encountered throughout this task, which seem to reinforce our hypothesis of a preference for a combined effort of automatic systems and human interpretation in character identification.

Keywords

distant reading, character identification, gender, profession

1. Introdução

Identificar personagens a partir da aplicação de sistemas automáticos num corpus de literatura em prosa de língua portuguesa é a questão central do projeto desenvolvido junto à Linguateca intitulado Desafio de Identificação de Personagens (DIP) (Santos et al., 2022). Esta tarefa tem nos revelado novos dados impossíveis de serem identificados numa leitura próxima (*close reading*), primeiramente porque a velocidade de identificação é imensamente inferior àquela realizada pelos sistemas e, em seguida, porque o volume de obras passíveis de serem trabalhadas é infinitamente inferior àquele que os sistemas podem dar conta. Já no início dessa atividade, um primeiro desafio foi levantado pelo grupo. Era preciso retomar o conceito de personagem e estabelecer diretrizes

DOI: 10.21814/lm.15.1.401

para o que os sistemas considerariam e o que descartariam como sendo personagens. Além disso, era preciso igualmente apontar as variadas formas pelas quais as personagens são 3referidas de modo a evitar o máximo possível as perdas que os sistemas poderiam ter em suas identificações.

Isso nos levou a renovar uma discussão sobre o que são as personagens literárias. Questionamentos como: o que faríamos com personagens que são animais? E com as figuras históricas mencionadas no interior da obra? Esta análise pode ser lida na íntegra no artigo sobre personagens publicado na Linguateca (Langfeldt et al., 2021). O que fica de importante para este artigo é que decidimos de que não haveria identificação de personagens que não contribuíam para o desenvolvimento da narrativa, como por exemplo personagens históricas que apenas são citadas no texto. Isso porque o nosso interesse justamente era o de observar personagens construídas dentro dessas obras e alguns elementos que faziam parte das escolhas dos autores nessas construções e que poderiam, portanto, indicar tendências de diferentes épocas, países e estilos nos textos analisados.

Dentre as categorias de nosso interesse, falaremos aqui sobre o gênero e a profissão das personagens. A identificação do gênero das personagens é um aspecto crucial da análise literária, uma vez que o modo como um gênero é apresentado na prosa ficcional pode revelar aspectos sobre o contexto cultural e histórico da obra, bem como as intenções e modos de pensar do seu autor. Entretanto, identificar manualmente o gênero das personagens de um largo conjunto de dados pode ser um processo exaustivo e passível de falhas e omissões. Neste sentido, o uso de sistemas de computador para identificar e analisar o gênero das personagens literárias tem se tornado cada vez mais comum na pesquisa literária (Bamman et al., 2014; Elsner, 2012).

A identificação das profissões viria complementar essa análise nos permitindo refletir sobre algumas outras questões como: o que seria considerado como profissão e o que seria considerado como uma mera ocupação ou estatuto social? Seria possível identificar profissões mais frequentes em obras brasileiras que portuguesas? E entre os gêneros, haveria alguns gêneros privilegiados em determinadas profissões? Encontraríamos alguma constância e/ou aspectos reveladores nos resultados?

1.1. A questão do gênero

O gênero na literatura se assemelha e se diferencia do seu conceito para- textual, no sentido em que, embora as personagens possam ser classificadas em gêneros mais ou menos precisos, o contexto literário no qual estão inseridos pode deslocá-las desta posição. Talvez o caso mais emblemático da literatura brasileira de uma personagem que muda de gênero seja o de Diadorim/Reinaldo, do romance *Grande Sertão: Veredas* (1956), de João Guimarães Rosa (1908–1967), em que a personagem Diadorim se traveste do jagunço Reinaldo, e mais recentemente, exemplos como os ilustrados no romance *Stella Manhattan* (1985), de Silviano Santiago e no livro reportagem *Ricardo e Vânia* (2019). No primeiro, não há exatamente uma mudança de gênero da personagem, mas a criação, pela personagem Eduardo da Costa e Silva, de um *alter ego* chamado Stella Manhattan. No último exemplo, Chico Felitti acrescenta ao conhecido relato sobre Ricardo a história de Vânia, que antes já tinha se chamado de Vagner.

É importante distinguir entre o gênero de uma personagem literária e o gênero ou o sexo de uma pessoa no mundo real, pois se tratam de dois conceitos diversos. Quando utilizamos sistemas de computação para identificar o gênero de uma personagem literária, estamos nos referindo à sua representação textual em uma obra literária, e não ao gênero ou sexo na vida real.

O gênero de uma personagem se refere ao modo como ela é representada em uma obra literária, inclusive os traços de personalidade, os comportamentos, as ações e os estereótipos de gênero. O gênero de uma personagem literária é, portanto, uma construção social que pode ser influenciada pela cultura, o período histórico e a intenção do autor.

1.2. A questão da profissão

A preocupação em identificar a ocupação surgiu do mesmo interesse em recolher informações que julgávamos básicas para caracterizar uma personagem. No início das nossas reflexões, logo nos demos conta que a escolha de um termo "guarda-chuva," como profissão ou ocupação, poderia apresentar complicações. Antes da equipe realizar a leitura das 43 obras que comporiam a coleção dourada e que depois passariam pelos sistemas que testariam as identificações, vários membros do grupo realizaram a leitura de quatro obras: *Dom Casmurro* (1899), *As Pupilas do Senhor Reitor* , *Quincas Borba* e *Dramas da Côrte* (1905). Essas leituras foram seguidas de

discussões que nos ajudaram a perceber o tipo de dificuldade que teríamos com as demais obras. Foi o caso, por exemplo, dos escravizados que apareceram em Dom Casmurro e do agregado, personagem que se faz passar por médico homeopata e depois assume que havia mentido. Em ambos os casos, decidiu-se rapidamente que não caberia classificar os escravizados ou o agregado como profissão ou ocupação, mas como algo que desse conta da condição social dos mesmos. Foi então que pensamos no termo estatuto social.

Em *Dramas da Côrte*, por sua vez, percebemos rapidamente outro desafio que os romances históricos, haviam vários do século XIX na nossa coleção, apresentariam: as personagens que possuíam títulos da nobreza ou da realeza ("príncipe", "duque", "conde", "baronesa"). Pensamos que igualmente o termo estatuto social poderia dar conta.

Restava, ainda, a diferença entre o que consideraríamos profissão ou ocupação. Havia muitos casos como "colecionador", "ladrão" ou "curandeiro." Esses pareciam ser mais atividades que as personagens exerciam para ocupar o seu tempo e que, por terem recebido a mesma, eram informações relevantes para o papel da personagem no desenvolvimento da narrativa.

Foi assim que decidimos arranjar os dados nesses três termos: profissão, ocupação e estatuto social. Para dar embasamento a essas escolhas, confirmamos a definição desses termos nos dicionários. Iniciando por profissão, na Infopédia, encontramos a seguinte definição: "exercício habitual de uma atividade econômica como meio de vida; ofício; mister; emprego; ocupação." No Dicionário de usos do Português Borba (2002), encontramos uma definição que complementava ainda este sentido "atividade ou ocupação especializada, e que supõe determinado preparo; ofício."

Já o termo ocupação apresentava como sinônimo também o termo profissão, mas apresentava a possibilidade de ser igualmente uma "atividade ou serviço em que se gasta algum tempo," o que ia ao encontro do que pensávamos sobre a distinção entre profissão e ocupação. Para terminar, procuramos a definição de "estatuto" e encontramos em uma das possibilidades "situação social; condição ou posição hierárquica; status." Decidimos, assim, que "estatuto social" daria conta de classificações ligadas a títulos de nobreza e funções dentro de um reino. Deixamos ainda uma quarta possibilidade para aquelas classificações que tinham sido realizadas de forma equivocada, agrupadas no termo NP, não profissão. Além disso, com relação às personagens que tinham tido uma profissão identificada, dividimos as mesmas em tipos, a fim de verificar se haveria grupos de profissões que eram mais frequentes e, também, grupos de profissões que eram mais comuns para personagens femininas ou masculinas. Dividimos as mesmas em profissões serviçais (PS), por exemplo "camareira", "ama", "cabra", "motorista"; liberais (PL), como "padeiro", "peixeiro", "pianista"; militares (PM), como "general", "infante" e "marechal"; e, por fim, as religiosas (PR) ao exemplo de "vigário", "arcebispo" e "padre."

A partir dessas classificações, passamos a aplicar os sistemas para obter alguns resultados que poderiam nos indicar algumas características de cada época ou literatura. Por exemplo, seria alguma profissão, ocupação ou estatuto social (POES) mais comum para as mulheres que para os homens? Haveria alguma diferença nas POES identificadas na literatura brasileira e portuguesa? Quais profissões eram mais frequentes em cada século?

Para responder a essas questões, julgamos que também seria interessante classificar as próprias profissões em tipos: profissões liberais, profissões militares, profissões religiosas e pessoal serviçal. Comentaremos os resultados dos sistemas aplicados tanto na CD quanto no PALAVRAS na Seção 4 "Alguns resultados do DIP."

2. Possibilidades e ganhos na identificação do gênero e das POES por sistemas automáticos

A primeira evidente vantagem do uso de sistemas automáticos na identificação de gênero e das POES das personagens é a velocidade da máquina. Imensamente maior do que aquela que os leitores humanos podem fazer, ainda que se tome aqui um grupo numeroso e experiente de pesquisadores.

A segunda grande qualidade desta abordagem, decorrente da primeira, é que ela permite que se identifique o gênero e a POES das personagens em um grande conjunto de dados, possibilitando aos pesquisadores que se concentrem em outros aspectos, tais como a análise do resultado encontrado. Isso significa não apenas uma maior amplitude do corpus a ser pesquisado, como um aprofundamento da análise em relação a um dado período, estilo de época ou conjunto de autores (possibilitando a comparação entre autores de países diversos, por exemplo).

É importante notar que a virada do estudo em larga escala proporcionada pela análise quantitativa dos sistemas computacionais gerou no-

vas formas de abordagens críticas, antes inimaginadas pela teoria literária (Piper et al., 2017). Tais análises demandam diversas formas de investigação e modelos, com a utilização de métricas, mapas, árvores e padrões sistematizados, menos comuns dentro da teoria literária até os anos 2000 (Moretti, 2005). Ou seja, o uso sistemático desta forma de abordagem necessariamente traz amplitude e um olhar diverso sobre as obras de um dado período ou região.

Além disso, a capacidade de identificar personagens automaticamente pode ter implicações significativas para os estudos literários, tais como a facilitação na análise da evolução da personagem na narrativa, assim como das suas relações, possibilitando estudos em larga escala de tendências e padrões, o que certamente abre novas avenidas para a crítica literária lusófona.

Neste sentido, o resultado desta abordagem pode ser utilizado para identificar tendências em relação às redes sociais e aos ambientes em que dados gêneros são descritos no texto, tais como quais personagens são mais centrais para a narrativa, quais são mais interconectadas do que outras etc. Associado a outros tópicos explorados também pelo DIP, como as relações familiares e a ocupação, estes insights fornecidos pelos sistemas automáticos podem contribuir para se lançar um novo olhar para a estrutura e a dinâmica das narrativas literárias.

3. Desafios

Em um primeiro momento, é preciso discutir a complexidade da tarefa de identificar uma personagem através de sistemas automáticos. Uma distinção primordial é entre os atos de ≪identificar≫ e ≪perceber≫, fundamental no que se refere a esta tarefa, já que o leitor humano realiza as duas operações simultaneamente. No Dicionário Online de Português,[1] a palavra identificar tem as seguintes definições: "conseguir comprovar ou definir a identidade de; saber quem é", "distinguir ou ter a capacidade de reconhecer (alguém ou alguma coisa)", enquanto perceber é definido assim: "entender o significado de algo através da inteligência". Falando de modo muito simplificado, a percepção implica a pré-existência de um dado objeto, enquanto a representação (literária) refere-se necessariamente a um elemento ausente, mas que aparece em cena graças a ela. A percepção de uma personagem literária ocorre quando o leitor produz uma imagem mental dela, ou seja, a personagem literária não existe a priori, tampouco ela é uma pura criação do seu autor, mas ela é necessariamente formada a partir da percepção do leitor, ela é sujeita a uma perspectiva, a um ponto de vista, a um "equipamento" intelectual e cultural, como observa Jouve (1998):

> É preciso observar que a identidade da personagem apenas pode ser concebida como o resultado de uma cooperação produtiva entre o texto e o sujeito leitor. O romance não dispõe dos meios, por conta própria, de dar uma percepção global da personagem. As razões são claramente formuladas por Umberto Eco na sua análise dos mundos narrativos. O universo induzido por um romance se caracteriza, em efeito, por uma ausência de autonomia, na medida em que: 1. de um ponto de vista formal, o texto não pode descrever exaustivamente um mundo; 2. de um ponto de vista semiótico, é inimaginável: a. estabelecer um mundo alternativo completo; b. Descrever como completo o mundo "real". Os universos narrativos, incapazes de constituir por eles mesmos os mundos possíveis, são obrigados de tomar emprestado algumas de suas propriedades do mundo de referência do leitor.[2]

Por consequência, uma personagem não é apenas um ator romanesco criado pelo autor, mas ela é também tudo que o leitor lhe atribui. Em uma pesquisa que pode se aproximar do DIP em alguns aspectos (embora com objetivos diversos), intitulada "Extração e análise de redes de personagens ficcionais", Labatut & Bost (2019) se depararam com alguns dilemas, no que diz respeito à identificação de personagens em literatura a partir de sistemas automáticos, em comparação à mesma tarefa realizada com textos não ficcionais, que seria interessante contemplar aqui. Segundo os autores, o mais desenvolvido programa

[1] https://www.dicio.com.br/identificar/ Acesso em 15/06/2023.

[2] Il convient de remarquer que l'identité du personnage ne peut se concevoir que comme le résultat d'une coopération productive entre le texte et le sujet lisant. Le roman n'a pas, à lui seul, les moyens de donner une perception globale du personnage. Les raisons en sont clairement formulées par Umberto Eco dans sons analyse des mondes narratifs. L'univers induit par un roman se caractérise, en effet, par une absence d'autonomie dans la mesure où 1/ d'un point de vue formel, le texte ne peut décrire exhaustivement un monde ; 2/ d'un point de vue sémiotique, il est inimaginable a/ d'établir un monde alternatif complet, b/ de décrire comme complet le monde ≪ réel ≫. Les univers narratifs, incapables de constituer par eux-mêmes des mondes possibles, sont obligés d'emprunter certaines de leurs propriétés au monde de référence du lecteur.

não pode produzir uma análise significativa de um texto literário, pois os dados precisam ser analisados por humanos (e, em particular, especialistas na área em questão). Além do mais, a intervenção humana é também necessária antes da execução da máquina, pois muitas vezes é preciso "preparar" os dados a serem analisados pelo computador. O que está completamente alinhado aos postulados do DIP, uma vez que a organização do desafio contava não apenas com especialistas na área de informática e processamento de linguagem natural, mas também com pesquisadores da área de literatura. Ainda segundo os autores Labatut & Bost (2019):

> Nos textos, a prosa literária é considerada mais complexa do que a prosa jornalística, ainda mais quando a obra é mais antiga. [...] Por exemplo, para detecção de personagens em romances: muitos personagens são parentes e compartilham o mesmo sobrenome; eles carregam apelidos; alguns personagens fictícios são objetos inanimados na vida real; escritores usam honoríficos específicos correspondentes a convenções sociais complexas, possivelmente desatualizadas e até imaginárias; e eles criam nomes para transmitir certo significado ou função. Para resolução de co-referência, o problema está relacionado a sentenças mais longas, uso mais frequente de pronomes e discurso direto, cadeias de co-referência mais numerosas e curtas.[3]

Além disso, há alguns outros desafios no uso de sistemas de computador para identificar o gênero na literatura. Um deles é a ambiguidade da identificação. Muitas obras literárias apresentam personagens que possuem características que podem ser interpretadas tanto como masculinas quanto femininas, ou que desafiam intencionalmente as normas de gênero. Nesses casos, os modelos de aprendizado de máquina e os que utilizam regras podem ter dificuldades para identificar com precisão o gênero da personagem.

Por exemplo, no livro *Ensaio sobre a Cegueira* (1995), do escritor português José Saramago (1922–2010), o gênero de várias personagens não é explicitamente declarado, deixando-o aberto à interpretação. Essa ambiguidade, também presente na leitura humana, pode tornar desafiador para os sistemas identificarem com precisão o gênero das personagens literárias.

Outro importante desafio de identificação das personagens é a variabilidade dos nomes e referências. As personagens podem ser referidas por nomes ou apelidos diferentes no decorrer da narrativa e podem ser identificadas através de inúmeras referências indiretas, tais como pronomes ou descrições. Além disso, se a única referência ao gênero da personagem for o nome e esta tiver um nome ambíguo, tal como ocorre com certos nomes de personagens indígenas, então não é possível determinar com precisão o gênero de uma personagem. Pensando nesses casos, que são ambíguos tanto para sistemas quanto para leitores reais, o DIP estabeleceu uma categoria em que uma personagem poderia ter os dois gêneros ou mesmo gênero nenhum (0). Os sistemas que poderiam concorrer ao desafio tiveram acesso a esse tipo de informação através das Perguntas Técnicas, disponibilizadas no site da avaliação conjunta.[4]

Do mesmo modo, em língua portuguesa, há inúmeras palavras que, sendo atributos de uma personagem ou a sua definição, não permitem estabelecer um gênero preciso. Este é o caso do substantivo sobrecomum, que é um substantivo uniforme, que apresenta apenas um termo para os dois gêneros que existem em língua portuguesa (masculino e feminino). São palavras tais como: a criança (para menino ou menina), o anjo (Maria é um anjo), cônjuge, defunto (o defunto era Maria), a estrela (de cinema), pessoa, monstro, testemunha, vítima, etc.

Além disso, as personagens podem ser introduzidas gradualmente no decorrer da narrativa e as suas relações evoluem, apresentando igualmente um caráter ambíguo e aberto à interpretação. Em outras palavras, como notou Mark Algee-Hewitt em sua pesquisa sobre a identificação sistemática de personagens dramáticas da literatura inglesa, é importante que o pesquisador entenda o que se deve medir através das análises quantitativas (Piper et al., 2017). O gênero das personagens, por exemplo, é um dos aspectos a se considerar quando se propõe a investigar as redes de relacionamento constituídas na narrativa.

Um outro aspecto limitante a se levar em

[3] In texts, literary prose is considered as more complex than journalistic prose, and even more so when the work is older. [...] For instance, for character detection in novels: many characters are relatives and share the same last name; they bear nicknames; some fictional characters are inanimate objects in real life; writers use specific honorifics corresponding to complex, possibly outdated, and even imaginary social conventions; and they craft names to convey certain meaning or function. For co-reference resolution, the problem comes from longer sentences, more frequent use of pronouns and direct speech, more numerous and shorter co-reference chains.

[4] https://www.linguateca.pt/aval_conjunta/dip/pjr.html#ptecnicas

conta é a interseccionalidade de identidade. Personagens literárias podem ter múltiplas intersecções identitárias, tais como raça e classe social, além do gênero. Modelos de pesquisa automática que apenas identifiquem o gênero, por exemplo, podem deixar escapar importantes aspectos de uma personagem que podem afetar a leitura dela.

Por exemplo, no romance brasileiro *O Bom-Crioulo* (1895), de Adolfo Caminha, a personagem Amaro é um marinheiro, negro, ex-escravizado, gay, descrito principalmente pela sua força física, "uma massa bruta de músculos", "um animal inteiro era o que ele era!". Um sistema automático que apenas identificasse o gênero de Amaro poderia perder uma importante interseccionalidade da identidade da personagem e o modo como isto afeta a sua representação no texto. Portanto, é de se destacar que apenas a identificação de um aspecto da personagem traz necessariamente um resultado incompleto e precisa ser colocado em relação a outros eixos que constituem a personagem ficcional. Essa foi a abordagem do DIP.

4. Alguns resultados do DIP

Descrevemos aqui os resultados obtidos por meio do único sistema que se candidatou à avaliação conjunta, o PALAVRAS-DIP (Bick, 2023), dando destaque para o gênero e a POES. Para outras categorias analisadas, conferir os outros artigos do volume.

4.1. Alguns resultados do DIP: o gênero das personagens

Confirmando aquilo que já conhecíamos em virtude da nossa prática de leitura próxima de literatura, a imensa maioria das obras literárias presentes no DIP traz um número maior de personagens masculinas, ou seja, mesmo observando um número considerável de textos, muitos dos quais distantes do cânone e do conhecimento de pesquisadores experientes da área de Letras, a constância da presença masculina nas obras literárias é mantida.

Mesmo em obras com um grande apelo a uma personagem feminina, como *Úrsula* (1859), *A Escrava Isaura* (1875), *A viuvinha* (1857), entre outras, a constância se mantém. O que reflete não só a realidade histórica, com o destaque do sujeito masculino quantitativa e qualitativamente, esse último, no que diz respeito à profissão, ocupação social ou estatuto social, como reforça a estereotipia de gênero, uma vez que a representação numérica masculina superior carrega, também, a ideia de que os homens são mais importantes e desenvolvem mais papeis do que as mulheres. Um estudo das personagens protagonistas, muito provavelmente, confirmaria o que aqui se levanta enquanto hipótese sobre o número total de personagens.

O PALAVRAS-DIP também identificou que em obras publicadas mais recentemente, continua havendo uma menção maior às personagens masculinas, ainda que essa diferença pareça diminuir, o que apenas seria possível de comprovar se tivéssemos no corpus um número maior de obras contemporâneas (o que pode ser feito em trabalhos futuros).

A nossa hipótese para o decréscimo no número de personagens em obras mais recentes está ligada ao gênero do texto: os romances históricos, muito comuns no século XIX e com uma grande quantidade de personagens, perdem força nos séculos seguintes, dando lugar a romances realistas/naturalistas, em que o foco recaía sobre os tipos sociais (Lukács, 2000 [1916]).

O gráfico abaixo, por exemplo, demonstra o peso que os romances históricos têm quando tomamos como base o número de personagens. Autores como Antônio José Coelho Lousada, Zeferino Norberto Gonçalves Brandão e Diogo de Macedo puxam para cima o número de personagens nas obras portuguesas, o mesmo acontecendo com a obra do brasileiro Franklin Távora.[5] Ademais, para alguns comentários sobre a escolha das obras, consultar Santos et al. (2023) neste mesmo volume.

Do ponto de vista da análise histórico-literária, alguns motivos justificam o grande número de personagens nos romances históricos, entre eles:

- A retratação de períodos históricos significativos para uma nação envolve uma gama ampla de atores (políticos, heróis nacionais, a realeza, as classes camponesas e militares, figuras importantes para o período, etc.);
- O tempo narrado costuma ser longo, o que, na maioria das vezes, envolve distintas gerações de personagens;
- A presença de intrigas políticas/culturais/ religiosas/étnicas que enriquecem a narrativa e ajudam a dar complexidade ao enredo.

[5] A lista completa com as obras usadas pelo DIP está em https://www.linguateca.pt/aval_conjunta/dip/colecao.html.

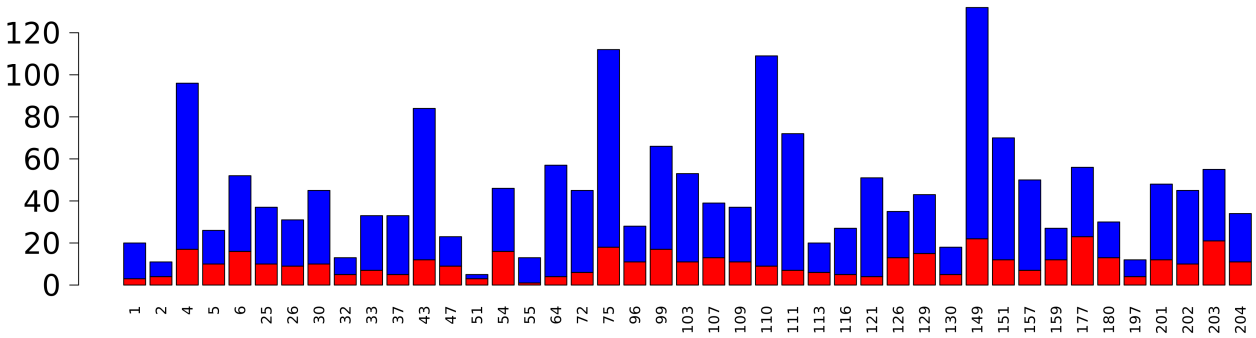

Figura 1: A distribuição de personagens por género na coleção dourada total: a vermelho, as personagens femininas; a azul, as masculinas

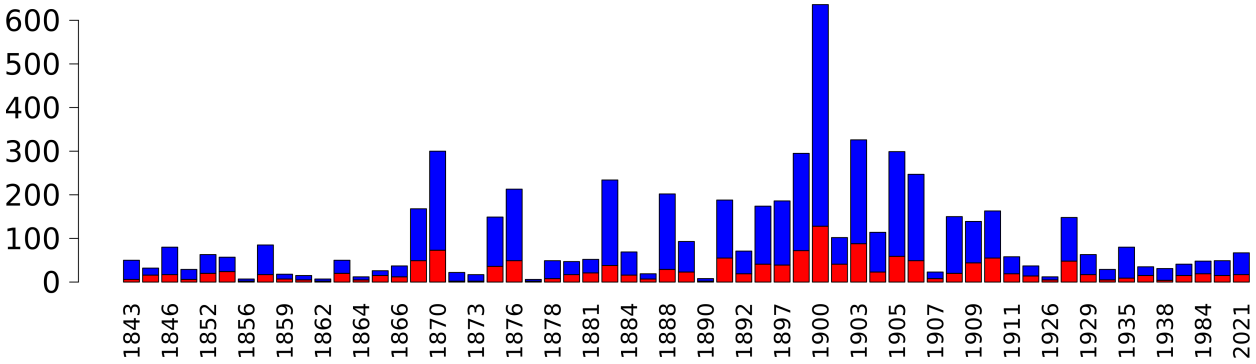

Figura 2: Distribuição do gênero das personagens ao longo do tempo, na resposta do PALAVRAS-DIP

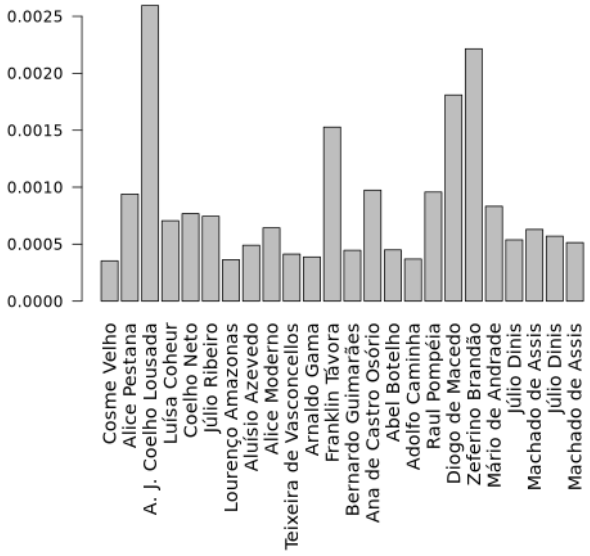

Figura 3: A densidade relativa do número de personagens por obra, na coleção dourada total

A segunda constatação, a diminuição não só de personagens, mas na distância entre o número de personagens masculinas e femininas nas obras mais recentes, pode ser explicada pela expansão social dos direitos femininos ao longo dos séculos XIX e XX, ainda que à personagem feminina recai criada como a POES mais frequente, reforçando a ideia de que a identidade de gênero está diretamente relacionada com outras esferas (estatal, institucional, trabalhista, educativa, doméstica, afetiva, sexual) (Carson, 1995).

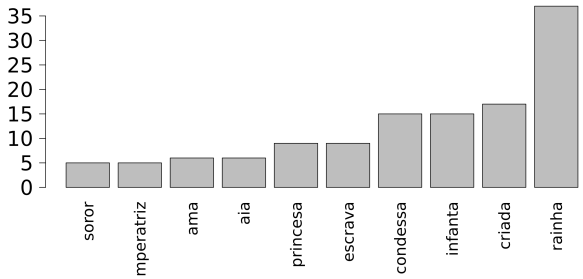

Figura 4: Profissões femininas com mais de 4 ocorrências, na resposta do PALAVRAS-DIP.

Mesmo que os autores, homens na sua grande maioria, tivessem as mulheres como provável público leitor, real ou imaginado, as personagens femininas das obras raramente conquistam posições de prestígio social, e, quando isso é feito, estão sujeitas a tios, maridos ou outra figura masculina, como no caso da personagem Aurélia Camargo, da obra *Senhora* (1875), de José de Alen-

car. Sobre esse público leitor feminino, Candido (2011, p. 94) afirma que:

> Daí um amaneiramento bastante acentuado que pegou em muito estilo; um tom de crônica, de fácil humorismo, de pieguice, que está em Macedo, Alencar e até Machado de Assis. Poucas literaturas terão sofrido tanto quanto a nossa, em seus melhores níveis, esta influência caseira e dengosa, que leva o escritor a prefigurar um público feminino e a ele se ajustar

Ou seja, apesar do leitor intencionado ser a figura feminina, as personagens masculinas são, não só as com maior peso numérico nas narrativas, mas, também, as que desenvolvem papéis de maior destaque social, relegando à mulher o ambiente doméstico.

Ainda no que diz respeito ao tamanho das obras e sua relação com a quantidade de personagens, a pesquisa realizada no contexto do DIP revela a tendência, imaginada, mas demonstrada agora de forma empírica em um grande número de obras, que o número de personagens tende a aumentar à medida que as obras se tornam mais extensas, mesmo que não se possa falar em uma tendência absoluta, uma vez que o gráfico abaixo nos demonstra obras que não apresentam esse padrão (*outliers*).

Também percebemos que as obras portuguesas não só possuem um número maior de personagens como contam com mais personagens femininas que as obras brasileiras, ainda que se repitam aqui POES de menor prestígio social. Como é muito raro um estudo sobre as obras portuguesas que levem em conta um número considerável de obras para, daí, tirar as suas conclusões, os resultados do DIP podem, por exemplo, ajudar a confirmar ou não o que diz Barreira (1986) sobre a personagem portuguesa oitocentista (tomando em conta apenas dois autores, Almeida Garret e Eça de Queiroz). Para a autora, a personagem portuguesa desse período é, em geral, fútil, vazia, quase ridícula, além de estar submissa aos caprichos e desejos masculinos.

4.2. Alguns resultados do DIP: a POES

Nesses primeiros dois gráficos (figuras 8 e 9) relativos a POES, vemos que tanto a classificação manual (CD) quanto a automática (PALAVRAS-DIP) identificaram em maior número ocupações que eram profissões que outros tipos de classificações, chegando ambas as avaliações em resultados muito semelhantes. Esse tipo de resultado nos dá confiança na aplicação dos sistemas, desde que direcionemos os mesmos de forma correta.

No grupo de figuras 10 a 15 vemos primeiramente que ambas as classificações chegaram a resultados muito semelhantes no que diz respeito ao volume de profissões classificadas em cada categoria (PL, PS, PM e PR). O que podemos perceber, e aí analisando a POES juntamente com o gênero, é a clara diferença entre as POES masculinas e femininas, tanto na coleção dourada, que foi atribuída manualmente, quanto pelo PALAVRAS-DIP. Se olharmos para profissões classificadas como serviçais (PS), ao exemplo de criada, vemos que é a POES feminina mais mencionada tanto pelo PALAVRAS-DIP quanto na CD. As outras posições com o maior número de ocorrências dizem respeito a POES de mais prestígio e/ou poder de controle social, como padres, reis, capitães e coronéis e são claramente mais atribuídas a personagens masculinas.

5. Conclusões e apontamentos para o futuro

Conforme visto no DIP, o uso de sistemas computacionais para a identificação do gênero e da POES em literaturas de língua portuguesa apresenta diversas vantagens e desafios. Enquanto a ambiguidade dos atributos, principalmente na identificação do gênero, e a falta de uma grande quantidade de dados anotados podem significar desafios consideráveis, a velocidade, a objetividade, a amplitude e o poder de análise que os sistemas computacionais apresentam os tornam uma ferramenta imprescindível para a pesquisa literária. Em síntese, o gênero das personagens pode ser colocado em perspectiva em relação a outros marcadores do DIP, tais como o gênero e as relações familiares, o gênero e as profissões/estatutos sociais (como demonstramos neste artigo), o gênero e os nomes (nomes que se repetem por gerações, por exemplo).

Com o avanço das pesquisas dedicadas à leitura distante e o desenvolvimento de ferramentas de pesquisa automáticas, esta abordagem na identificação dos diversos atributos de uma personagem literária irá cada vez mais proporcionar novas possibilidades e perspectivas à crítica que se refere à questão do gênero na representação literária lusófona. Proposta para o qual esta iniciativa do DIP vem a ser uma importante etapa fundadora.

Além das dificuldades inerentes à execução da tarefa, cabe ainda abordar a pertinência de tais estudos dentro do âmbito da literatura. Ora,

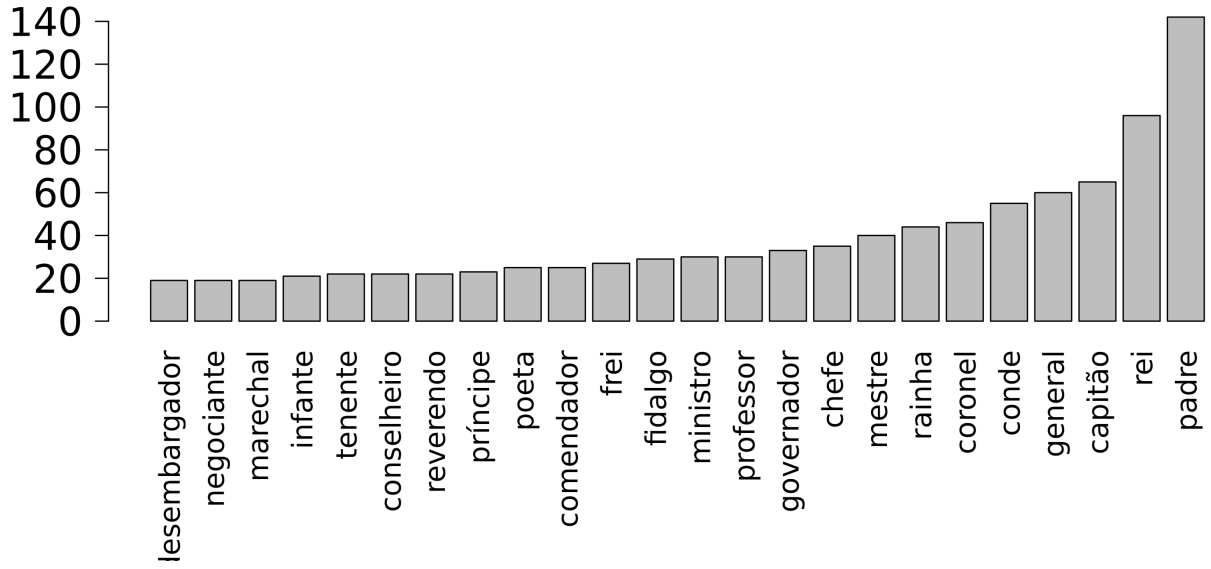

Figura 5: Profissões masculinas e femininas com mais de 18 ocorrências, na resposta do PALAVRAS-DIP.

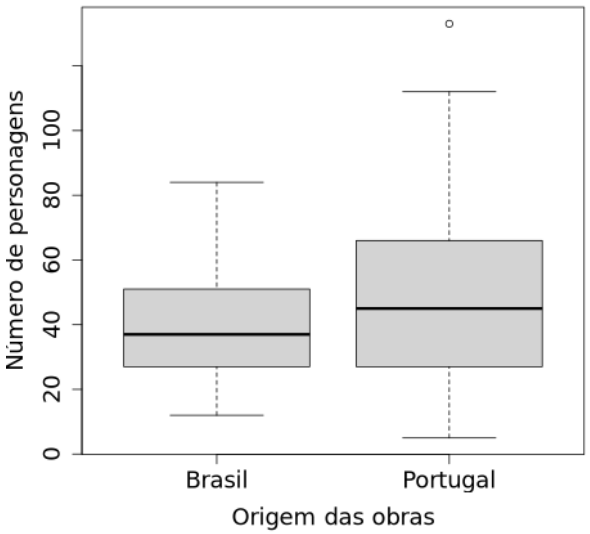

Figura 6: O número de personagens por obra, nas 43 obras brasileiras e portuguesas a que atribuímos uma solução.

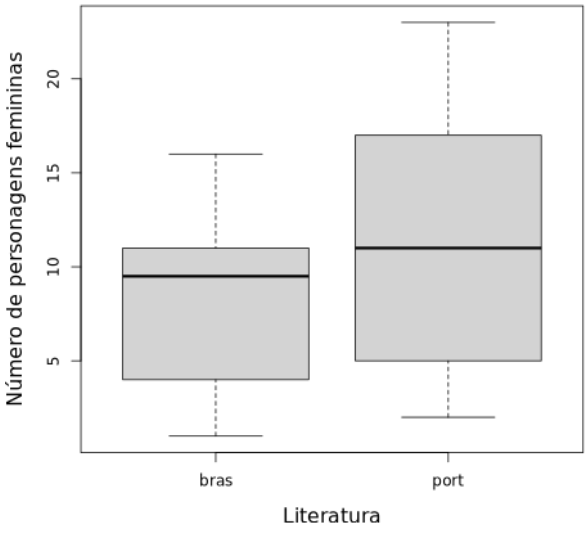

Figura 7: O número de personagens femininas por obra, por literatura.

neste aspecto, Véronique Parenteau traz uma constatação fundamental. As pessoas que geralmente se interessam pelo uso de sistemas automáticos na análise literária não são especialistas de literatura. Embora haja alguns que atuem de fato como professores de literatura em universidades, a grande maioria que se interessa pelo distant reading são não-especialistas de literatura, mas matemáticos, físicos, engenheiros de TI, psicólogos, linguistas, entre outros. Os teóricos da literatura por seu lado ainda fazem pouco uso da informática em suas pesquisas e frequentemente põem esta iniciativa em questão.

Parenteau (1998) afirma sobre este aspecto:

> A análise dos textos literários pelo computador é marginalizada pelos literatos. A maior parte deles não acredita que a informática possa lhes trazer uma ajuda efetiva nos seus trabalhos e parecem não ter a curiosidade de descobrir as possibilidades desta ferramenta. É preciso dizer que uma boa parte dos textos dentro da área da análise de textos por computação são bastante técnicas e um tanto rebarbativas para quem não é muito familiarizado com as estatísticas e a informática. Por outro lado, os especialistas em análise de textos li-

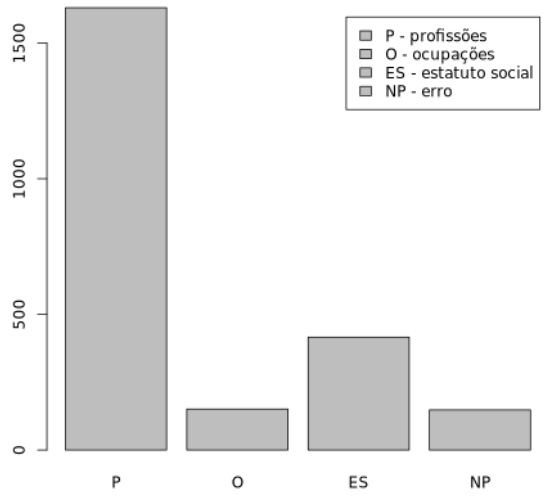

Figura 8: Número de profissões, ocupações e estatutos sociais por grupo na resposta do PALAVRAS-DIP.

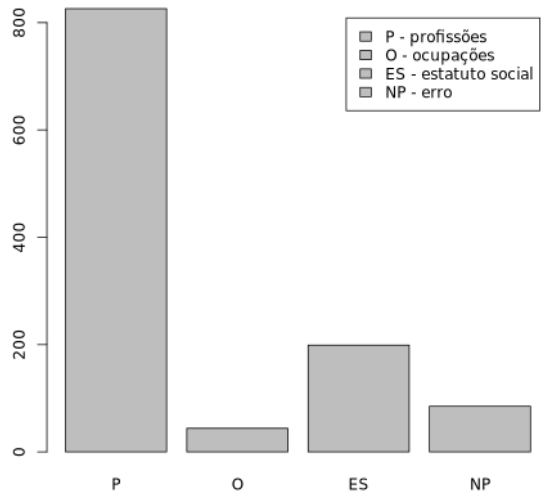

Figura 9: Número de profissões, ocupações e estatutos sociais por grupo na CD total.

> terários assistida por computador não fazem sempre um uso muito pertinente das ferramentas computacionais. Uma grande parte dos estudos se limitam à análise de aspectos muito simples, tais como o tamanho das palavras e das frases, a frequência de certas palavras etc. Em si mesmos, estes resultados de tais análises não são suficientemente interessantes de um ponto de vista estritamente literário. Por outro lado, eles podem ser práticos quando utilizados para fins comparativos, à condição, certo, que a comparação seja pertinente, que o seu autor tenha um objetivo específico.[6]

[6] L'analyse de textes littéraires par ordinateur est marginalisée par les littéraires. La plupart d'entre eux ne croient pas que l'informatique puisse leur apporter une aide réelle dans leurs travaux et ne semblent pas avoir la curiosité de découvrir les possibilités de cet outil. Il faut dire qu'une bonne partie des écrits dans le domaine de l'analyse de textes par ordinateur sont assez techniques et quelquefois rébarbatifs pour qui n'est pas très familier avec les statistiques et l'informatique. D'un autre côté, les experts en analyse de textes littéraires assistée par ordinateur ne font pas toujours un usage très pertinent des outils informatiques. Bien des études se limitent à l'analyse d'aspects très simples comme la longueur des mots et des phrases, la fréquence de certains mots, etc. En eux-mêmes, les résultats de telles analyses ne sont pas très intéressants d'un point de vue strictement littéraire. Par contre, ils peuvent être pratiques lorsqu'utilisés pour fins de comparaison; à la condition, bien sûr, que la comparaison soit pertinente, que son auteur ait un objectif précis.

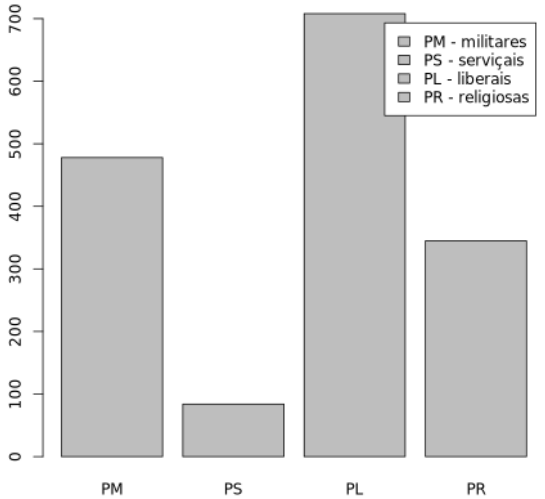

Figura 10: Número de profissões por subgrupo na resposta do PALAVRAS-DIP.

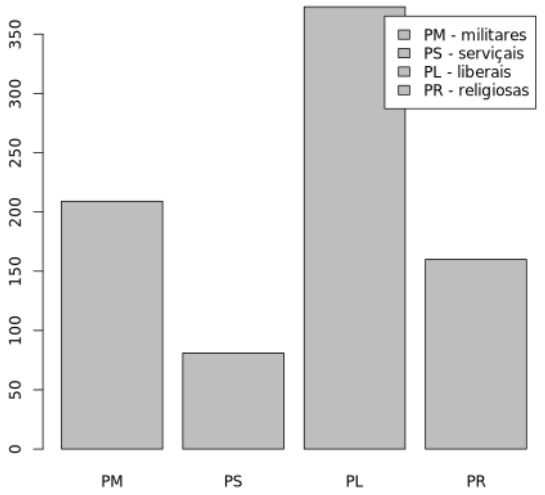

Figura 11: Número de profissões por subgrupo na coleção dourada total.

Dentre as utilidades que Parenteau enumera para o uso da computação nos estudos literários, a comparação é, de fato, a primeira qualidade.

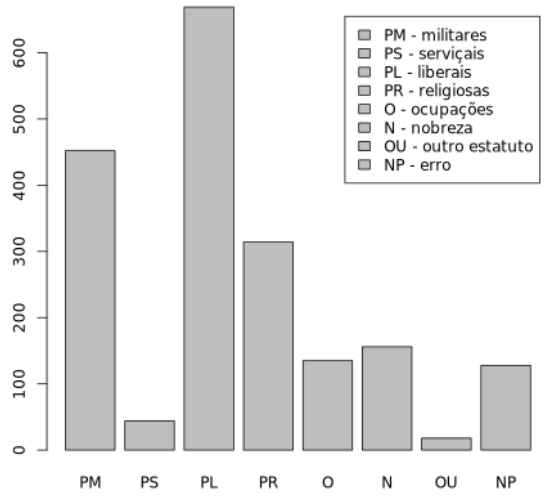

Figura 12: Número de profissões masculinas por subgrupo na resposta do PALAVRAS-DIP.

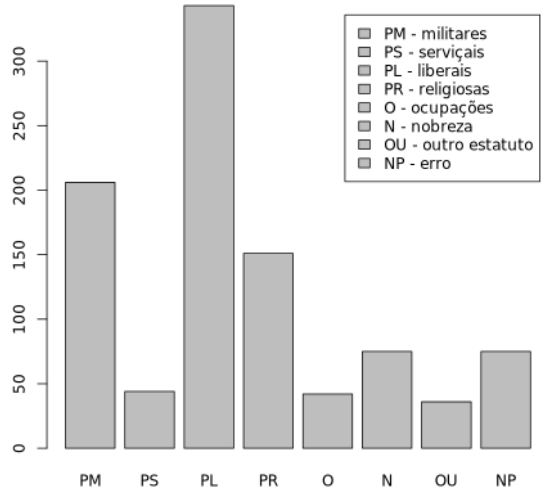

Figura 13: Número de profissões masculinas por subgrupo na coleção dourada total.

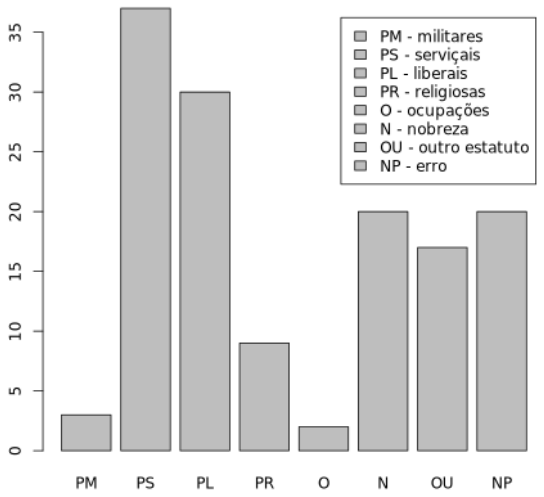

Figura 14: Número de POES femininas por grupo na resposta do PALAVRAS-DIP.

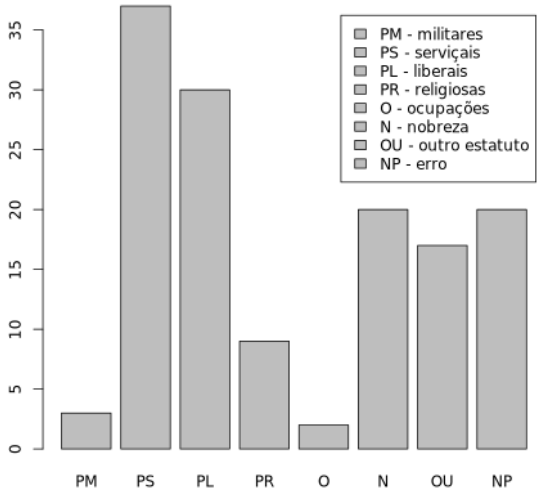

Figura 15: Número de POES femininas por subgrupo na coleção dourada total.

Além desta, ela cita também que estes estudos podem contribuir a: determinar a paternidade de um texto, diferenciar as imitações das obras autênticas, estudar os motivos rítmicos em versos e descobrir as marcas de um autor na evolução da língua. Ou seja, estes são alguns temas que o uso das ferramentas de computação pode contribuir com novas pesquisas.

Seguindo estas indicações de Parenteau, parece-nos que o DIP ainda poderá oferecer muitas oportunidades de novos caminhos dentro dos estudos literários no que se refere às personagens de romances em literatura em língua portuguesa. Vejamos a seguir alguns exemplos que podem vir a ser explorados.

- O nome da personagem como indicativo de informações relevantes à narrativa em questão. Tomemos como exemplo o autor brasileiro Machado de Assis, cuja obra encontra-se em domínio público, digitalizada e disponível, portanto, para o uso de ferramentas computacionais. Em várias de suas obras, os nomes das personagens são indicativos do que irá se desenrolar na narrativa. Por exemplo: em Dom Casmurro (1889), a personagem Capitu seria aquela que capitula? Bentinho seria aquele que crê sem pensamento crítico? Em Memórias póstumas de Brás Cubas (1881), por que a personagem teria o nome de um famoso bandeirante brasileiro? Sendo o bandeirante aquela figura considerada heroica em alguns manuais de história do Brasil que, porém, era um caçador de escravos fugitivos e escravizador de indígenas. Seria coincidência que a personagem de Machado de Assis chicoteia seu escravo desde pequeno e é um tipo de protótipo

de "dono do mundo," no universo machadiano? Será que encontraríamos outros casos de personagens deste tipo, em que o nome tem um duplo sentido com um personagem histórico e com a etimologia das palavras dentro de toda a imensa obra em prosa deste autor? Outra questão interessante é a do livro Esaú e Jacó (1904), em que o título do livro e o nome dos protagonistas — Pedro e Paulo — remetem à bíblia. Teria esta escolha recaído sobre o fato de que o perfil do leitor deste autor seriam mulheres, de classe média, cuja principal leitura, além dos romances, era a bíblia? Este é apenas um caso, mas poderíamos ampliar para outros, como o autor contemporâneo amazonense Milton Hatoum, cuja obra é repleta de nomes de personagens que significam algo na história.

Ora, ocorre que um autor dá nome a seus personagens de um modo diverso como um pai dá a seu filho. Um personagem literário é geralmente "batizado" segundo alguma intenção relacionada à narrativa. Neste sentido, o DIP teria muitas possibilidades a serem exploradas no estudo de inúmeros autores de língua portuguesa.

- Os nomes das personagens como atributos de um dado estilo literário. Aqui pensamos, por exemplo, em um tipo de obra tal como *O Paroara* (1889), de Rodolfo Teófilo, em que a personagem principal da narrativa tem a alcunha de paroara, que dá título ao livro. Vejamos que no caso desta obra, típica do naturalismo brasileiro, os nomes dão frequentemente lugar a apelidos, indicativos de "tipos". Estes tipos por si só sintetizam uma história, como o paroara — aquele retirante do Ceará que emigra para o Amazonas em busca de fazer fortuna na época da borracha e retorna a sua terra natal. Pois bem, haveria muitos outros tipos dentro da literatura brasileira do século XIX? Seria isso uma marca da literatura brasileira ou isso poderia ser encontrado em obras de outros países lusófonos? Está aqui mais uma instigante relação entre narrativa e estudo dos estilos literários que poderia ser explorada com a ajuda do DIP.
- O texto e seu autor — aqui os nomes das personagens, suas relações de parentesco e suas ocupações/posições sociais teriam também muito a revelar. Por exemplo, dentro das relações familiares ou a ocupação das personagens, haveria muitas pistas a explorar sobre as condições de realização da obra literária e das intenções do seu autor. Personagens que têm nomes curtos e simples, gente simples do povo, poderiam indicar, para além do estilo literário (naturalista ou realista), o engajamento do seu autor e a sua intenção em denunciar algum tipo de injustiça social, sua ligação com ideias políticas ou movimentos sociais etc. A simples busca através do sistema desenvolvido pelo DIP possibilitaria traçar um mapeamento deste ramo da literatura dentro do universo lusófono. Inversamente, relações de parentesco com pessoas de classes mais altas ou de profissões de prestígio à época (como o pároco ou o prefeito), todas são informações preciosas e que poderiam ser utilizadas para a análise literária com a ajuda do DIP, que contribuíram muito para uma compreensão mais refinada das obras e de seus autores.

Enfim, assim como indicou Parenteau, haveria certamente muitas outras oportunidades abertas por esta iniciativa precursora dentro dos estudos literários em língua portuguesa no mundo. O fundamental é perceber que o DIP foi uma etapa capital no desenvolvimento de uma nova ferramenta de pesquisa dentro dos estudos literários, que amplia as pesquisas que já vêm sendo realizadas no âmbito das personagens para um número muito maior de obras e que abre novas portas para indagações e caminhos futuros dentro dos estudos de literatura em língua portuguesa.

Agradecimentos

Agradecemos o apoio da FAPEMA pelo financiamento de uma bolsa de pós-doutorado a Emanoel Pires.

Referências

Bamman, David, Ted Underwood & Noah A. Smith. 2014. A Bayesian mixed effects model of literary character. Em *52nd Annual Meeting of the Association for Computational Linguistics*, 370–379. doi 10.3115/v1/P14-1035.

Barreira, Cecília. 1986. Imagens da mulher na literatura portuguesa oitocentista. *Análise Social* 22(92/93). 521–525.

Bick, Eckhard. 2023. Extraction of literary character information in Portuguese. *Linguamática* 15(1). 31–40. doi 10.21814/lm.15.1.397.

Borba, Francisco da Silva. 2002. *Dicionário dos usos do português no Brasil*. Ática.

Candido, Antonio. 2011. *Literatura e sociedade*. Ouro sobre Azul.

Carson, Alejandro Cervantes. 1995. Entrelaçando consensos: reflexões sobre a dimensão social da identidade de gênero da mulher. *Cadernos Pagu* 4. 187–218.

Elsner, Micha. 2012. Character-based kernels for novelistic plot structure. Em *13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 634–644.

Jouve, Vincent. 1998. Le personnage comme produit de l'interaction texte/lecteur. Em *L'effet-personnage dans le roman*, 25–39. Presses universitaires de France.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Langfeldt, Marcia Caetano, Emanoel Pires, Rebeca Schumacher Fuão & Ricardo Gaiotto. 2021. Considerações sobre a personagem literária. https://www.linguateca.pt/aval_conjunta/dip/personagem.html.

Lukács, Georg. 2000 [1916]. *A teoria do romance*. Duas Cidades/Editora 34. Traduzido por José Marcos Mariani de Macedo.

Moretti, Franco. 2005. *Graphs, maps, trees: Abstract models for a literary history*. Verso.

Parenteau, Véronique. 1998. L'analyse de textes littéraires assistée par ordinateur: une introduction. *Cursus* 4(1). em linha.

Piper, Andrew, Mark Algee-Hewitt, Koustuv Sinha, Derek Ruths & Hardik Vala. 2017. Studying literary characters and character networks. Em *Digital Humanities*, 119–121.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP - desafio de identificação de personagens: objectivo, organização, recursos e resultados. *Linguamática* 15(1). 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational Processing of the Portuguese Language (PROPOR)*, 413–419.

Proposta recebida em Abril 2023 e aceite para publicação em Junho 2023.

Avaliação no Desafio de Identificação de Personagens

Evaluation in DIP, the Character Identification Challenge in Portuguese

Roberto Willrich
Departamento de Informática e Estatística
Universidade Federal de Santa Catarina

Diana Santos
Linguateca & ILOS, Universidade de Oslo

Resumo

A primeira edição do Desafio de Identificação de Personagens (DIP) foi uma avaliação conjunta de soluções computacionais para a identificação de personagens em textos literários, bem como a extração de características destas personagens e seus relacionamentos. Para esta avaliação, foi necessária a definição de uma metodologia de avaliação, incluindo a seleção de métricas adequadas ao problema da identificação de personagens em textos literários. Este artigo apresenta uma panorâmica de avaliação na área de identificação de personagens em textos literários, assim como as escolhas concretas que foram realizadas pela comissão organizadora do DIP. Estas escolhas resultaram na definição da metodologia de avaliação do DIP. O uso da metodologia de avaliação proposta é ilustrado pela avaliação da solução candidata submetida ao DIP. Ao final, são apresentadas críticas e sugestões de melhorias à metodologia de avaliação proposta.

Palavras chave

estudos literários computacionais, identificação de personagens, avaliação conjunta, métricas de avaliação

Abstract

The first edition of the Character Identification Challenge (DIP) was a joint evaluation of computational solutions for the identification of characters in literary texts, as well as the extraction of characteristics and relationships of these characters. For this evaluation, it was necessary to define an evaluation methodology, including the selection of appropriate metrics for the problem of identifying characters in literary texts. This article surveys the evaluation methods employed in character identification in literary works and presents the concrete choices done in DIP. These choices resulted in the definition of the DIP evaluation methodology. The use of the proposed evaluation methodology is illustrated by the evaluation of the candidate solution submitted to DIP. We end the paper with some critical remarks and suggestions for improvement.

Keywords

computational studies of literature, shared task, character identification, evaluation metrics

1. Introdução

As personagens, suas características (como género e profissão), relacionamentos familiares, profissionais, e suas interações, são elementos importantes para o estudo de vários tipos de textos literários, como romances e contos. Por exemplo, como apontado por Labatut & Bost (2019), no contexto da análise da narrativa, ou análise narratológica, vários trabalhos realizam a extração manual das personagens e suas interações. Esta operação manual é muito dispendiosa, principalmente caso seja realizada em um corpus muito grande. Neste sentido, diversas iniciativas buscam aplicar técnicas do domínio da inteligência artificial ou outras para a identificação de personagens em textos literários.

Esta identificação de personagens apresenta uma série de desafios. O primeiro é como classificar uma entidade no texto como personagem. Muitos trabalhos (Finkel et al., 2005; Elson et al., 2010; Lee & Yeung, 2012) consideram como personagem apenas pessoas mencionadas no texto. Outros trabalhos, como Valls-Vargas et al. (2014), consideram, além de pessoas, personagens de outras classes, como animais, objetos antropomórficos e criaturas fantásticas. O segundo desafio é que textos literários geralmente utilizam várias formas de mencionar uma mesma personagem, e não apenas pelo seu nome completo, prenome, ou uso apenas do sobrenome (acompanhado ou não da forma de um título pessoal) (de Does et al., 2017). Além destes, textos literários podem se referir a uma personagem de diversas outras formas, como diminutivos do nome, pseudônimos (alcunhas), uso apenas do título pessoal, pronomes pessoais, e descrições nominais (*o lavrador*, *a tal menina*, *o seu professor*). Esta característica torna a identificação de

DOI: 10.21814/lm.15.1.398

personagens em texto literário um grande desafio. Santos et al. (2023) abordam diversos destes desafios em profundidade.

Visando contribuir para o desenvolvimento da área de identificação de personagens, propomos o primeiro Desafio de Identificação de Personagens (DIP)[1] (Santos et al., 2022a). Trata-se de uma avaliação conjunta de soluções computacionais para identificação de personagens de textos literários, bem como a extração de características destas personagens e de seus relacionamentos. O DIP foi organizado de forma conjunta pela Linguateca,[2] NUPILL[3] da Universidade Federal de Santa Catarina (UFSC), Universidade Estadual do Maranhão (UEMA) e Universidade de Oslo (UiO).

Este artigo tem por objetivo apresentar a metodologia de avaliação definida para aferir a qualidade das soluções submetidas ao DIP. Para tal, foi necessária a adoção de métricas de avaliação adequadas para aferir a qualidade de soluções de identificação de personagens, suas características e relacionamentos. De forma a ilustrar o uso da metodologia adotada no DIP, este trabalho apresenta a avaliação da solução submetida à primeira edição do DIP.

O restante deste artigo está organizado na forma que segue. A Seção 2 apresenta um enquadramento teórico acerca das principais etapas envolvidas no processo de identificação de personagens e das métricas clássicas de avaliação. A Seção 3 descreve diversas propostas de solução para identificação (semi-)automatizadas de personagens, suas características e relacionamentos de diferentes tipos. Esta descrição foca principalmente nos processos e métricas de avaliação adotados por estes trabalhos. Em seguida, as Seções 4 e 5 detalham nossa experiência na primeira edição do DIP e a metodologia de avaliação adotada. De forma a ilustrar o uso desta metodologia, a Seção 6 apresenta uma avaliação do sistema participante do DIP. Na sequência, a Seção 7 apresenta uma análise crítica da metodologia adotada, e aponta melhorias a serem realizadas. Finalmente, a Seção 8 apresenta as conclusões e as perspectivas futuras deste trabalho.

2. Enquadramento teórico

Esta seção visa revisar alguns conceitos importantes para o entendimento do problema de identificação de personagens literárias, bem como apresentar as métricas clássicas de avaliação de técnicas de extração de conhecimento envolvendo personagens, suas características, relações pessoais e interações.

2.1. Processo de Identificação de Personagens

As técnicas de identificação de personagens em geral dividem o processo nas seguintes etapas (Labatut & Bost, 2019):

- **Detecção de menções a personagens**
  nesta etapa normalmente são usadas técnicas de Processamento de Linguagem Natural (PLN) do tipo Reconhecimento de Entidades Mencionadas (NER, *Named Entity Recognition*). Estas técnicas permitem identificar e classificar as entidades mencionadas em um texto escrito. Um desafio nesta etapa é classificar corretamente que uma entidade mencionada no texto literário seja classificada como personagem.
- **Co-identificação de personagens**
  esta etapa, também chamada de resolução de correferência, tem por objetivo identificar o conjunto de menções utilizadas no texto que co-identificam uma mesma personagem. A co-identificação de personagem é uma tarefa desafiadora, especialmente dada a multiplicidade de formas que o autor utiliza no texto para se referir às personagens.
- **Extração de características das personagens**
  uma vez identificadas as personagens, alguns trabalhos tentam extrair características destas personagens, como seu tipo (principal/secundária), género, profissão/ocupação/estatuto social, e faixa etária (p.e., bebê, jovem, adulto e idoso). Para isto, podem ser utilizadas técnicas de PLN.
- **Extração de relações entre personagens**
  esta etapa visa extrair os diversos tipos de relacionamentos entre personagens. No domínio de literatura, o objetivo é extrair relações que representem traços de personagens e elementos-chave da narrativa (Chu et al., 2021). Estas relações podem ser do tipo familiares (p.e., pai, mãe, filho, tio, etc.), de sentimentos (p.e., ama, odeia, amigo, inimigo, amante), profissionais (p.e., patrão/empregado, amo/escravo), ou outros tipos relacionados à narrativa (p.e., aliado, membro do clã, traído, assassino).

[1] https://www.linguateca.pt/DIP
[2] https://www.linguateca.pt/
[3] https://nupill.ufsc.br/

- **Extração de interações entre personagens**
  existem diversas iniciativas (incluindo Elson et al. (2010), Lee & Yeung (2012) e Agarwal et al. (2013)) que tentam extrair automaticamente as interações de diálogo entre personagens. O objetivo destes trabalhos é a extração da rede social (também chamada de rede conversacional) da obra, que representa as personagens por nodos e as suas interações por arestas ligando as personagens.

2.2. Métricas de Avaliação

A identificação de personagens literárias é de fato um processo de extração de conhecimento a partir de textos em linguagem natural (obras literárias). Para a avaliação de tais técnicas, é possível utilizar métricas clássicas, como: precisão (*precision*); revocação (*recall*), também chamada de abrangência, sensibilidade ou cobertura; acurácia (*accuracy*), também chamada de acerto; e a medida-F (*f-score* ou *f-measure*), que conjuga as duas primeiras. Estas métricas clássicas podem ser adaptadas para avaliar tarefas do tipo classificação ou de recuperação/recolha de informação.

2.2.1. Métricas do contexto de recuperação de informação

Em geral, os trabalhos relacionados ao DIP adotam as métricas clássicas aplicadas ao contexto da recuperação de informação. Neste contexto, a precisão (P) é determinada pela razão entre o número de instâncias relevantes (*IRel*) que foram recuperadas e todas as instâncias recuperadas (*IRec*), conforme Equação 1.

$$P = \frac{\{IRel\} \cap \{IRec\}}{\{IRec\}} \tag{1}$$

Instância no contexto de identificação de personagens pode se referir às personagens em si, ou então alguma outra informação, por exemplo, um certo fato, como um relacionamento entre duas personagens (por exemplo, Pedro é irmão de Maria).

A revocação R, por sua vez, é a fração de instâncias relevantes que são recuperadas com êxito, e calculada conforme Equação 2.

$$R = \frac{\{IRel\} \cap \{IRec\}}{\{IRrel\}} \tag{2}$$

A *medida-F* é determinada pela média harmônica de precisão e revocação, conforme expressa na Equação 3.

$$F = \frac{2 \times P \times R}{P + R} \tag{3}$$

2.2.2. Métricas do contexto de classificação

Já no contexto de classificação, estas métricas são quantificadas usando os termos verdadeiros-positivos (VP), verdadeiros-negativos (VN), falsos-positivos (FP) e falso-negativos ou falsos-negativos (FN). As equações 4, 5, 6 definem a acurácia, precisão, revocação no contexto de classificação. Para cálculo da medida-F é utilizada a Equação 3.

$$A = \frac{VP + VN}{VP + VN + FP + FN} \tag{4}$$

$$P = \frac{VP}{VP + FP} \tag{5}$$

$$R = \frac{VP}{VP + FN} \tag{6}$$

2.2.3. Métricas de avaliação de resolução de correferências

Como visto na Seção 2.1, a resolução de correferências, no contexto da identificação de personagens, é o processo que permite co-identificar os diversos termos e expressões, chamadas de menções, pelos quais uma personagem (uma mesma entidade mencionada) pode ser referenciada no texto literário. O conjunto de menções a uma mesma personagem no texto é denominado de conjunto de equivalência. Em outras palavras, um conjunto de equivalência contém o grupo de menções no texto que são semanticamente equivalentes e se referem à mesma personagem.

Em geral, a avaliação de técnicas de resolução de correferências é baseada em métricas que permitem comparar os conjuntos de equivalência obtidos por um sistema automático com os conjuntos de equivalência anotados manualmente, chamada aqui de coleção dourada.

Existem diversas métricas de avaliação da resolução de correferências (Pradhan et al., 2014). A primeira métrica de avaliação de sistemas de resolução de correferências foi a MUC (Vilain et al., 1995). Esta métrica mede a eficiência da ligação entre as menções através da Medida-F MUC.

Afim de ilustrar a métrica MUC, considere aqui o seguinte exemplo de ligações de correferência de personagens de uma obra presentes

na coleção dourada: <Pedro – Pedro da Silva, Pedro da Silva – Dr. Silva, Maria – Maria da Silva, Maria da Silva – Mariazinha, Mariazinha – Sra. Silva>. Neste caso, são identificados dois conjuntos de equivalências, definido como K = {Pedro, Pedro da Silva, Dr. Silva}, {Maria, Maria da Silva, Mariazinha, Sra. Silva }.

Também considere as seguintes ligações de correferências preditas por um sistema de identificação de personagens: <Pedro – Pedro da Silva, Dr. Silva – Mariazinha, Maria – Maria da Silva, Maria da Silva – São Paulo>. Neste caso, são preditos três conjuntos de equivalências S = {Pedro, Pedro da Silva}, {Dr. Silva, Mariazinha}, {Maria, Maria da Silva, São Paulo}.

Analisando as ligações de correferência acima, observa-se que o sistema em análise identificou corretamente 2 ligações, e 3 são incorretas, de um total de 5 ligações da coleção dourada. Neste caso, a revocação será de $\frac{2}{5} = 0,4$, enquanto que a precisão será de $\frac{2}{4} = 0,5$. Finalmente, pode-se determinar a medida-F MUC utilizando a equação 3, que resulta em 0,44.

A medida-F MUC tem diversas limitações para avaliação da resolução de correferências (Luo, 2005): ela ignora menções únicas (sem outras menções que co-identifiquem a mesma entidade); ela não permite comparar efetivamente o desempenho de sistemas, pois esta medida favorece sistemas que produzem poucas entidades, e com isto pode resultar medidas-F mais altas para sistemas de pior desempenho.

Surgiram algumas propostas para sanar as limitações do MUC. Uma das primeiras resultou na métrica *B-cubed* (Bagga & Baldwin, 1998), onde a precisão e revocação de um sistema são calculadas com base na precisão e revocação obtidas para cada menção i, conforme formalizado nas equações 7 e 8. Nestas equações, K_i é o conjunto de equivalência da coleção dourada com todas as menções de uma entidade mencionada, que inclui a menção i, e S_i é o conjunto de equivalência predita por um sistema com todas as menções de uma entidade mencionada, que inclui a menção i. Com base nestas definições, a precisão de uma menção i é determinada pela razão entre o número de menções corretas da entidade referenciada por i (presentes em K_i e também na saída do sistema S_i) contendo a menção i e o número de menções em S_i. A revocação de uma menção i é determinada pela razão entre o número de menções corretas da entidade em S_i sobre o número de menções em K_i.

$$P_i = \frac{|S_i \cap K_i|}{|S_i|} \tag{7}$$

$$R_i = \frac{|S_i \cap K_i|}{|K_i|} \tag{8}$$

A medida-F *B-cubed* é calculada com base nas precisões e revocações obtidas pela soma ponderada das medidas calculadas para cada personagem, como formalizado nas equações 9 e 10. O peso associado a cada entidade dependeria da tarefa a ser cumprida pelo sistema. No caso do DIP, de extração de informação, o peso poderia igual para todas as entidades (personagens), ou então personagens mais importantes poderiam ter pesos maiores.

$$P = \frac{\sum_{i=1}^{N_k} w_i \times P_i}{|K|} \tag{9}$$

$$R = \frac{\sum_{i=1}^{N_k} w_i \times R_i}{|K|} \tag{10}$$

A título de exemplo, e usando os mesmos conjuntos K e R da métrica MUC, a precisão usando *B-cubed* resultaria em:

$$\begin{aligned} P_{Pedro\ da\ Silva} = \frac{2}{2}; P_{Pedro} = \frac{2}{2} \\ P_{Dr.\ Silva} = \frac{1}{2}; P_{Mariazinha} = \frac{1}{2} \\ P_{Maria\ da\ Silva} = \frac{2}{4}; P_{Maria} = \frac{2}{4} \\ P_{Sao\ Paulo} = \frac{0}{4}; \\ P = \frac{1}{7}\left(\frac{2}{2} + \frac{2}{2} + \frac{1}{2} + \frac{1}{2} + \frac{2}{4} + \frac{2}{4}\right) \end{aligned} \tag{11}$$

A métrica *B-cubed* também tem suas limitações. Como apontado por Luo (2005), o *B-cubed* calcula a precisão e revocação de menções comparando entidades contendo a menção e portanto uma entidade pode ser usada mais que uma vez. Motivado pelas limitações das métricas *MUC* e *B-cubed*, Luo (2005) propõe a métrica medida-F CEAF (*Constrained Entity-Alignment*), que é calculada com base no melhor mapeamento um-a-um entre as correferências da coleção dourada e da saída gerada pelo sistema.

3. Trabalhos Relacionados

Existem diversas iniciativas que tratam da identificação de personagens literárias. A maior parte dos trabalhos citados aqui visam, além da identificação de personagens, a extração das chamadas redes sociais, ou também chamadas de redes

conversacionais. Estas redes são derivadas a partir das interações de diálogo entre personagens, e são representadas por grafos, onde os nodos representam as personagens e as arestas indicam a existência de diálogo entre duas personagens.

Esta seção apresenta as principais iniciativas de identificação de personagens literárias, detalhando principalmente as métricas adotadas no processo de avaliação. A Tabela 1 sumariza as principais características das iniciativas analisadas, em termos de tipo de informação extraída, os corpora (*datasets*) usados nos experimentos, e as métricas adotadas. As métricas citadas nesta tabela são aquelas adotadas pelas iniciativas para avaliar os resultados das etapas da identificação de personagens apresentadas na Seção 2.1. Alguns trabalhos listados utilizam outras métricas para análise dos grafos gerados, que não são apresentadas por estarem fora do escopo deste artigo.

3.1. Extração de redes sociais

Elson et al. (2010) propõem um método para extração de redes sociais de texto literários. Os textos considerados foram 60 romances britânicos do século XIX. Neste trabalho, os nodos da rede social são ponderados de acordo com o nível de interação da personagem representada pelo nodo com as outras personagens, e as arestas são ponderadas de acordo com o nível de interação entre duas personagens.

Neste trabalho, os autores utilizaram a ferramenta Stanford NER tagger (Finkel et al., 2005) para extrair menções classificadas como pessoas ou organizações. Em seguida, as menções são agrupadas (clusterizadas) em correferentes para a mesma entidade (pessoa ou organização). Para identificar a personagem que está interagindo em cada caso de discurso direto, os autores utilizaram uma técnica de aprendizagem baseada em regras e estatísticas proposta por Elson & McKeown (2010).

A avaliação da proposta foi realizada considerando quatro romances. Para cada livro, foram selecionados aleatoriamente 4 a 5 capítulos. Estes capítulos foram anotados manualmente as personagens e as interações existentes. Para calcular o acordo entre os anotadores, cada anotador atribuiu "sim" ou "não" para cada par de personagens participando de uma interação. A partir disto, foi determinado o coeficiente de concordância de Kappa entre os anotadores.

Para avaliação, este trabalho adotou as métricas precisão, revocação e medida-F na determinação das conversações. Os autores também realizaram extrações de características das redes conversacionais, como o número de personagens e número de personagens que fala, o grau médio do grafo e variação da densidade do grafo, entre outros.

3.2. Extração de redes sociais de pessoas e lugares

Lee & Yeung (2012) também propõem um método para extração de redes sociais de textos literários. Neste trabalho, redes sociais, de forma similar às redes conversacionais adotadas por Elson & McKeown (2010), são grafos onde: os nodos representam personagens do tipo pessoa e suas localizações (nomes dos lugares); arestas entre personagens representam a existência de interação pessoal entre elas; e finalmente arestas entre personagens e lugares são presentes se a pessoa esteve fisicamente presente na localização.

Lee & Yeung (2012) também utilizaram a ferramenta Stanford NER tagger, agora para extração das entidades do corpus adotado (traduções em inglês dos 5 primeiros livros do Velho Testamento) classificadas como pessoas e nomes geográficos. Para resolução de correferência, foi utilizado o sistema *Stanford Deterministic Coreference Resolution* (Lee et al., 2011) e o método de agrupamento proposto por Elson & McKeown (2010) para associar as entidades mencionadas com suas diversas menções.

Para avaliação do método proposto, Lee & Yeung (2012) adotaram uma coleção dourada, onde as arestas personagem-personagem e personagem-localização foram anotadas manualmente. Além disso, o processo de avaliação considerou apenas as personagens principais (mencionadas ao menos 10 vezes no corpus). O trabalho também utilizou as métricas precisão, revocação e medida-F para avaliar de forma independente o reconhecimento das entidades mencionadas, o conjunto de arestas personagens-personagens, e o conjunto de arestas personagem-lugar.

3.3. Extração da rede social de Alice no país das maravilhas

Agarwal et al. (2013) apresentam um procedimento para extração da rede social da obra *Alice no país das maravilhas*. No trabalho, a rede social é definida por um grafo onde os nodos representando personagens e arestas indicam eventos sociais (evento em que duas personagens interagem de forma deliberada e consensual). Para esta tarefa, foi adotado um sistema pré-treinado com um corpus de notícias usando *tree kernel* e SVM (*Support Vector Machines*).

Iniciativa	Informação extraída	Corpus	Métricas
(Elson et al., 2010)	Rede de interações entre personagens	4 romances	Precisão Revocação Medida-F
(Lee & Yeung, 2012)	Rede de interação entre personagens e lugares	5 livros do Velho Testamento	Precisão Revocação Medida-F
(Agarwal et al., 2013)	Rede de interações entre personagens	1 livro	Acurácia Precisão Revocação Medida-F
(Valls-Vargas et al., 2014)	Personagens e seus tipos	Contos folclóricos (269 sentenças)	Acurácia Precisão Revocação
(Groza & Corde, 2015)	Personagens, seus tipos, relacionamentos e papéis	7 contos populares	Acurácia Precisão Revocação Medida-F
(Vala et al., 2015)	Grafo de correferência de personagens	88 obras	Precisão Revocação Medida-F
(Chaturvedi et al., 2017)	Personagens e seus relacionamentos	50 sentenças de romances	Precisão Revocação Medida-F
(Dekker et al., 2019)	Redes de interações de personagens e seus relacionamentos	40 romances	Precisão Revocação Medida-F
(Łajewska & Wróblewska, 2021)	Personagens do tipo pessoa	13 romances	Precisão Revocação Medida-F
(Chu et al., 2021)	Personagens, características e relações	Triplas compiladas dos infoboxes de 142 wikis da fandom.com	Precisão Revocação Medida-F HITs@k MRR
(Srinivasan & Power, 2022)	Personagens e seus tipos	20 sumários de estórias de ficção	Precisão Revocação Medida-F

Tabela 1: Sumário das iniciativas analisadas

Agarwal et al. (2013) avaliam a detecção de eventos sociais e extração da rede social usando as métricas clássicas de acurácia, precisão, revocação e medida-F. Além disso, para avaliar a rede social, são usadas métricas aplicadas à análise de redes sociais para comparar a rede extraída e a rede dourada.

3.4. Identificação de personagens com o sistema *Voz*

citevalls2014toward propõem um método que usa Raciocínio Baseado em Casos (CBR, *Case-Based Reasoning*) para identificar personagens em textos usando o sistema *Voz*. Este sistema primeiro extrai entidades mencionadas do texto e em seguida determina um vetor de características (*feature-vector)* para cada uma delas usando informações linguísticas e conhecimento externo. Este trabalho também propõe uma métrica para determinar a similaridade entre cada entidade com aquela da base de casos (*case-base*), que é utilizada para inferir se a entidade é uma personagem ou não. Este sistema utiliza as ferramentas Stanford CoreNLP[4] para segmentar o texto em sentenças e anotá-las com diversas informações.

[4] https://github.com/stanfordnlp/CoreNLP

Para a realização da avaliação, Valls-Vargas et al. (2014) utilizaram uma coleção de contos folclóricos russos traduzidos para o inglês. Como operação manual, foram retiradas no texto diálogos e passagens onde o narrador dialoga com o leitor diretamente. Ao todo, o corpus foi formado de 269 sentenças segmentadas pela Stanford CoreNLP. Este corpus contém diferentes tipos de personagens (humanos, animais, e objetos antropomórficos e criaturas fantásticas). Para avaliação, foram anotados manualmente 1122 entidades no corpus, sendo 614 personagens e 507 não são personagens. Como métricas, os autores utilizaram acurácia, precisão e revocação.

3.5. Identificação de personagens em contos populares

Groza & Corde (2015) propõem uma técnica de identificação de personagens em contos populares, além da extração dos relacionamentos entre estas personagens e seus papéis no desenvolvimento da estória. A extração do conhecimento dos contos é obtida pela combinação de um módulo de PLN, baseado na ferramenta de engenharia de textos GATE (Bontcheva et al., 2004), e na inferência de ontologia do domínio de contos populares. Esta ontologia é processada usando a OWLAPI (Horridge & Bechhofer, 2011) para a geração das classes de personagens para a GATE. O corpus de contos é analisado visando popular esta ontologia e anotar cada conto com as entidades mencionadas que foram identificadas.

Groza & Corde (2015) não apresentam explicitamente uma definição de personagens de contos populares, mas a ontologia de domínio define diferentes tipos de personagens, como pessoa (com diversas subclasses, como homem, mulher, menino, menina, rei, princesa, etc.), animal (com subclasses do tipo urso, pássaro, cão, e outros), planta, e entes sobrenaturais (como gigante).

Neste trabalho, a técnica de identificação de personagens foi testada usando sete contos populares. Para determinação da acurácia, os autores escolheram manualmente cerca de 3 personagens por conto, totalizando 20 personagens. Estas personagens foram classificadas como personagem principal ou secundária. Groza & Corde (2015) adotaram as métricas acurácia, precisão, revocação e medida-F no contexto de tarefas de classificação, onde: VP representa o número de sentenças que são encontradas tanto no conjunto manualmente anotado como no conjunto de teste; VN representa o número de sentenças que não estão no conjunto anotado manualmente, nem no conjunto de teste; FP representa o número de sentenças que estão no conjunto de teste e não no conjunto manualmente anotado; e FN representa o número de sentenças que estão no conjunto manualmente anotado, mas não no conjunto de teste.

3.6. Técnica de extração de grafos de correferência

Vala et al. (2015) propõem um pipeline de oito estágios para detectar personagens literárias e construir um grafo de correferência. Neste grafo, os nodos representam menções e as arestas entre menções indicam que as menções co-identificam uma mesma personagem. Este trabalho também utiliza as ferramentas Stanford CoreNLP tanto para NER quanto para resolução de correferência.

Para a avaliação, Vala et al. (2015) utilizaram dois conjuntos de dados (*dataset*): uma coleção manualmente anotada de 58 obras com a lista completa de todas as personagens e suas possíveis menções. O segundo conjunto é formado por 30 romances e a lista de suas personagens obtidas do site Sparknotes[5], sendo que foram adicionados manualmente as listas das possíveis menções que co-identificam cada personagem (i.e, os conjuntos de equivalência).

Para avaliação, foram considerados os conjuntos de equivalência gerados pelo sistema proposto, anotada por $E = \{E_1, \ldots, E_n\}$, onde E_i é o conjunto de equivalência contendo as possíveis menções para uma personagem i. Esta lista é confrontada com os conjuntos de equivalência K da coleção dourada contendo todas as menções corretas para cada personagem. Para avaliação, os autores formalizaram o problema em como determinar uma combinação bipartida máxima (*maximum bipartite matching*). Para precisão, a combinação é a medida da pureza de um conjunto de equivalência, E_i, em relação às menções da coleção dourada, K_j: $1 - \frac{|E^i - K^j|}{|E^i|}$. A revocação é definida de forma mais flexível da combinação, com o objetivo de medir se uma personagem K_j foi identificada. Esta combinação é medida como a seguinte função binária: 1 se $E_i \cap K_j \neq \emptyset$, e 0 caso contrário. O trabalho também utiliza a métrica medida-F com base nas definições de precisão e revocação anteriores.

[5] https://www.sparknotes.com/

3.7. Aprendizagem não supervisionada para extração de relações entre personagens

Chaturvedi et al. (2017) utilizam o pipeline BookNLP[6] para obter as tags POS (*Part Of Speech*), análises de dependência, resolução de correferências, e identificação das personagens principais. Em seguida, cada sentença envolvendo personagens é representada por um vetor de características que é associado a um estado latente. Para este último, o problema é tratado como uma tarefa de aprendizado não supervisionado, e os estados latentes são aprendidos utilizando suposições Markovianas para extração do fluxo de informações entre sentenças individuais.

Para avaliação, Chaturvedi et al. (2017) utilizaram 50 sentenças manualmente anotadas, e a métrica adotada foi a média da medida-F.

3.8. Avaliação de ferramentas para identificação de personagens e extração de redes sociais

Dekker et al. (2019) citam que extração de personagens no domínio da literatura é desafiador, pois os nomes não seguem as mesmas "regras" do mundo real, como apontado por de Does et al. (2017). Os autores citam que a resolução de correferência também é um desafio no domínio da literatura, devido à variedade de alcunhas que uma personagem pode receber.

Dekker et al. (2019) avaliam quatro ferramentas NER para identificação de personagens e redes sociais em romances, aferindo o desempenho destas ferramentas frente aos desafios apontados. As ferramentas NER avaliadas foram BookNLP[7], Stanford NER,[8] Illinois tagger[9] e IXA-Pipe-NERC[10].

Para a avaliação das ferramentas, Dekker et al. (2019) utilizaram um corpus de 40 romances clássicos e modernos, obtidos do projeto Gutenberg ou comprados em formato ebook. A avaliação foi realizada com base em uma coleção dourada composta de anotações realizadas manualmente em 10 romances (a partir de 300 sentenças em média selecionadas de cada livro). Durante o processo de identificação manual das personagens, foram adotadas as seguintes regras: ignorar pronomes genéricos (p.e., você, ele); ignorar frases exclamativas (p.e., Por Cristo!); ignorar sintagmas nominais genéricos (p.e., Mário não sabia o que dizer **ao mago**); e incluir personagens não humanas.

Para a avaliação foram utilizadas as métricas de precisão, revocação e medida-F. Os experimentos realizados por Dekker et al. (2019) demonstraram que a ferramenta BookNLP superou o desempenho das demais.

3.9. Anotação de Protagonistas

Łajewska & Wróblewska (2021) propõem a *protagonistTagger*, uma ferramenta para marcação (*tagging*) de personagens do tipo pessoa. O método implementado por esta ferramenta compreende duas etapas: (1) reconhecimento de entidades mencionadas (NER) da classe pessoa; e (2) desambiguação de entidades nomeadas (NED, *Named Entity Disambiguation*). O trabalho utilizou o modelo de linguagem pré-treinada oferecida pela biblioteca em código aberto SpaCy.[11]

Para a realização de experimentos foram considerados um conjunto de dados composto por 1300 sentenças de 13 romances clássicos. O processo de criação deste corpus foi o seguinte: (1) obtenção de um corpus inicial com os textos dos romances sem anotações; (2) obtenção de uma lista com nomes completos de todos os protagonistas (usando um parser da Wikipedia), que são considerados como etiquetas (*tags*) pré-definidas; (3) reconhecimento das entidades mencionadas da classe pessoa usando um modelo NER pré-treinado usando textos manualmente anotados; (4) e uso de um algoritmo de emparelhamento para anotar cada entidade nomeada da categoria pessoa a uma das etiquetas pré-definidas em (2).

Łajewska & Wróblewska (2021) citam que um dos casos mais difíceis de tratar no processo de emparelhamento são os diminutivos e alcunhas. Para tratar este problema, o *protagonistTagger* considera uma lista completa de diminutivos com mais de 3300 diferentes formas de nomes. Outro desafio foi a identificação de entidades mencionadas pelo seu sobrenome. No caso, para identificar se a menção ao sobrenome se refere à toda a família, ou a uma única personagem, foi considerada a palavra precedente ao sobrenome. Se é um título pessoal (por exemplo, *Mr., Mrs., Ms. ou Miss*) a citação identifica uma única pessoa.

Em Łajewska & Wróblewska (2021), os autores utilizaram as métricas clássicas de precisão, revocação e medida-F para avaliar o algoritmo de emparelhamento da etapa (4).

[6] https://github.com/booknlp/booknlp
[7] https://github.com/booknlp/booknlp
[8] https://nlp.stanford.edu/software/CRF-NER.shtml
[9] https://github.com/kordjamshidi/illinois-cogcomp-nlp
[10] https://github.com/ixa-ehu/ixa-pipe-nerc
[11] https://spacy.io/

3.10. KnowFi

Chu et al. (2021) propõem o método *KnowFi* para extração de conhecimento de textos de ficção longos. Este método combina a aprendizagem neural aprimorada por BERT (*Bidirectional Encoder Representations for Transformers*) com o algoritmo de aprendizado com seleção e agregação criteriosa de passagens de texto. Além disso, *KnowFi* determina os tipos semânticos das entidades usando SpaCy, e assim provendo um tipo para as menções (pessoa, nacionalidade/religião, evento, etc.).

Para a avaliação da extração de relações entre personagens, Chu et al. (2021) criaram o corpus LoFiDo *Long Fiction Documents*, composto de triplas Sujeito-Predicado-Objeto (SPO) compiladas dos *infoboxes* de 142 wikis de comunidades fãs em fandom.com. Este corpus especifica 64 relações, como inimigo, amigo, aliado, religião, arma, etc. Para avaliação, foi criada uma coleção dourada checada manualmente.

Chu et al. (2021) utilizaram as métricas clássicas de média de precisão, de revocação e de medida-F. Além destas, foram utilizadas as métricas HITs@k, para avaliar com que frequência um resultado correto aparece entre as k principais extrações por par entidade-relação, e a métrica MRR (*Mean Reciprocal Rank*) para obter a classificação recíproca média da primeira extração.

Neste trabalho foram realizadas duas avaliações: automática e manual. A avaliação manual foi necessária devido à baixa eficiência demonstrada pela avaliação automática, quando foi utilizada uma coleção dourada incompleta, gerada automaticamente. Na avaliação manual, foram selecionadas as top 100 extrações do resultado sobre as quais foram realizadas avaliações manuais para checar a correção de cada uma.

3.11. Extração de Personagens e seus tipos

Srinivasan & Power (2022) propõem uma solução para extração e classificação de personagens a partir de sumários da estória. A extração de personagens e a resolução de correferências foi realizada através da ferramenta StanfordNLP. Usando as tags POS a partir do texto anotado, foram identificados os nomes próprios. Ao final, é aplicado um conjunto de heurísticas para eliminação de entidades que não são personagens, considerando, entre outros, o fato que a entidade não é sujeito de nenhuma sentença, não tem um nome pessoal ou título pessoal, ou cujo nome não é reconhecido pela base de dados WordNet.

A solução proposta permite classificar uma personagem como protagonista, antagonista ou personagem auxiliar. Para esta classificação foram utilizados algoritmos de aprendizado supervisionados. Além do tipo, também foram analisados os pronomes pessoais para a identificação do género das personagens, que foi confirmado através de regras heurísticas.

Para a avaliação, este trabalho adotou um corpus formado por 20 sumários de estórias de ficção que ao total fazem menção a 218 personagens. Como métricas de avaliação, Srinivasan & Power (2022) utilizaram a precisão, revocação, e medida-F para o contexto de sistemas de classificação (Seção 2.2.2). Elas foram calculadas para duas classes: personagens e não-personagens. Para avaliação da extração de personagens, os autores consideraram *VP* aquelas instâncias que foram identificadas corretamente, *FP* aquelas instâncias que foram incorretamente identificadas como personagens, *VN* aquelas instâncias que foram corretamente identificadas como não-personagens, e *FN* aquelas instâncias que a abordagem falhou na identificação delas como personagem apesar delas serem de fato personagens.

Para avaliação da classificação das personagens, os autores também usaram as médias de precisão, revocação e medida-F entre os três tipos de personagens (protagonista, antagonista e suporte).

3.12. Outros sistemas

Conhecemos vários outros sistemas que descrevem tarefas relacionadas com o DIP, mas não os descrevemos aqui por não descreveram a forma como são avaliados: por exemplo Jayakumar et al. (2022) extraem redes sociais dinâmicas, Santos & Freitas (2019) criam redes para a literatura lusófona.

4. O DIP

Muito brevemente, visto que já foi apresentado em várias outros contextos (Santos et al., 2022b,a, 2023), o DIP, Desafio de Identificação de Personagens, é uma tarefa que pretende desenvolver e avaliar programas, que, para uma dada obra literária em português, consiga identificar:

- as personagens nessa obra;
- as relações familiares entre personagens.[12]

[12] Especificamente as seguintes: *mãe, pai, filho, filha, neto, neta, avó, avô, irmã, irmão, cunhado, cunhada,*

As personagens são representadas por um conjunto de menções (ou nomes) que lhes correspondem, pelo seu género, e pela sua profissão/ocupação/estatuto social. É possível que uma personagem tenha várias profissões ao longo da história, ou que nenhuma seja mencionada.

Para este desafio, foram criados dois conjuntos de dados compostos de 200 obras literárias de domínio público, disponibilizados no formato TXT e PDF. Estes conjuntos de dados estão disponíveis no site do DIP.[13] O desafio também definiu um formato de representação dos dados extraídos que seriam extraídos pelas soluções participantes do desafio. Esta padronização foi necessária para simplificar o processo de avaliação.

5. Metodologia de Avaliação Adotada

Para o processo de avaliação, foi necessário criar uma Coleção Dourada (CD). Para tal, todas as personagens, suas diversas menções, atributos e relações de 38 obras foram anotadas manualmente.[14]

No processo de avaliação, os resultados das soluções seriam confrontados com aqueles disponíveis na CD. Para aplicar métricas mais tradicionais, e porque não tínhamos ideia das possíveis interligações entre as diferentes subtarefas, escolhemos avaliar separadamente as cinco subtarefas do DIP, a saber:

- reconhecimento da entidade do tipo personagens mencionados na obra;
- resolução de correferência, para a identificação da conjunto de menções que identificam unicamente uma personagem;
- classificação do género da personagem entre masculino (M), feminino (F) e ambos os sexos (A);
- identificação das possíveis profissões/ocupações/estatutos sociais mencionados na obra;
- extração das relações familiares (um conjunto pré-determinado) entre as personagens

Para avaliar estas cinco tarefas, definimos cinco medidas parcelares, e consideramos a medida final como a média aritmética das cinco medidas (ou das quatro, se a obra avaliada não inclui qualquer relação familiar):

AI avaliação da identificação, usando a medida-F sobre o conjunto de todos os nomes identificados pelo sistema, e coligidos na CD

ACI avaliação da co-identificação, expandindo todos os conjuntos, e usando a medida-F sobre esses conjuntos

AG avaliação do género, entrando em conta com a co-identificação

APOES avaliação da profissão, ocupação e estatuto social, usando uma medida-F adaptada

AR avaliação das relações, também usando uma medida-F

As seções que seguem descrevem como essas medidas são calculadas. Para ilustrar estas métricas, considere os seguintes dados anotados em uma obra fictícia da CD:

- Co-identificações de uma personagens, seus gêneros e profissões: {Bento – Padre Bentinho – Dom Casmurro, M, advogado}, {Capitu – Capitolina, F,}, {Dona Fortunata, F,}, {Thomaz, M, escravo}, {Doutor João da Costa – João da Costa, M, médico};
- Relações entre personagens: (Bento, marido, Capitu),(Dona Fortunata, mãe, Capitu).

Também considere que um sistema de identificação de personagens extraia da mesma obra as seguintes informações:

- Co-identificações de uma personagens, seus gêneros e profissões: {Dom Casmurro, M, escravo - advogado}, {Bento, Padre Bentinho, M}, {Capitu, M,}, {São Paulo, M, }, {Thomaz, M, escravo};
- Relações entre personagens: (Bento, marido, Capitu),(Thomaz, primo, Capitu).

5.1. Avaliação da identificação (AI)

Para avaliar a identificação de personagens com esta métrica, é calculada inicialmente a precisão P_i e revocação R_i de cada obra i usando as equações 12 e 13. Nestas equações S_i é o conjunto de menções a personagens identificadas por um sistema para uma obra i, e K_i é o conjunto de menções presentes na CD para esta mesma obra. A precisão para uma obra i então determinada pela razão entre o número de menções corretamente identificadas pelo sistema nesta obra e o

primo, prima, tio, tia, sobrinho, sobrinha, bisavó, bisavô, bisneto, bisneta, nora, genro, sogro/a, mulher, marido, padrinho, madrinha, compadre, comadre, afilhado e *afilhada.*

[13] https://www.linguateca.pt/aval_conjunta/dip/colecao.html

[14] Disponíveis em https://www.linguateca.pt/aval_conjunta/dip/nova_colecao_dourada/

número total de menções identificadas pelo sistema de uma entidade. Por sua vez, a revocação para uma obra i é determinada pela razão entre o número de menções corretamente identificadas pelo sistema e o número de menções realmente presentes na obra.

$$P_i = \frac{|S_i \cap K_i|}{|S_i|} \tag{12}$$

$$R_i = \frac{|S_i \cap K_i|}{|K_i|} \tag{13}$$

No exemplo ilustrativo, temos as seguintes menções identificadas:

- K_1 = {Bento, Padre Bentinho, Dom Casmurro, Capitu, Capitolina, Dona Fortunata, Thomaz, Doutor João da Costa, João da Costa};
- S_1 = {Dom Casmurro, Bento, Padre Bentinho, Capitu, São Paulo, Thomaz};

Neste caso, $P_1 = \frac{5}{6}$ e $R_1 = \frac{5}{9}$. Sendo assim, a medida-F será de 0,67.

Com base na precisão e revocação de cada obra i, é determinada a média da medida-F_i usando a Equação 3. Ao final, a nossa métrica AI é determinada pela média das Medidas-F obtidas no conjunto de obras N utilizadas para avaliação, conforme Equação 14.

$$AI = \frac{\sum_{i=1}^{N} Medida_{F_i}}{N} \tag{14}$$

Durante o ensaio do DIP e do processo de avaliação, os organizadores e participantes se depararam com diversos desafios, alguns já citados na literatura, e outros. Um caso que tivemos de tratar foi a existência — inesperadamente, em muitas obras — de personagens com o mesmo nome, quer totalmente (dois Franciscos), quer parcialmente (ou seja o Sr. João da Esquina, e o Dr. João Semana podem ser ambos apenas tratados por João em contextos específicos).

Embora estes casos — sobretudo os de sobreposição total — sejam praticamente impossíveis de identificar automaticamente, quisemos que a CD incluísse o número certo (e as formas certas) das personagens, para podermos caracterizar a literatura que era o nosso objeto de estudo.

Assim, a primeira passagem para a avaliação de identificação é a "tradução" de nomes semelhantes em nomes diferentes, para depois aplicar a medida-F, como já indicado.

5.2. Avaliação da co-identificação (ACI)

Para avaliação da co-identificação das personagens, criamos a métrica ACI, que se trata de uma medida-F obtida comparando-se pares de co-identificação das personagens.

Para determinar a ACI, inicialmente o conjunto de co-identificação de cada personagem é convertida em pares de co-identificação. Para tal, inicialmente as menções de cada personagem são organizadas em ordem alfabética. A primeira menção identificará a personagem, e cada par de co-identificação da personagem relaciona esta primeira menção com cada outra do conjunto de menções. Para considerar personagem referenciadas com apenas uma menção, é utilizado o termo ZERO para formar o par com esta menção única. No exemplo ilustrativo, os pares de co-identificação da CD seriam: (Bento – Dom casmurro), (Bento – Padre Bentinho), (Capitolina – Capitu), (Dona Fortunata – ZERO), (Thomaz – ZERO), (Doutor João da Cota – João da Costa). E os pares identificados pelo sistema seriam: (Dom Casmurro – ZERO), (Bento – Padre Bentinho), (Capitu – ZERO), (São Paulo – ZERO), (Thomaz – ZERO).

Na métrica ACI proposta, a precisão e revocação são obtidas confrontando os pares de co-identificação de cada obra i considerando apenas os pares da CD de personagens corretamente identificadas pelo sistema. No exemplo anterior, a CD indica a existência de 5 personagens, e o sistema identificou 3 delas. Uma personagem é identificada pelo sistema se ao menos uma das menções que co-identificam uma personagem na CD esteja presente no resultado. Assim, apenas os seguinte pares da CD serão considerados: (Bento, Dom casmurro), (Bento, Padre Bentinho), (Capitolina – Capitu), (Thomaz – ZERO).

No exemplo ilustrativo, segundo as equações 1 e 2), $P_1 = \frac{2}{5}$ e $R_1 = \frac{2}{4}$. Sendo assim, a medida-F será de 0,44.

A exemplo da métrica AI, a medida ACI será calculada como a média da medida-F obtida considerando as obras da CD.

5.3. Avaliação do género (AG)

Para avaliar a identificação do género de uma personagem proposta pelo sistema, consideramos as suas diferentes menções, e confirmamos com o género presente na CD para cada menção. Para tal, para cada obra i presente na CD é criado um conjunto de pares <menção, género>, designado por KD_i.

Para cada personagem j proposta pelo sistema, usamos as menções que também se encontram na CD e descartamos as não existentes, formando assim o conjunto KS_j. KD_j corresponde aos pares com as mesmas menções na CD.

O resultado da avaliação usa a seguinte função:

$$AG_j = \begin{cases} -1, & \text{se } \bigcup \text{gen}(KD_j) = \{M, F\} \\ -1, & \text{se } \text{gen}(KS_j) \neq \bigcup \text{gen}(KD_j) \\ 1, & \text{se } \text{gen}(KS_j) = \bigcup \text{gen}(KD_j) \\ 0, & \text{se } MF \in \bigcup \text{gen}(KD_j) \end{cases} \quad (15)$$

E o valor para uma obra é a média de AG_j naquela obra, ou seja:

$$AG_i = \frac{\sum_{j=1}^{|numpers|} AG_j}{|numpers|} \quad (16)$$

Ao contrário do que se poderia pensar, esta é a subtarefa cuja avaliação é mais original, devido às várias possibilidades que tivemos de contemplar.

Explanando em português a equação 15, se a personagem contiver géneros distintos segundo a CD (estiver marcada com MF, ambos), não recebe pontuação (note-se que não houve nenhum caso destes nas obras do DIP).

Se o sistema tiver proposto géneros incongruentes, ou seja, tanto M como F, baseados na separação de diferentes nomes de uma mesma personagem, segundo a CD, em personagens diferentes, também conta como -1.

Além disso, recebe 1 ou -1 conforme o género concorde com o da CD ou não.

Atente-se no seguinte exemplo ilustrativo: A partir da CD são extraídos os seguintes conjuntos: {(Bento, M), (Padre Bentinho, M), (Dom Casmurro, M), (Capitu, F), (Capitolina, F), (Thomaz, M), (José Bento, M)}. Já o sistema hipotético identificou as seguintes três personagens com seus géneros: {(Bento|Padre Bentinho|Dom Casmurro|Capitu, M), (Capitolina, F), (Thomaz|José Bento, M)}.

A primeira recebe -1 (porque na CD corresponde a diferentes personagens com géneros diferentes), e as duas segundas recebem 1. O valor de AG será de (-1+2)/3= 0,333.

A métrica AG é determinada pela Equação 17, onde $|G|$ é o número de obras na CD, e AG_i é a avaliação da qualidade da identificação do género das personagens da obra i.

$$AG = \frac{\sum_{i=1}^{|numpers|} AG_i}{|G|} \quad (17)$$

De notar que as nossas escolhas obedeceram ao seguinte princípio: não devemos penalizar, ou avaliar, uma mesma questão mais do que uma vez. Por isso aceitamos (e consideramos corretas, do ponto de vista do género) personagens incorretas, por exemplo juntando várias diferentes, desde que o seu género seja semelhante.

5.4. Avaliação das profissões, ocupações e estatutos sociais (POES)

Neste caso, que podemos considerar um caso de classificação múltipla (pode haver mais de uma POES para uma personagem), também usamos a medida-F, mas apenas relativa às personagens que existem na CD. Ou seja, não penalizamos nem premiamos a atribuição de POES a personagens não existentes. Isto é mais uma aplicação do princípio mencionado na seção anterior. Essa penalização terá sido feita por altura da avaliação da identificação.

Para avaliar a identificação dos POES das personagens, e ao mesmo tempo considerando a existência de diferentes menções identificando cada personagem, consideramos o POES atribuído a cada menção. Para tal, para cada obra i presente na CD são criados dois conjuntos de pares <menção, POES>, anotados por $POESG_i$ e $POESS_i$. $POESG_i$ mantém os pares identificando os POES de cada menção da obra i presente na CD (K_i) e cujas menções também foram identificadas pelo sistema (S_i). $POESS_i$ mantém os pares das menções presentes em K_i, mas agora com os POES identificados pelo sistema.

No exemplo ilustrativo, a partir da CD são extraídos o conjunto $POESG_1$ = {(Bento, advogado), (Padre Bentinho, advogado), (Dom Casmurro, advogado), (Capitu, ZERO), (Thomaz, escravo)}. Já para o sistema hipotético do exemplo $POESS_1$ = {(Bento, ZERO), (Padre Bentinho, ZERO), (Dom Casmurro, escravo), (Dom Casmurro, advogado), (Capitu, ZERO), (Thomaz, escravo)}.

Confrontando os dois conjuntos anteriores, pode-se calcular a precisão e revocação de cada obra. No exemplo, a precisão seria de 3/6 e a revocação de 3/5, e assim a medida-F APOE da obra será de 0,55.

A medida POES da solução é determinada pela média da medida-F POES das obras da CD.

5.5. Avaliação das relações (AR)

A avaliação das relações familiares, anotada por AR, foi de longe a que exigiu um processamento mais complicado, porque foi preciso converter — ou alinhar — os identificadores das personagens propostos pelo sistema para os identificadores na CD, antes de poder avaliar as relações familiares entre as personagens.

É depois necessário expandir as relações para todas as formas possíveis de exprimir a mesma relação, entrando em conta com o género da personagem. Assim, se `X neta Y`, adicionar-se-á `Y avó X` se Y for do género feminino, e Y avô X se for do género masculino.

Após estas duas operações, usamos simplesmente a medida-F sobre o conjunto das relações (expandido).

Há três questões que devemos levantar:

- A contagem faz-se sempre sobre pares de relações (a relação proposta pelo sistema e a usa inversa). Mas há um caso apenas, o de `X viúva Y`, em que não há relação inversa. Esse caso deveria contar a dobrar para ser igualmente premiado ou penalizado.
- Não fazemos uma expansão transitiva, apenas reflexiva. Ou seja, de `A irmão B` e `B irmão C` não concluímos, para efeitos de avaliação, que `A irmão C`. Nem de `D mãe E` e `E mãe F` obtemos `D avó F`. (Mas veja-se Mota & Santos (2023) para outros processamentos e conclusões.)
- No caso de não haver nenhuma relação familiar na CD,[15] não calculamos esta medida de avaliação.

A fim de ilustrar o cálculo da medida AR, considere novamente o exemplo ilustrativo desta seção. Os pares de relação da CD são os seguintes (considerando a menção que identifica a personagem como descrita na Seção 5.2): {(Bento, marido, Capitu)}. E o sistema identificou as seguintes relações: {(Bento, marido, Capitu),(Thomaz, primo, Capitu)}. Neste caso, a precisão será 1/2 e revocação será 1/1.

5.6. Exemplos concretos

Exemplos de aplicação de cada uma destas medidas com base numa resposta fictícia sobre a obra *Dom Casmurro*, de Machado de Assis,[16] podem ser consultados na página associada à avaliação do DIP.[17]

obra	AI	ACI	AG	APOES	AR
001	0,596	0,913	1,000	0,000	0,000
002	0,690	0,750	1,000	0,133	0,000
004	0,711	0,688	0,818	0,370	0,308
005	0,479	0,611	1,000	0,250	0,000
006	0,800	0,854	0,941	0,182	0,222
025	0,455	0,520	0,882	0,143	0,000
026	0,405	0,683	0,647	0,304	0,286
030	0,773	0,708	0,800	0,094	0,250
032	0,704	0,871	0,778	0,200	0,000
033	0,535	0,182	0,909	0,178	0,133
037	0,683	0,427	0,789	0,189	0,073
043	0,660	0,646	0,850	0,391	0,338
047	0,854	0,660	1,000	0,143	0,632
051	0,667	0,889	1,000	0,222	0,000
054	0,642	0,690	0,857	0,581	0,235
055	0,588	0,727	1,000	0,235	–
064	0,628	0,685	1,000	0,189	0,000
072	0,561	0,780	0,889	0,405	0,000
075	0,598	0,672	0,882	0,362	0,145
096	0,560	0,868	0,750	0,200	0,242
099	0,716	0,484	1,000	0,386	0,231

Tabela 2: Avaliação do PALAVRAS no DIP.

6. Avaliação do sistema participante

Apenas um sistema participou no DIP, o PALAVRAS-DIP (ver Bick (2023)), sobre o qual apresentamos aqui as medidas que alcançou em outubro de 2022. Fazendo a média sobre as várias obras, em 21 obras, obteve um AI médio de 0,634, um ACI médio de 0,681, um AG médio de 0,895, e um APOES médio de 0,246. Como uma obra na CD não tinha nenhumas relações, a média de AR, feita sobre 20 obras, foi de 0,155.

Na Figura 1 mostramos as diferentes classificações para a identificação. A obra em que obteve melhores resultados na métrica AI, 0,854, foi a obra 47: *A escrava Isaura*, de Bernardo Guimarães. A obra em que funcionou pior, com 0,405, foi a obra 26: *Simá*, de Lourenço Amazonas.

A Figura 2 contém as diferentes classificações para a co-identificação (ou avaliação de correferência). A obra em que obteve melhores resultados, 0,913, foi a obra 1: *Miss Kate*, de Cosme

[15] Na primeira edição do DIP, isso apenas aconteceu para a obra 55: *O bom crioulo* de Adolfo Caminha.

[16] No âmbito do DIP, foram disponibilizadas as soluções para quatro obras: *Dom Casmurro* e *As Pupilas do Senhor Reitor*, de Júlio Dinis, na apresentação da tarefa; e *Quincas Borba* de Machado de Assis e *Dramas da Corte* de Alberto Osório de Castro, como resultado do ensaio.

[17] https://www.linguateca.pt/aval_conjunta/dip/avaliacao.html

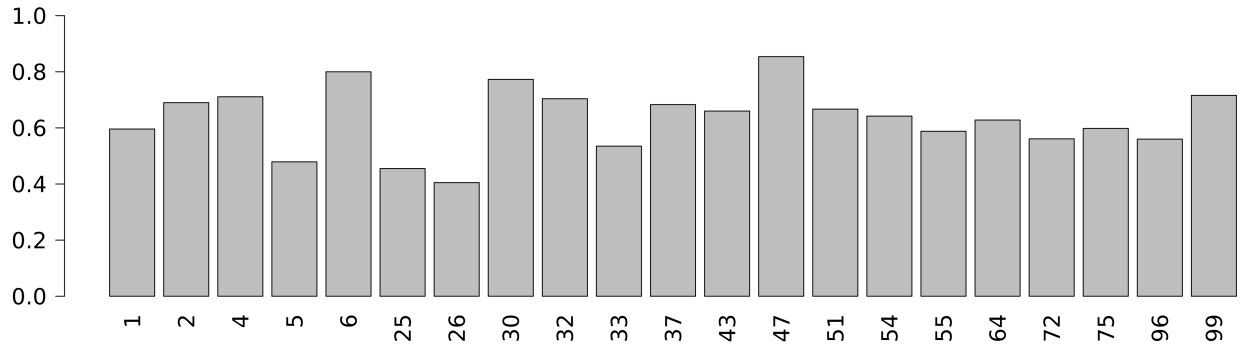

Figura 1: Resultados da avaliação da identificação

Velho, e a pior, com apenas 0,182, foi a 33: *A ermida de Castromino*, de Teixeira de Vasconcellos. Estes casos representam diferentes situações: na primeira obra, com poucas (21) personagens, a maior parte das correferências são variações dum mesmo nome, tendo ou não formas de tratamento. A segunda obra, com 33 personagens, é talvez o caso da obra com mais erros provenientes do reconhecimento ótico de carateres.

A Figura 3 contém as diferentes classificações para a avaliação do género. É claramente a tarefa em que o PALAVRAS tem melhor desempenho, e que podemos também claramente considerar a tarefa mais fácil. Mesmo assim houve algumas obras com uma pontuação fraca, como 0,647 para a obra 26: *Simá*, de Lourenço Amazonas, e 0,750 para a obra 99 *Amar, verbo intransitivo*, de Mário de Andrade.

A Figura 4 contém as diferentes classificações para a avaliação da profissão, ocupação e estatuto social.

Para a avaliação de APOES, a melhor pontuação foi de 0,581 na obra 54: *O Barão de Lavos*, de Abel Botelho. A pior, de zero, refere-se à obra 1, *Miss Kate*, de Cosme Velho. Esta obra ilustra uma das fragilidades, ou problemas, da forma de fazer a avaliação da profissão, ocupação e estatuto social no DIP: o facto de não haver uma normalização, ou ontologia das várias formas de definir esses termos. Por exemplo, a personagem Tiburtino Mendes está descrita na coleção dourada como "sextanista da faculdade de Medicina," e na resposta do sistema como "estudante." A resposta está evidentemente certa, e é até talvez mais útil de um ponto de vista de leitura distante, em que quereríamos juntar todos os estudantes e não distinguir o ano ou a faculdade em que estudam.

Mas para conseguir que a avaliação automática considerasse a resposta do sistema como certa, teríamos de fazer um trabalho considerável de padronização e definição do que se consideraria correto ao nível do campo semântico "descrição profissional."

Finalmente, a Figura 5 contém as diferentes classificações para a avaliação da extração de relações.

Esta foi a área em que o PALAVRAS teve pior desempenho, e que foi aliás mencionada pelo seu autor como ainda não estando suficientemente desenvolvida na altura da participação no DIP. Não admira, portanto, que não tenha conseguido identificar relações em oito obras. O melhor resultado foi de 0,632 para a obra 47: *A escrava Isaura*, de Bernardo Guimarães.

Se quisermos apreciar o desempenho global do PALAVRAS em relação a cada obra, e fazendo portanto a média aritmética das cinco (ou quatro) medidas, vemos na Figura 6 que o desempenho varia entre um máximo de 0,6578 e um mínimo de 0,3874 para os já mencionados *A escrava Isaura* e *A ermida de Castromino*.

Colocámos, na figura, numa cor diferente as obras escritas por autoras. Na figura 7 parece que as obras escritas por autoras tiveram pior classificação global, mas a diferença não é estatisticamente significativa.

6.1. Nova tentativa do PALAVRAS

Para ver se estas medidas podem contribuir para uma comparação racional de diferentes sistemas, pedimos novos resultados ao PALAVRAS em abril de 2023, sabendo que o autor tinha melhorado o sistema desde o DIP, que relembramos, ocorreu de 15 a 17 de setembro de 2022. Os novos resultados encontram-se na Tabela 3.

O valor médio de AI, ACI, AG e APOES, calculado sobre 21 obras, foi de 0,633, 0,702, 0,934 e 0,262. Quanto à avaliação de relações, a média sobre as 20 obras é de 0.195.

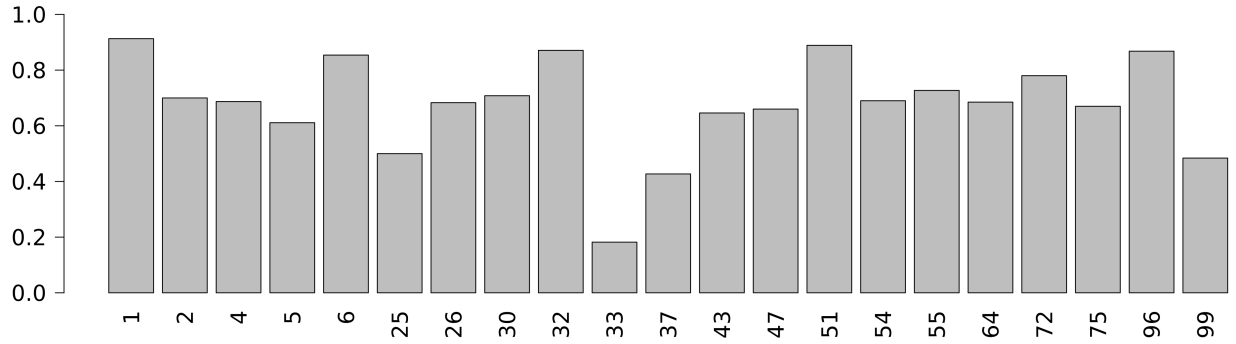

Figura 2: Resultados da avaliação da correferência, ou co-identificação

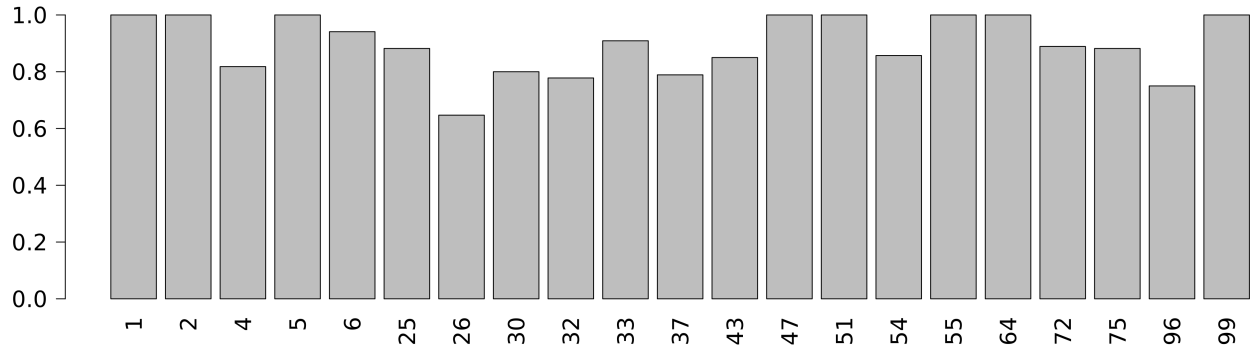

Figura 3: Resultados da avaliação do género

obr	AI	ACI	AG	APO	AR
001	0,485	**0,942**	0,857	0,000	0,000
002	0,647	**0,824**	1,000	0,091	0,000
004	**0,777**	0,585	**0,911**	0,303	0,093
005	**0,514**	0,514	1,000	0,080	**0,400**
006	0,724	0,659	0,852	0,053	**0,308**
025	0,446	0,485	**0,913**	**0,182**	0,000
026	0,404	**0,706**	**0,789**	**0,379**	0,286
030	0,765	**0,825**	**0,939**	**0,211**	**0,429**
032	**0,765**	0,730	**0,833**	**0,205**	0,000
033	**0,567**	**0,591**	**1,000**	0,114	**0,174**
037	0,626	**0,489**	**0,909**	**0,246**	**0,087**
043	**0,686**	**0,712**	0,822	**0,400**	0,314
047	0,796	**0,976**	1,000	**0,196**	0,381
051	**0,833**	0,889	1,000	**0,750**	0,000
054	0,614	**0,820**	**0,938**	0,439	0,000
055	**0,632**	0,462	1,000	**0,364**	–
064	0,567	**0,716**	1,000	0,167	0,000
072	**0,650**	0,762	**0,920**	0,333	**0,222**
075	0,446	0,543	**0,933**	0,324	0,114
096	**0,687**	**0,939**	**1,000**	**0,421**	**0,846**
099	0,656	**0,577**	1,000	0,243	**0,242**

Tabela 3: Avaliação do PALAVRAS em maio de 2023: a negrito estão os casos em que melhorou

De notar que, tal como na avaliação conjunta propriamente dita, tivemos de fazer alguns ajustes à CD para não penalizar novas menções que não tinham sido encontradas pelos revisores humanos.

Embora não vejamos grande diferença no desempenho do sistema, convém também sublinhar que o PALAVRAS não seguiu completamente as diretivas do DIP, em duas questões que podem ter consequências negativas nas medidas de avaliação: normalizou os nomes das profissões, ocupações e estatutos sociais – por exemplo, *creada* passou para *criada*; e também os (poucos) casos em que um nome de relação familiar era usado como forma de tratamento (ou seja, *viuva Maria* para *viúva Maria*, ou *mãi Joana* para *mãe Joana*).

Poderíamos — ou talvez deveríamos mesmo, para um medição mais justa do desempenho do sistema — fazer uma avaliação mais relaxada, para realmente identificar, e premiar, os casos que apenas são devidos a falta de normalização na CD.

7. Avaliação crítica da metodologia

Há várias razões para sermos críticos em relação à metodologia de avaliação usada no primeiro DIP, embora tenha sido por nós proposta com a me-

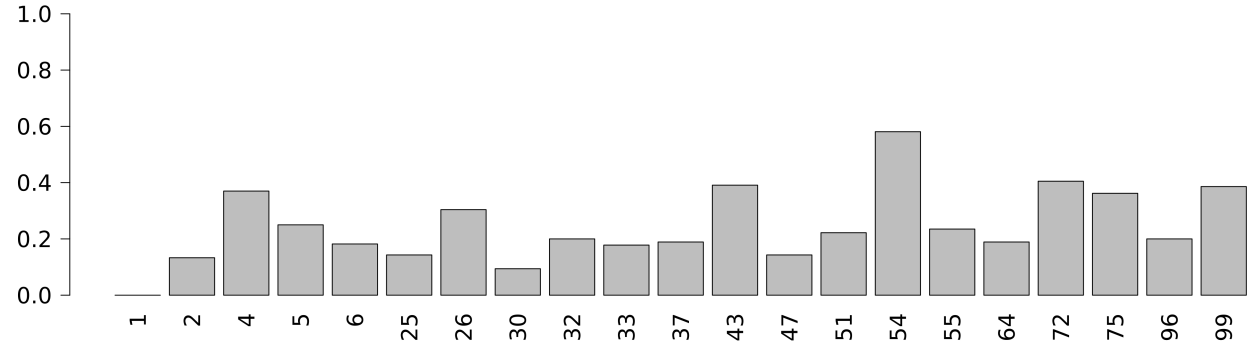

Figura 4: Resultados da avaliação da profissão, ocupação e estatuto social

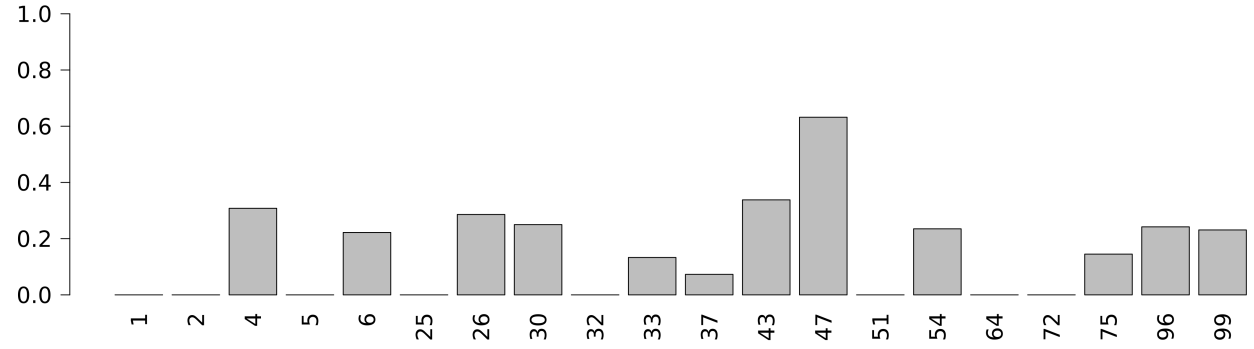

Figura 5: Resultados da avaliação da extração de relações familiares

lhor das intenções. Em primeiro lugar, o termos dividido a avaliação em cinco avaliações distintas impede uma visão global da dificuldade da tarefa, e da relação entre as diferentes subtarefas. Ou seja, é plausível que o facto de que uma personagem tenha muitos nomes torne a deteção do seu género, ou da sua profissão, mais difícil. Ou que uma personagem muito frequente tenha mais formas de ser mencionada do que outra que só aparece duas vezes. Contudo, todas são tratadas da mesma maneira.

Queremos com isto dizer que as medidas de avaliação abstraem da dificuldade intrínseca de uma dada obra, e de uma dada personagem de uma dada obra.

Idealmente deveríamos pesar a avaliação de um sistema pela dificuldade da obra e da personagem: uma obra com cinquenta personagens cada uma delas com mais do que uma forma de serem mencionadas é certamente mais complicada do que outra apenas com três personagens só com uma forma de serem identificadas. Contudo, ambas as obras contam igualmente.

A mesma coisa se passa para personagens com nomes genuinamente diferentes, e que mudam de profissão ao longo de uma obra. O acerto de um sistema sobre elas conta o mesmo que os casos de uma personagem que só aparece uma vez, e que portanto só tem uma forma de ser mencionada.

Deveríamos ser capazes de medir a dificuldade de identificação das personagens de uma obra, com base na coleção dourada, e depois avaliar um sistema entrando em conta com isso. Contar pouco os casos fáceis, e dar mérito aos casos difíceis.

Em segundo lugar, o uso de medidas separadas para as subtarefas pode ser enganador, visto que elas não medem, por exemplo, a capacidade global de identificar profissões de um dado sistema. E isto porque apenas nos atemos às profissões das personagens bem identificadas.

Considerando uma situação em que há dez personagens com 10 profissões, e que o sistema apenas identifica duas dessas personagens, e as profissões estão certas, terá uma medida 1 na APOES, mas só encontrou duas das 10 profissões que um leitor estaria interessado em encontrar.

Parece-nos que o problema da avaliação do DIP como a concebemos é que não toma (excepto na medida do género), as personagens certas como ponto de partida.

O ideal, parece-nos agora, seria ter uma medida por personagem. Quantas suas formas foram identificadas e juntas? Quantas suas profissões e relações com outras personagens foram identificadas? Quantas personagens faltam? Quantas são espúrias?

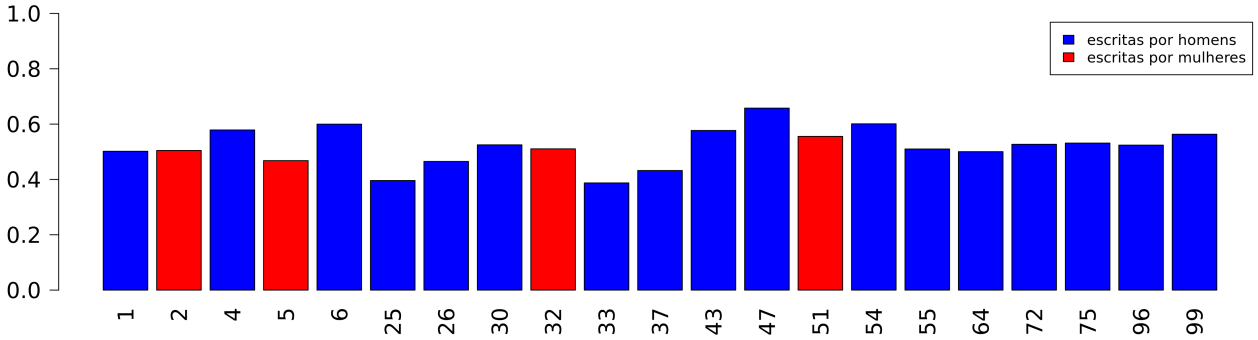

Figura 6: Resultados da avaliação total

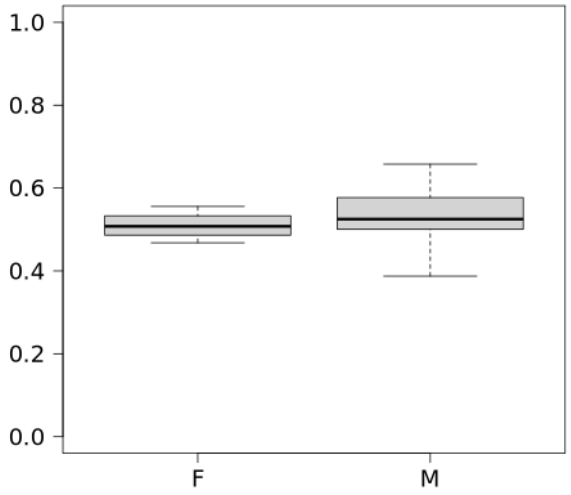

Figura 7: Avaliação total pelo género do autor

Em vez de concentrar a avaliação em pares de nomes correferentes, como fazemos em ACI, ou a pares profissão nome, em APOES.

No entanto, definir uma métrica que consiga pontuar todos os casos possíveis de uma forma satisfatória não é fácil, e terá de ficar para trabalho futuro.

8. Conclusões

A tarefa do DIP é um pouco diferente das tarefas tratadas na literatura que revisamos, porque junta várias tarefas (como identificação e correferência), não se aplica aos próprios textos (como é o caso da marcação de protagonistas) mas apenas extrai essa informação da obra completa. Embora também compare/avalie um tipo de redes de personagens, são diferentes das usadas nos trabalhos acima descritos, porque não dependem da interação ou da co-ocorrência das personagens no texto, são apenas representações das suas relações familiares.

Seja como for, existem muitos pontos de contacto com os trabalhos mencionados anteriormente, e todos focam o estudo de personagens em textos literários.

Neste artigo, apresentámos detalhadamente as métricas de avaliação usadas no primeiro DIP, os resultados com elas obtidos pelo único sistema participante, o PALAVRAS-DIP, no DIP e seis meses mais tarde, e fizemos várias críticas no sentido de esclarecer os potenciais problemas, e futuramente desenvolver outras medidas.

Agradecimentos

Agradecemos vivamente à Cristina Mota a bateria de testes que desenvolveu para testar os programas de avaliação, e aos outros organizadores, participantes e observadores do DIP, pelo retorno dado em variadas ocasiões.

Agradecemos também a Rebeca Schumacher e Cristina Mota os seus comentários que nos permitiram melhorar o texto.

Agradecemos à FCCN – Fundação para a Computação Científica Nacional (Portugal), o alojamento da Linguateca nos seus servidores, e ao UNINETT Sigma2 - the National Infrastructure for High Performance Computing and Data Storage in Norway pelos recursos computacionais.

Referências

Agarwal, Apoorv, Anup Kotalwar & Owen Rambow. 2013. Automatic extraction of social networks from literary text: A case study on Alice in wonderland. Em *6th International Joint Conference on Natural Language Processing*, 1202–1208.

Bagga, Amit & Breck Baldwin. 1998. Algorithms for scoring coreference chains. Em *1st International Conference on Language Resources and Evaluation (LREC)*, vol. 1, 563–566.

Bick, Eckhard. 2023. Extraction of literary character information in Portuguese. *Linguamática* 15(1). 31–40. 10.21814/lm.15.1.397.

Bontcheva, Kalina, Valentin Tablan, Diana Maynard & Hamish Cunningham. 2004. Evolving GATE to meet new challenges in language engineering. *Natural Language Engineering* 10(3–4). 349–373. 10.1017/S1351324904003468.

Chaturvedi, Snigdha, Mohit Iyyer & Hal Daume III. 2017. Unsupervised learning of evolving relationships between literary characters. Em *AAAI Conference on Artificial Intelligence*, vol. 31 1, 3159–3165. 10.1609/aaai.v31i1.10982.

Chu, Cuong Xuan, Simon Razniewski & Gerhard Weikum. 2021. KnowFi: Knowledge extraction from long fictional texts. Em *3rd Conference on Automated Knowledge Base Construction*, on line. 10.24432/C51S38.

Dekker, Niels, Tobias Kuhn & Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Computer Science* 5. e189.

de Does, Jesse, Katrien Depuydt, Karina Van Dalen-Oskam, Maarten Marx et al. 2017. Namescape: Named entity recognition from a literary perspective. Em *CLARIN in the Low Countries*, chap. 30, 361–370. Ubiquity Press London.

Elson, David & Kathleen McKeown. 2010. Automatic attribution of quoted speech in literary narrative. Em *AAAI Conference on Artificial Intelligence*, vol. 24 1, 1013–1019. 10.1609/aaai.v24i1.7720.

Elson, David K, Nicholas J. Dames & Kathleen McKeown. 2010. Extracting social networks from literary fiction. Em *48th Annual Meeting of the Association for Computational Linguistics*, 138–147.

Finkel, Jenny Rose, Trond Grenager & Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. Em *43rd Annual Meeting of the Association for Computational Linguistics*, 363–370. 10.3115/1219840.1219885.

Groza, Adrian & Lidia Corde. 2015. Information retrieval in falktales using natural language processing. Em *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 59–66. 10.1109/ICCP.2015.7312606.

Horridge, Matthew & Sean Bechhofer. 2011. The OWL API: A Java API for OWL ontologies. *Semantic Web* 2(1). 11–21.

Jayakumar, Archana, Vedica Rao, AS Rohit Kumar, Prithwijit Banerjee & Roopa Ravish. 2022. Analyzing the development of complex social systems of characters in a work of literary fiction. Em *3rd International Conference for Emerging Technology (INCET)*, 1–7. 10.1109/INCET54531.2022.9824015.

Labatut, Vincent & Xavier Bost. 2019. Extraction and analysis of fictional character networks: A survey. *ACM Computing Surveys* 52(5). 1–40. 10.1145/3344548.

Łajewska, Weronika & Anna Wróblewska. 2021. Protagonists' tagger in literary domain–new datasets and a method for person entity linkage. *arXiv preprint arXiv:2110.01349* .

Lee, Heeyoung, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu & Dan Jurafsky. 2011. Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. Em *15th Conference on Computational Natural Language Learning: Shared task*, 28–34.

Lee, John & Chak Yan Yeung. 2012. Extracting networks of people and places from literary texts. Em *26th Pacific Asia Conference on Language, Information, and Computation*, 209–218.

Luo, Xiaoqiang. 2005. On coreference resolution performance metrics. Em *Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 25–32.

Mota, Cristina & Diana Santos. 2023. Pais, filhos, e outras relações familiares no DIP. *Linguamática* 15(1). 41–53. 10.21814/lm.15.1.402.

Pradhan, Sameer, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng & Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. Em *52nd Annual Meeting of the Association for Computational Linguistics*, 30–35. 10.3115/v1/P14-2006.

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

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2022a. Introduction to DIP: goal, setup, resources and results. Apresentação. https://www.linguateca.pt/aval_conjunta/dip/apr_encontro/DIPpresentation.pdf.

Santos, Diana, Cristina Mota, Emanoel Pires, Marcia Caetano Langfeldt, Rebeca Schumacher Fuão & Roberto Willrich. 2023. DIP: Desafio de identificação de personagens: objectivo, organização, recursos e resultados. *Linguamática* 15(1). 3–30. 10.21814/lm.15.1.399.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher & Paulo Silva Pereira. 2022b. Identifying literary characters in Portuguese: Challenges of an international shared task. Em *Computational processing of the Portuguese language, (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39.

Srinivasan, Vardhini & Aurelia Power. 2022. Character extraction and character type identification from summarised story plots. *Journal of Computer-Assisted Linguistic Research* 6. 19–41. 10.4995/jclr.2022.17835.

Vala, 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 (EMNLP)*, 769–774. 10.18653/v1/D15-1088.

Valls-Vargas, Josep, Santiago Ontañón & Jichen Zhu. 2014. Toward automatic character identification in unannotated narrative text. Em *7th Intelligent Narrative Technologies Workshop*, 38–44.

Vilain, Marc, John D Burger, John Aberdeen, Dennis Connolly & Lynette Hirschman. 1995. A model-theoretic coreference scoring scheme. Em *6th Message Understanding Conference (MUC)*, 45–52.

Artigos de Investigação

Proposta recebida em Outubro 2022 e aceite para publicação em Janeiro 2023.

Extracção de Relações de Apoio e Oposição em Títulos de Notícias de Política em Português

Extraction of Support and Opposition Relationships in Portuguese Political News Headlines

David S. Batista

Resumo

Títulos de notícias de política relatam com frequência relações de apoio ou oposição entre personalidades, por exemplo: *"Marques Mendes critica estratégia de Rui Rio"* ou *"Costa reafirma confiança em Centeno."* Neste trabalho analisámos milhares de títulos arquivados, identificando os que expressam relações de apoio ou oposição e associando as personalidades políticas com o seu identificador na Wikidata, resultando assim num grafo semântico. O grafo permite responder a interrogações envolvendo personalidades políticas e partidos. Descrevemos o processo de geração do grafo e tornamo-lo disponível, assim como uma colecção de dados anotada manualmente, que permitiu treinar classificadores de aprendizagem supervisionada para identificar as relações expressas nos títulos e ligar as personalidades com a Wikidata.

Palavras chave

extracção de relações semânticas, dados anotados, web semântica, ciência política

Abstract

Political news headlines often report supportive or opposing relationships between personalities, for example: *"Marques Mendes criticizes Rui Rio's strategy"* or *"Costa reaffirms confidence in Centeno."* In this work we analyzed thousands of archived titles, identifying those that express supportive or opposing relationships, and associating the political personalities with their identifier on Wikidata, thus resulting in a semantic graph. The graph allows answering questions involving political personalities and parties. We describe the graph generation process and make it available together with a labelled dataset, which allowed supervised learning classifiers to be trained to identify the relationships expressed in the titles and link the personalities with Wikidata.

Keywords

semantic relationship extraction, annotated dataset, semantic web, political science

1. Introdução

Os títulos de notícias relacionados com política ou políticos relatam com frequência interacções envolvendo duas ou mais personalidades políticas. Muitas dessas interacções correspondem a relações de apoio ou oposição de uma personalidade para uma outra personalidade, por exemplo:

- *"Marques Mendes critica estratégia de Rui Rio"*
- *"Catarina Martins pede a demissão do governador Carlos Costa"*
- *"Sócrates foi às bases apelar ao voto em Soares"*

A análise de um grande número deste tipo de relações ao longo do tempo permite vários estudos, por exemplo: encontrar quais as grandes comunidades de apoio ou oposição em função dos governos no poder, ou encontrar as grandes alianças e oposições e as suas dinâmicas. Pode-se também explorar individualmente uma personalidade ao longo do tempo, por exemplo, comparando as relações de apoio ou oposição antes de tomar posse em determinado cargo público com as relações depois de ter assumido o cargo, ou ver que relações de apoio subitamente emergiram. Uma base de dados reunindo notícias expressando relações de apoio ou oposição entre personalidades políticas pode ser usada para rapidamente reunir uma colecção de notícias contendo ou envolvendo personalidades e partidos políticos específicos, por exemplo, para auxiliar numa tarefa de jornalismo de investigação.

Tendo um método automático para extrair relações e podendo aplicá-lo a uma colecção de dados abrangendo longos períodos de tempo permitiria concretizar os exemplos descritos anteriormente.

Neste trabalho apresentamos um método para extrair relações de apoio ou oposição entre personalidades políticas e descrevemos os resultados da aplicação do mesmo a uma colecção

DOI: 10.21814/lm.15.1.386

de notícias abrangendo um período de cerca de 25 anos. Durante o processo de extracção das relações ligámos as personalidades políticas envolvidas com o seu identificador na Wikidata (Malyshev et al., 2018) enriquecendo assim a relação com informação associada à personalidade (e.g.: afiliação política, cargos públicos exercidos, legislaturas, relações familiares, etc.).

Todas as relações extraídas são representadas sob a forma de triplos semânticos seguindo a norma *Resource Description Framework* (RDF) (Schreiber & Raimond, 2014). As personalidades políticas envolvidas, representadas pelo seu identificador na Wikidata, são ligadas através de uma relação de oposição ou apoio representada pela notícia que dá suporte à relação. Esta estrutura dá assim origem a um gráfico semântico, sendo então possível formular interrogações SPARQL (Prud'hommeaux et al., 2013) envolvendo a informação da Wikidata associada a cada personalidade e as relações extraídas dos títulos de notícias, por exemplo:

- Listar todas as notícias onde a personalidade X se opõe à personalidade Y
- Listar os membros de um determinado partido que apoiaram alguma personalidade específica
- Listar os membros de um determinado partido apoiados/opostos por membros de um outro partido
- Listar personalidades que estão ligadas através de uma relação familiar e de uma relação de oposição/apoio
- Listar personalidades que fazem parte do mesmo Governo e que estão envolvidas numa relação de oposição/apoio

As principais contribuições deste trabalho são:

- um grafo semântico ligando personalidades políticas representadas na Wikidata através de uma relação de oposição ou apoio suportada por uma notícia
- um conjunto de dados anotados utilizado para treinar os classificadores de extracção de relações sentimento direccionado de títulos de notícias, e também para ligar as personalidades mencionadas à Wikidata
- um interface web que permite explorar o gráfico semântico

Este artigo está organizado da seguinte forma: na Secção 2 referimos trabalho relacionado, na Secção 3 descrevemos a base de conhecimento usada no suporte de ligação das personalidades políticas à Wikidata. A Secção 4 refere e descreve as fontes de notícias utilizadas. Na Secção 5 detalhamos o conjunto de dados anotados e na Secção 6 os classificadores de aprendizagem supervisionada desenvolvidos. Na Secção 7 descrevemos o processo de extracção de triplos RDF e a construção do gráfico semântico. Finalmente, na Secção 8 reunimos as conclusões deste trabalho e apresentamos algumas ideias para trabalho futuro.

2. Trabalho Relacionado

A análise de sentimento, no contexto de Processamento de Linguagem Natural, tem sido maioritariamente alvo de estudo em conteúdo gerado em redes sociais (Zimbra et al., 2018) ou na avaliação de produtos ou serviços (Pontiki et al., 2016). Nestes domínios o autor do texto e o alvo da opinião são explícitos. No contexto de análise de notícias de política, onde existe com frequência um sentimento expresso entre actores políticos sob a forma de relações de apoio ou oposição (Balahur et al., 2009, 2010), as abordagens de análise de sentimento a produtos ou serviços não se aplicam, dado que a direcção da relação de sentimento tem que ser considerada.

Nesta secção descrevemos recursos semelhantes aos que produzimos neste trabalho, que tornamos públicos, e abordagens para a tarefa de extracção de sentimento direccionado em texto de notícias de política.

2.1. Recursos e dados anotados

Sarmento et al. (2009) propõem um método para a criação automática de um corpus para detecção de um sentimento positivo ou negativo para com uma personalidade política, e aplicam o método a comentários a notícias de jornais *on-line*. Neste recurso a origem do sentimento, pressupõem-se, é do comentador.

Moreira et al. (2013) disponibilizam uma ontologia descrevendo actores políticos, os seus cargos e partidos políticos afiliados, usando fontes de informação oficiais e informação recolhida da *web* para adicionar nomes alternativos às personalidades presentes na ontologia.

de Arruda et al. (2015) criaram um corpus de notícias políticas em português do Brasil, anotando cada parágrafo com o sentimento segundo duas dimensões: o actor político referido pelo parágrafo, e o sentimento dessa referência: positivo, negativo ou neutro. Neste recurso fica em aberto qual é a origem do sentimento. (Baraniak & Sydow, 2021) disponibilizam corpora se-

melhante, anotando o sentimento para com uma personalidade política em textos de jornais *online*, para o Inglês e Polaco.

2.2. Extracção de sentimento direccionado em texto de notícias

Vários autores exploraram métodos para extrair sentimento envolvendo actores políticos. De notar que muitos dos trabalhos transformam a tarefa de detecção de sentimento numa tarefa de detecção de uma relação entre entidades mencionadas (Bassignana & Plank, 2022).

Alguns exploram estas relações num contexto de política internacional, i.e.: os actores são nações referidas em texto de notícias de política, sendo que algumas dessas relações implicitamente têm um sentimento positivo ou negativo. O'Connor et al. (2013) propõem um modelo não supervisionado baseado em *topic models* e padrões linguísticos para identificar relações, de forma aberta, descrevendo conflitos entre nações referenciadas em artigos de notícias em Inglês. Han et al. (2019) propõem também um modelo não supervisionado para gerar descritores de relações para pares de nações mencionadas em notícias em Inglês. O modelo proposto estende o trabalho de Iyyer et al. (2016) integrando informação linguística (i.e.: predicados verbais e substantivos comuns e próprios) por forma a identificar o contexto das relações.

Liang et al. (2019) define a tarefa de extracção de relações de culpabilidade para textos em Inglês: dado um artigo d e um conjunto de entidades E, presentes no artigo, detectar se existe uma relação de culpabilidade (s, t), onde $s, t \in E$, quando s culpa t com base no artigo d, sendo que há $|E| \cdot (|E| - 1)$, possíveis relações de culpabilidade. Para detectar estas relações os autores propõem três modelos. O modelo *Entity Prior* extrai informação sobre entidades, tentando capturar um *prior* sobre quem é susceptível de culpar quem sem informação adicional. O modelo *Context* faz uso da informação de contexto da frase onde duas entidades ocorrem para determinar a presença de uma relação de culpabilidade. O modelo *Combined* combina a informação dois modelos anteriores num único modelo. Os autores aplicam esta abordagem num corpus com 998 artigos de notícias e com cerca de 3 entidades por artigo reportando uma macro-média F_1 de 0,70 com o modelo *Combined*.

Park et al. (2021) propõe uma estrutura de relações para detectar o sentimento e a direcção: dada uma frase s referindo duas entidades p e q, detectar qual a relação de sentimento entre p e q de entre cinco possíveis: neutra, p tem uma opinião positiva ou negativa em relação a q, ou q tem uma opinião positiva ou negativa em relação a p. No seu trabalho os autores usam múltiplos modelos transformando a tarefa de extracção de sentimentos em sub-tarefas que respondem a perguntas de sim/não para cada um dos 5 sentimentos possíveis, combinando depois os vários resultados num resultado final. Esta abordagem é aplicada para Inglês num corpus criado pelos autores contendo frases de artigos de notícias contento pelo menos duas entidades. Os pares de entidades estão anotadas com um dos 5 sentimentos possíveis. Os autores reportam uma macro-média de F_1 de 0,68.

3. Base de Conhecimento

Dado que as personalidades envolvidas nas relações a extrair são personalidades políticas relevantes, começámos por construir uma base de conhecimento a partir da Wikidata (Malyshev et al., 2018). Fazendo interrogações SPARQL ao *endpoint* público[1] recolhemos o identificador de todas as:

- pessoas que são ou foram afiliadas a algum partido político português
- pessoas portuguesas nascidas depois de 1935 cuja profissão seja: *juiz, economista, advogado, funcionário público, político, empresário ou banqueiro*
- pessoas que têm ou tiveram pelo menos um cargo de uma lista de cargos públicos portugueses previamente seleccionados (e.g.: *ministro, líder do partido, embaixador*, etc.)

Para além dos resultados destas interrogações, seleccionámos manualmente alguns identificadores de personalidades não abrangidas pelas interrogações SPARQL definidas acima, muitas delas de um contexto político internacional, mas que interagem com personalidades portuguesas. Acrescentámos também todos os identificadores de partidos políticos a que as personalidades recolhidas estão afiliadas. Este processo resultou num total de 1 757 personalidades e de 37 partidos políticos. De notar que alguns dos partidos incluídos são partidos já extintos e/ou de um contexto internacional. Para cada um dos identificadores das personalidades e partidos descarregámos a página correspondente na Wikidata utilizando um outro *endpoint*[2] público.

[1] https://query.wikidata.org

[2] https://www.wikidata.org/wiki/Special:EntityData?

Título	**Relação**
Sá Fernandes acusa **António Costa** de defender interesses corporativos	Ent_1-opõe-se-Ent_2
Joana Mortágua: declarações de **Cavaco** são "uma série de disparates"	Ent_1-opõe-se-Ent_2
Passos Coelho é acusado de imaturidade política por **Santos Silva**	Ent_2-opõe-se-Ent_1
Durão Barroso defende **Paulo Portas** como "excelente ministro"	Ent_1-apoia-Ent_2
Armando Vara escolhido por **Guterres** para coordenar autárquicas	Ent_2-apoia-Ent_1
Manuel Alegre recebe apoio de **Jorge Sampaio**	Ent_2-apoia-Ent_1
Rui Tavares e **Ana Drago** eleitos nas primárias do LIVRE	outra
Teresa Zambujo reconhece vitória de **Isaltino Morais**	outra
CDS acusa **Marcelo Rebelo de Sousa** de pôr em causa relação com **Cavaco**	outra

Tabela 1: Exemplos de títulos e das relações manualmente anotadas correspondentes.

Para cada personalidade política seleccionámos: o seu identificador na Wikidata, o seu nome mais comum e os nomes alternativos, i.e.: combinações de nomes próprios e apelidos. Com base nestes três campos, criámos um índice no ElasticSearch (Gormley & Tong, 2015) usando a sua configuração de omissão, não fazendo uso de qualquer funcionalidade extra tais como analisadores de n-gramas.

4. Fontes de dados

A principal fonte de notícias foi o arquivo da *web* portuguesa (Gomes et al., 2013). Usando a API pública de pesquisa recolhemos páginas arquivadas restringindo os resultados a ocorrências de nomes reunidos na Secção 3 e a 45 domínios `.pt` associados a diversas fontes de informação: jornais *on-line*, *websites* de estações de televisão e rádio, e portais agregadores de conteúdos.

Uma segunda fonte de notícias foi a colecção CHAVE[3](Santos & Rocha, 2004, 2001), contendo os artigos do jornal PÚBLICO publicados entre 1994 e 1995. Finalmente, foram também adicionados alguns artigos não arquivados pelo arquivo.pt, retirados directamente das secções *Mundo*, *Política* e *Sociedade* do site publico.pt.

Deste processo resultou uma colecção de cerca de 13,7 milhões de títulos de artigos publicados entre 1994 e 2022. De seguida foi aplicado um pré-processamento de modo a remover notícias com: títulos duplicados, títulos com menos de 4 palavras, e títulos ou URLs contendo palavras que fazem partem de uma lista pré-definida (e.g.: *desporto*, *celebridades*, *artes*, *cinema*, etc.) que sugerem um outro contexto que não política. Este pré-processamento resultou em 1,3 milhões de títulos distintos, cerca de 10% dos dados inicialmente recolhidos.

[3] https://www.linguateca.pt/CHAVE/

5. Colecção de Relações Anotadas

De forma a poder treinar classificadores de aprendizagem supervisionada para identificar as relações presentes nos títulos das notícias, e fazer a ligação das personalidades com a Wikidata, anotámos manualmente títulos com: as menções a personalidades, os identificadores na Wikidata e a relação entre as personalidades mencionadas.

Começámos por pré-processar todos os títulos recolhidos usando o pacote de software spaCy 3.0 (Honnibal et al., 2020), com o modelo `pt_core_news_lg-3.0.0` fizemos o reconhecimento de entidades mencionadas do tipo `PESSOA`. Para cada entidade reconhecida tentamos encontrar o seu identificador correspondente na Wikidata fazendo uma interrogação ao índice descrito na Secção 3 e assumindo que na lista de resultados o primeiro é o identificador correcto associado à entidade. Seleccionámos depois os títulos para anotação, incluindo apenas títulos referindo pelos menos duas personalidades.

No processo de anotação todos os títulos foram carregados para ferramenta de anotação Argilla,[4] e usando o interface gráfico fomos seleccionando títulos para anotar.

Para cada título corrigimos sempre que necessário as entidades reconhecidas e os seus identificadores na Wikidata, caso existam. Anotámos a relação existente: **oposição** ou **apoio**, e a direcção da mesma. Quando nenhuma das duas se verifica a relação é anotada como **outra**. A Tabela 1 demonstra alguns exemplos das relações anotadas. O processo de anotação foi feito por um anotador. Nas situações mais ambíguas, por exemplo, onde é necessária a informação no texto da notícia para decidir, as relações foram anotados como **outra**.

Este processo resultou num conjunto de dados contendo 3.324 títulos anotados. Para cada título anotámos apenas duas personalidades e a

[4] https://github.com/argilla-io/argilla

relação entre elas, mesmo que os títulos contenham referências a mais do que duas personalidades. A Tabela 2 caracteriza os dados em termos de número de relações e direcção. A maioria dos títulos contém uma relação de **oposição** ou **outra**, e a grande maioria das relações têm uma direcção da primeira para a segunda entidade, $\text{Ent}_1 \rightarrow \text{Ent}_2$.

Relação	**$\text{Ent}_1 \rightarrow \text{Ent}_2$**	**$\text{Ent}_1 \leftarrow \text{Ent}_2$**	**Total**
opõe-se	1 155	102	1 257
apoia	717	44	761
outra	-	-	1 306
Total	1 872	146	3 324

Tabela 2: Relações por classe e direcção.

O rácio de relações de oposição para com relações de apoio é de 1,6, este valor é semelhante com os dados para Inglês disponibilizados por Park et al. (2021), onde esta mesma relação entre as duas classes é de 1,8. Em termos de representatividade das classes, agregadas por sentimento, os dois conjuntos de dados também são semelhantes, sendo **outra** a classe mais presente, seguindo-se **oposição** e por último **apoio**.

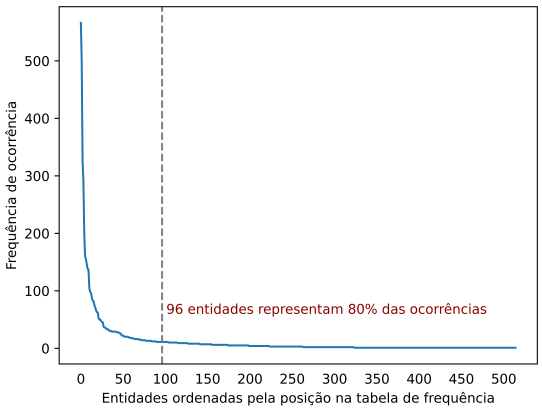

Figura 1: Distribuição de frequências de ocorrências das personalidades nos títulos anotados.

Das 6.648 menções a nomes de personalidades políticas anotadas, 515 são distintas e têm um identificador na Wikidata. Um total de 129 entidades distintas, identificadas por agregação da *string* que as refere no título, não estão associadas a um identificador por não estarem presentes na Wikidata.

Analisando a frequência de ocorrência de cada entidade observa-se que há pequeno número de entidades responsáveis por uma grande parte de todas as ocorrências de entidades nos dados anotados. Como mostra a Figura 1 existe um número pequeno de entidades frequentes, e uma longa lista de entidades pouco frequentes, em concreto, 96 personalidades distintas, i,.e.: 19% das personalidades, são responsáveis por 80% das menções a personalidades nos dados.

Em termos de número de palavras contidas nos títulos, excluindo palavras que fazem parte das entidades, há uma mediana de 8 palavras com um máximo de 22 e um mínimo de 1. Este conjunto de dados anotados encontram-se online[5] sob o formato JSON como ilustrado na Figura 2.

```
{"title": "Ana Gomes defende Durão Barroso",
 "label": "ent1_supports_ent2",
 "date": "2002-05-11 08:26:00",
 "url": "http://www.publico.pt/141932",
 "ent1": "Ana Gomes",
 "ent2": "Durão Barroso",
 "ent1_id": "Q2844986'",
 "ent2_id": "Q15849"}
```

Figura 2: Exemplo de um título anotado sob o formato JSON.

6. Processo de Extracção de Relações

O processo de extração de triplos RDF a partir dos títulos das notícias envolve 4 sub-processos:

- reconhecimento de entidades-mencionadas do tipo `PESSOA`
- ligação das entidades com um identificador na Wikidata
- classificação do tipo de relação
- classificação da direcção da relação

6.1. Reconhecimento de Entidades

O reconhecimento de entidades mencionadas é baseado num método híbrido, combinando regras com um modelo supervisionado.

Usando a componente `EntityRuler`[6] do spaCy 3.0, definimos uma série de regras combinando padrões baseados nos nomes de todas as personalidades da base de conhecimento descrita na Secção 3. Para detectar as entidades do tipo `PESSOA` este classificador aplica primeiro as regras e de seguida o modelo supervisionado para Português do spaCy. Em situações de desacordo entre as duas abordagens as entidades marcadas com regras têm prioridade.

[5] https://github.com/politiquices/data-releases
[6] https://spacy.io/usage/rule-based-matching

A Tabela 3 mostra a performance para as três abordagens sob o conjunto de dados anotado.

Abordagem	**P**	**A**	**F_1**
Regras	0,99	0,42	0,59
Modelo	0,97	0,91	0,94
Regras+Modelo	0,97	0,92	0,94

Tabela 3: (P)recisão, A(brangência) e F_1 para a componente de REM combinando regras e o modelos supervisionado.

6.2. Ligação com a Wikidata

O algoritmo para associar personalidades com identificadores na Wikidata tem duas fases. Numa primeira fase o algoritmo tenta apenas usar o título da notícia, se este processo falhar, tenta então usar possíveis referências às personalidades no texto da notícia.

O algoritmo começa por fazer uma interrogação à base de conhecimento (BC), usando a referência à personalidade presente no título, gerando assim uma lista de candidatos para uma determinada personalidade. Se a lista contém apenas um candidato e a similaridade (Jaro, 1989) para com a personalidade referida no título é de pelo menos 0,8 esse candidato é seleccionado. Se houver mais do que um candidato, o algoritmo filtra apenas aqueles com uma similaridade de 1,0 e se houver apenas um esse é o candidato seleccionado. Em qualquer outro caso nenhum candidato é retornado.

O Algoritmo 1 descreve o procedimento que usa apenas o título da notícia.

```
def title_only(ent, candidates):
    if len(candidates) == 1:
      if jaro(ent,candidates[0]) >= 0.8:
        return candidates[0]
    else:
      filtered = exact(ent, candidates):
        if len(filtered) == 1:
          return candidates[0]
    return None
```

Algoritmo 1: Ligação com a Wikidata usando apenas o título.

Se nenhum candidato for gerado na primeira fase ou nenhum for seleccionado da lista de candidatos o algoritmo tenta expandir as entidades mencionadas no título com base no texto da notícia, explorando um padrão: uma personalidade mencionada no título por uma versão curta do seu nome (e.g.: apenas o apelido) é normalmente referida no texto da notícia por um nome mais completo.

O algoritmo identifica todas as pessoas mencionadas no texto da notícia, usando a componente descrita na Secção 6.1, e seleciona apenas as que têm pelo menos um nome em comum com o nome da personalidade referida no título, gerando assim uma entidade expandida, e assumindo que corresponde à mesma entidade referida no título.

Se do processo resulta apenas uma entidade expandida e se há uma similaridade de 1.0 com um dos candidatos anteriormente seleccionados da BC, esse candidato é escolhido. Caso contrário a entidade expandida é usada para fazer uma interrogação à BC e recolher uma nova lista de candidatos. Se nessa lista apenas há um candidato e a sua similaridade é de pelo menos 0.8 para com a entidade expandida, esse candidato é escolhido. Se há mais do que um candidato e apenas um tem uma similaridade 1.0 com a entidade expandida, esse é escolhido.

```
def article_text(expanded, candidates):
  if len(expanded) == 1:
    filtered = exact(expanded[0], candidates)
        if len(filtered) == 1:
          return filtered[0]

    x_candidates = get_candidates(expanded)
    if len(x_candidates) == 1:
      if jaro(expanded, x_candidates[0])>=0.8:
        return x_candidates[0]

    filtered = exact(expanded, x_candidates)
    if len(filtered) == 1:
      return matches[0]

  if len(expanded) > 1:
    filtered = []
    for e in expanded:
      exact_candidates = exact(e, candidates)
          for c in exact_candidates:
            filtered.append(c)
    if len(filtered) == 1:
      return filtered[0]

  return None
```

Algoritmo 2: Ligação com a Wikidata usando o texto da notícia para expandir as entidades reconhecidas no título.

Se do processo de expansão resultam várias entidades expandidas, filtramos candidatos da BC com similaridade 1.0 para com a entidade expandida, se existir apenas um, esse candidato

é o escolhido. Em qualquer outro caso aqui não descrito nenhum candidato é selecionado.

O Algoritmo 2 descreve este procedimento usando o texto da notícia.

Os resultados desta abordagem sobre o conjunto de dados anotados são descritos na Tabela 4. A classificação *incorrecta* corresponde a personalidades que não foram associadas ao identificador correcto na Wikidata, *não desambiguada* para as que o algoritmo não conseguiu seleccionar um identificador único de entre todos os candidatos ou a BC não retornou nenhum resultado.

Na Tabela 4 são reportadas duas avaliações, a primeira coluna descreve os resultados para o algoritmo base, sem mapeamentos. A segunda coluna considera a ambiguidade que uma referência pode ter em termos de personalidades que representa. Por exemplo, nos dados anotados, todas as menções a *Cavaco* correspondem à personalidade *Cavaco Silva*, com base nisto o algoritmo mapeia todas as referências a *Cavaco* para *Cavaco Silva*. Da mesma forma, todas as menções a *Marques Mendes* correspondem à personalidade *Luís Marques Mendes*. Fazendo uso destes mapeamentos reduzimos o número de entidades para as quais o algoritmo não consegue encontrar um identificador.

Classificação	**base**	**mapeamentos**
correcta	5 059	5 136
incorrecta	43	43
não desambiguada	246	169
Exactidão	0,93	0,96

Tabela 4: Resultados da avaliação do algoritmo de ligação com a Base de Conhecimento.

6.3. Classificador de Tipo de Relação

Optámos por decompor a tarefa de classificação da relação em duas tarefas: classificação do tipo de relação e direcção da relação, por oposição a desenvolver um único classificador que teria que distinguir de entre 5 classes possíveis, e com classes muito desequilibradas em termos de representatividade. Esta secção descreve o classificador desenvolvido para detectar o tipo de relação presente num título, tendo três classes possíveis: **opõe-se**, **apoia** e **outra**. Todas as experiências foram feitas com uma avaliação cruzada de quatro partições.[7]

[7] https://github.com/politiquices/data-releases

Avaliámos diferentes abordagens para a classificação supervisionada das relações presentes nos títulos, nomeadamente: um classificador SVM (Cortes & Vapnik, 1995) com um *kernel* linear, uma rede neuronal recorrente do tipo LSTM (Hochreiter & Schmidhuber, 1997), e uma rede neuronal do tipo *transformer*, o DistilBERT (Sanh et al., 2019).

Para o classificador SVM utilizámos como *features* uma abordagem baseada em vectores TF-IDF (Salton & Buckley, 1988), fazendo um pré-processamento do título, usando um padrão, de modo a identificar o contexto relevante, i.e.: o contexto no título que contém informação que descreve a relação: `<Ent1 X Ent2 contexto>` onde `X` = {*"diz a", "responde a", "sugere a", "diz que", "afirma que", "espera que", "defende que", "considera que", "sugere que", "quer saber se", "considera", "manda"*}. Sempre que o padrão não se verifica usamos todas as palavras do título para construir o vector, excepto o nome das personalidades.

A rede neuronal recorrente LSTM foi usada numa arquitectura bidireccional, ou seja, são usadas duas redes LSTM, ambas com uma dimensão de 128, uma lendo o título da primeira para a última palavra e outra da última para a primeira palavra, sendo que os dois estados finais de cada LSTM são concatenados e passados a um *layer* linear. Usamos *embeddings* pré-treinados para Português baseados no método FastText (*skip-gram*) de dimensão 50 (Hartmann et al., 2017). A rede foi treinada por 5 *epochs* com um *batch size* de 8.

O modelo DistilBERT foi treinado tendo como base um modelo pré-treinado para o Português (Abdaoui et al., 2020), sendo depois afinado no conjunto de dados anotado, i.e.: os pesos de todos os *layers* pré-treinados foram actualizados tendo em conta a tarefa de classificação da relação. A rede foi treinada durante 5 *epochs* com um *batch size* de 8.

A Tabela 5 descreve os resultados para os vários classificadores. Não há diferenças muito acentuadas em termos de performance entre os 3 classificadores, embora a abordagem usando o DistilBERT tenha alcançado os melhores resultados. Ao analisar os resultados notámos que há relações difíceis de classificar correctamente, nomeadamente as que contêm expressões idiomáticas, por exemplo:

- *José Lello diz que Nogueira Leite quer "abifar uns tachos"*
- *Louçã diz que "António Borges é o grilo falante" de Passo Coelho*

Relação	P	A	F_1
opõe-se	0,71	0,69	0,70
outra	0,69	0,69	0,69
apoia	0,65	0,69	0,67
Macro-Média	0,69	0,69	0,69

(a) SVM com um kernel linear.

Relação	P	A	F_1
opõe-se	0,75	0,64	0,69
outra	0,65	0,75	0,70
apoia	0,65	0,62	0,63
Macro-Média	0,69	0,68	0,68

(b) LSTM bidireccional.

Relação	P	A	F_1
opõe-se	0,74	0,76	0,75
outra	0,72	0,71	0,72
apoia	0,72	0,71	0,71
Macro-Média	0,73	0,72	0,72

(c) DistilBERT pré-treinado em Português.

Tabela 5: (P)recisão, (A)brangência e F_1 para uma avaliação com 4-partições e validação cruzada com diferentes classificadores.

Outras relações são ambíguas e difíceis de classificar sem mais nenhum outro contexto do que aquele presente no título. No conjunto de dados que tornamos público, todos os títulos contêm um URL para o texto da notícia.

Os resultados obtidos com as abordagens descritas, para os dados em Português, estão em linha com os resultados reportados anteriormente em dados em Inglês (Liang et al., 2019; Park et al., 2021).

6.4. Classificador de Direcção da Relação

O classificador da direcção tem duas classes possíveis. Como mostra a Tabela 2, o conjunto de dados tem um viés para com a classe $Ent_1 \rightarrow Ent_2$ representando 91,5% dos dados. Assim, optámos por desenvolver uma abordagem baseada em regras para detectar apenas a classe $Ent_1 \leftarrow Ent_2$, e sempre que nenhuma das regras se verifica o classificador atribui a classe $Ent_1 \rightarrow Ent_2$.

Definimos regras baseadas em padrões construídos com informação morfológica e sintáctica (Nivre et al., 2020) extraída do título com o spaCy, usando o mesmo modelo que o descrito na Secção 5. Extraímos a informação morfo-sintáctica de todas as palavras, incluindo para os verbos informação sobre a conjugação: a pessoa e o número. Os padrões definidos foram os seguintes:

- **VOZ_PASSIVA**: procuramos por padrões `<VERB><ADP>`, um verbo seguido de uma proposição. Verificamos se a voz passiva está presente e envolve as personalidades mencionadas no título: se a entidade Ent_1 tem uma dependência para com o verbo do tipo **acl**, se o verbo tem uma dependência para com a Ent_1 do tipo **nsubj:pass** ou se o verbo tem uma dependência para com a Ent_2 do tipo **obl:agent**.
- **VERBO_ENT2**: detecta o padrão morfológico `<PUNCT><VERB>`Ent_2`<EOS>`, um sinal de pontuação seguido de um verbo, e terminando com a Ent_2, restringido o verbo a ser conjugado na 3ª pessoa do singular do presente do indicativo, e onde `<EOS>` representa o final do título, significando que Ent_2 é a última palavra no texto do título.
- **NOUN_ENT2**: verifica se o padrão `<ADJ>?<NOUN><ADJ>?<ADP>`Ent_2`<EOS>` está presente no título, i.e.: um substantivo podendo ser precedido ou sucedido de um ou mais adjectivos terminando com a Ent_2, sendo que o substantivo é restrito a uma lista de substantivos pré-definida.

A Tabela 6 mostra alguns exemplos de títulos de notícias e das regras que foram aplicadas para detectar a direcção $Ent_1 \leftarrow Ent_2$. As regras são aplicadas de forma sequencial, pela mesma ordem aqui descritas, se nenhum dos padrões é detectado no título o classificador atribui a classe $Ent_1 \rightarrow Ent_2$.

A Tabela 7 contém os resultados deste classificador para o conjunto de dados anotados.

Os resultados mostram que o método proposto classifica correctamente grande parte da direcção das relações $Ent_1 \leftarrow Ent_2$, a única classe para as quais foram desenvolvidas regras, sem prejuízo para com a classe $Ent_1 \rightarrow Ent_2$.

7. Grafo Semântico

Os componentes descritos na secção anterior formam o processo de extracção de triplos RDF a partir dos títulos de notícias recolhidos.

O processo de extracção começa por fazer o reconhecimento de personalidades no título da notícia e a sua ligação com o identificador de cada personalidade na Wikidata. O processo de extracção continua se ambas as personalidades reconhecidas foram ligadas com um identificador

Título	Regra Aplicada
Marques Júnior elogiado por **Cavaco Silva** pela "integridade de carácter"	VOZ_PASSIVA
Passos Coelho é acusado de imaturidade política por **Santos Silva**	VOZ_PASSIVA
António Costa vive no "país das maravilhas"acusa **Assunção Cristas**	VERBO_ENT2
Passos Coelho "insultou 500 mil portugueses", acusa **José Sócrates**	VERBO_ENT2
Maria Luís Albuquerque sob críticas de **Luís Amado**	NOUN_ENT2
André Ventura diz-se surpreendido com perda de apoio de **Cristas**	NOUN_ENT2

Tabela 6: Exemplos de títulos e respectivas regras usadas para detectar a direcção da relação.

Direction	P	A	F_1	#Títulos
$Ent_1 \rightarrow Ent_2$	0,99	1,00	0,99	1 488
$Ent_1 \leftarrow Ent_2$	0,95	0,84	0,89	129
weighted avg.	0,98	0,98	0,98	1 517

Tabela 7: (P)recisão, A(brangência) e F_1 usando 3 regras baseadas em padrões.

na Wikidata, caso contrário o título é descartado. O tipo de relação presente no título é detectado com o modelo DistilBERT. Se a relação entre as personalidades no título da notícia não for classificada como **outra** o classificador da direcção da relação é também aplicado ao título, caso contrário o título é descartado.

Para todos os títulos considerados o resultado final é um triplo RDF ligando as personalidades através de uma relação de oposição ou apoio suportada por uma notícia. Os triplos RDF gerados são indexados num motor SPARQL (Jena, 2015) juntamente com um sub-grafo da Wikidata descrito na Secção 3.

O grafo gerado tem um total de 680 personalidades políticas, 107 partidos políticos e 10.361 notícias cobrindo um período de 25 anos, Está disponível *on-line* no formato *Terse RDF Triple Language*[8] e poder ser também explorado através de um interface textitweb.[9]

8. Conclusões e Trabalho Futuro

Este trabalho descreve em detalhe o processo de construção de um grafo semântico a partir de títulos de notícias de política.

Através de interrogações SPARQL e fazendo referência às várias propriedades, retiradas da Wikidata de cada personalidade, consegue-se explorar relações de apoio e oposição através de agregações por partidos políticos, cargos públicos, governos constitucionais, assembleias constituintes, entre outras, podendo assim formular-se interrogações mais complexas, por exemplo: *"Ministros do XXII Governo Constitucional que foram opostos por personalidades do PCP ou BE."*, obtendo-se como resposta a lista de ministros e os artigos que dão suporte às relações de oposição vindas do BE.

Umas das limitações deste trabalho prende-se com o título da notícia não conter informação suficiente para perceber que tipo de relação ou sentimento existe de uma personalidade para outra, ou a presença de expressões idiomáticas, que tornam difícil a classificação automática. Como trabalho futuro gostaríamos de explorar o texto da notícia de forma a complementar o título para melhorar a detecção da relação. Com base também no texto da notícia as relações poderiam ser enriquecidas, categorizando-as em tópicos, dando mais uma dimensão à relação, um contexto para o sentimento de suporte ou oposição.

Alguns títulos contêm uma relação mútua, por exemplo: *"Sócrates e Alegre trocam acusações sobre co-incineração"* ou *"Pinto da Costa rebate críticas de Pacheco Pereira"*, poderiam ser classificados com a direcção $Ent_2 \leftrightarrow Ent_1$, indicando neste caso que ambas as personalidades se acusam mutuamente.

Este trabalho também deixa em aberto oportunidades de realizar diversos estudos com base na estrutura do grafo, por exemplo: encontrar comunidades de apoio e oposição em função do tempo e verificar quais as mudanças dentro dessas comunidades. Pode-se também estudar triângulos políticos: se duas personalidades políticas, X e Y, sempre acusam ou defendem uma terceira personalidade Z, qual será a relação típica expectável entre X e Y?

Agradecimentos

Gostaríamos de agradecer ao Nuno Feliciano por todos os comentários dados durante a elaboração deste trabalho e à equipa do Arquivo.PT por disponibilizar acesso aos dados arquivados através de uma API e pela consideração deste trabalho para os prémios Arquivo.PT 2021. Ao Edgar Fe-

[8] https://github.com/politiquices/data-releases
[9] https://www.politiquices.pt

lizardo e ao Tiago Cogumbreiro pelas revisões extensivas ao artigo, e também aos revisores Sérgio Nunes e José Paulo Leal por todos os comentários e correcções apontadas.

Referências

Abdaoui, Amine, Camille Pradel & Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual BERT. Em *Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, 119–123. 10.18653/v1/2020.sustainlp-1.16.

de Arruda, Gabriel Domingos, Norton Trevisan Roman & Ana Maria Monteiro. 2015. An annotated corpus for sentiment analysis in political news. Em *10th Brazilian Symposium in Information and Human Language Technology (STIL)*, 101–110.

Balahur, Alexandra, Ralf Steinberger, Erik van der Goot, Bruno Pouliquen & Mijail Kabadjov. 2009. Opinion mining on newspaper quotations. Em *International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, 523–526. 10.1109/WI-IAT.2009.340.

Balahur, Alexandra, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen & Jenya Belyaeva. 2010. Sentiment analysis in the news. Em *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 655–662.

Baraniak, Katarzyna & Marcin Sydow. 2021. A dataset for sentiment analysis of entities in news headlines (SEN). *Procedia Computer Science* 192. 3627–3636. 10.1016/j.procs.2021.09.136.

Bassignana, Elisa & Barbara Plank. 2022. What do you mean by relation extraction? A survey on datasets and study on scientific relation classification. Em *60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 67–83. 10.18653/v1/2022.acl-srw.7.

Cortes, Corinna & Vladimir Vapnik. 1995. Support-vector networks. *Machine learning* 20(3). 273–297. 10.1007/BF00994018.

Gomes, Daniel, David Cruz, João Miranda, Miguel Costa & Simão Fontes. 2013. Search the past with the Portuguese Web Archive. Em *22nd International World Wide Web Conference*, 10.1145/2487788.2487934.

Gormley, Clinton & Zachary Tong. 2015. *Elasticsearch: The definitive guide*. O'Reilly Media.

Han, Xiaochuang, Eunsol Choi & Chenhao Tan. 2019. No permanent Friends or enemies: Tracking relationships between nations from news. Em *Conference of the North American Chapter of the ACL (NAACL)*, 1660–1676. 10.18653/v1/N19-1167.

Hartmann, Nathan, Erick Fonseca, Christopher Shulby, Marcos Treviso, Jéssica Silva & Sandra Aluísio. 2017. Portuguese word embeddings: Evaluating on word analogies and natural language tasks. Em *11th Brazilian Symposium in Information and Human Language Technology (STIL)*, 122–131.

Hochreiter, Sepp & Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8). 1735–1780. 10.1162/neco.1997.9.8.1735.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem & Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in Python. 10.5281/zenodo.1212303.

Iyyer, Mohit, Anupam Guha, Snigdha Chaturvedi, Jordan Boyd-Graber & Hal Daumé III. 2016. Feuding families and former Friends: Unsupervised learning for dynamic fictional relationships. Em *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 1534–1544. 10.18653/v1/N16-1180.

Jaro, Matthew A. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association* 84(406). 414–420. 10.1080/01621459.1989.10478785.

Jena, Apache. 2015. A free and open source java framework for building semantic web and linked data applications. Available online: https://jena.apache.org/ (accessed on 20 November 2022).

Liang, Shuailong, Olivia Nicol & Yue Zhang. 2019. Who blames whom in a crisis? Detecting blame ties from news articles using neural networksa. Em *AAAI Conference on Artificial Intelligence*, vol. 33 01, 655–662. 10.1609/aaai.v33i01.3301655.

Malyshev, Stanislav, Markus Krötzsch, Larry González, Julius Gonsior & Adrian Bielefeldt. 2018. Getting the most out of Wikidata: Semantic technology usage in Wikipedia's knowledge graph. Em *17th International*

Semantic Web Conference (ISWC), 376–394. 10.1007/978-3-030-00668-6_23.

Moreira, Silvio, David S Batista, Paula Carvalho, Francisco M Couto & Mario J Silva. 2013. Tracking politics with POWER. *Program: electronic library and information systems* 47(2). 120–135. 10.1108/00330331311313708.

Nivre, Joakim, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers & Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. Em *12th Language Resources and Evaluation Conference (LREC)*, 4034–4043.

O'Connor, Brendan, Brandon M. Stewart & Noah A. Smith. 2013. Learning to extract international relations from political context. Em *51st Annual Meeting of the ACL*, 1094–1104.

Park, Kunwoo, Zhufeng Pan & Jungseock Joo. 2021. Who blames or endorses whom? entity-to-entity directed sentiment extraction in news text. Em *Findings of the Association for Computational Linguistics (ACL-IJCNLP)*, 4091–4102. 10.18653/v1/2021.findings-acl.358.

Pontiki, Maria, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia Loukachevitch, Evgeniy Kotelnikov, Nuria Bel, Salud María Jiménez-Zafra & Gülşen Eryiğit. 2016. SemEval-2016 task 5: Aspect based sentiment analysis. Em *10th International Workshop on Semantic Evaluation (SemEval)*, 19–30. 10.18653/v1/S16-1002.

Prud'hommeaux, Eric, Steve Harris & Andy Seaborne. 2013. SPARQL 1.1 query language. W3C Technical Report, http://www.w3.org/TR/sparql11-query.

Salton, Gerard & Chris Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* 24(5). 513–523. 10.1016/0306-4573(88)90021-0.

Sanh, Victor, Lysandre Debut, Julien Chaumond & Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Em *Fifth Workshop on Energy Efficient Training and Inference of Transformer Based Models (EMC^2)*, on–line.

Santos, Diana & Paulo Rocha. 2001. Evaluating CETEMPúblico, a free resource for Portuguese. Em *39th Annual Meeting of the Association for Computational Linguistics*, 450–457. 10.3115/1073012.1073070.

Santos, Diana & Paulo Rocha. 2004. CHAVE: Topics and questions on the Portuguese participation in CLEF. Em *Working Notes for CLEF*, on–line.

Sarmento, Luís, Paula Carvalho, Mário J. Silva & Eugénio de Oliveira. 2009. Automatic creation of a reference corpus for political opinion mining in user-generated content. Em *1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion*, 29–36. 10.1145/1651461.1651468.

Schreiber, Guus & Yves Raimond. 2014. RDF 1.1 Primer. W3C Technical Report, https://www.w3.org/TR/rdf11-primer/.

Zimbra, David, Ahmed Abbasi, Daniel Zeng & Hsinchun Chen. 2018. The state-of-the-art in Twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems* 9(2). 10.1145/3185045.

Proposta recebida em Outubro 2022 e aceite para publicação em Março 2023.

Classificação da qualidade da argumentação em *tweets* no domínio da política brasileira

Argument Quality Assessment in Brazilian political tweets

Cássio Faria da Silva
Rede Gonzaga de Ensino Superior / Universidade Federal de São Carlos

Vânia Paula de Almeida Neris
Universidade Federal de São Carlos

Helena de Medeiros Caseli
Universidade Federal de São Carlos

Resumo

A argumentação é uma habilidade inerente à comunicação humana, tanto em situações orais quanto escritas. Argumentos bem fundamentados são importantes para amparar a tomada de decisões e aprendizado, assim como para a obtenção de conclusões amplamente aceitas. Como área de pesquisa, a argumentação é um campo multidisciplinar que estuda os processos de debate e raciocínio. Em linguística computacional, investigações têm sido realizadas para (i) identificar argumentos e suas unidades e (ii) gerar ou (iii) avaliar a qualidade dos argumentos. No entanto, a maioria dos trabalhos atuais se concentra na mineração de argumentos em textos formais em inglês. Neste artigo, foi avaliada a qualidade da argumentação em *tweets* de domínio político, escritos em português do Brasil, usando algoritmos tradicionais de aprendizado de máquina – como Regressão Logística, *K-Nearest Neighbors*, Árvores de Decisão, Máquinas de Vetores Suporte (SVM), Floresta Aleatória e *Naive Bayes* – e também um ajuste fino de dois modelos neurais (BERTimbau e RobertaTwitterBR). Além de trazer resultados práticos para a avaliação da qualidade da argumentação em um gênero textual desafiador, como o Twitter, e em um domínio controverso, como a política brasileira, este artigo também visa suprir a carência de trabalhos que avaliem automaticamente a qualidade dos argumentos em português. Dentre os algoritmos de classificação avaliados, o modelo obtido a partir do ajuste fino do BERTimbau apresentou os melhores resultados com uma precisão de 69,65% quando foram consideradas todas as classes e de 100,00% para as mensagens de alta qualidade de argumentação.

Palavras chave

avaliação da qualidade da argumentação, *tweet*, BERT, política brasileira

Abstract

Argumentation is an inherent skill in human communication, both in oral and written situations. Well-founded arguments are important to support decision-making and learning, as well as to reach widely accepted conclusions. As a research area, argumentation is a multidisciplinary field that studies the processes of debate and reasoning. In computational linguistics, investigations have been carried out to (i) identify arguments and their units and (ii) generate or (iii) evaluate the quality of arguments. However, most current work focuses on argument mining in formal English texts. In this article, we evaluated the quality of argumentation in political domain tweets, written in Brazilian Portuguese, using traditional machine learning algorithms – such as Logistic Regression, KNearest Neighbor, Decision Trees, Support Vector Machines (SVM), Random Forest and Naive Bayes – and also a fine-tuning of two neural models (BERTimbau and RobertaTwitterBR). In addition to bringing practical results for the assessment of argumentation quality in a challenging textual genre, such as Twitter, and in a controversial domain, such as Brazilian politics, this article also aims to fill in the lack of works that automatically assess the quality of arguments in Portuguese. Among the evaluated classification algorithms, the model obtained from the fine-tuning of BERTimbau presented the best results, with an accuracy of 69.65% when all classes were considered and 100.00% for messages with high quality of argumentation.

Keywords

argument quality assessment, tweet, BERT, Brazilian politics

DOI: 10.21814/lm.15.1.387

1. Introdução

Um argumento é uma afirmação (ou conclusão) acompanhada por um número arbitrário de premissas que justificam, fundamentam, apoiam, defendem ou explicam a afirmação (Potthast et al., 2019). Argumentos bem fundamentados são importantes para amparar a tomada de decisões e aprendizado, assim como para a obtenção de conclusões amplamente aceitas. A argumentação (capacidade de produzir argumentos) é uma habilidade inerente à comunicação humana tanto em situações orais quanto escritas. Como área de pesquisa, a argumentação é um campo multidisciplinar que estuda os processos de debate e raciocínio (Habernal & Gurevych, 2017). Para Eemeren & Grootendorst (2003), a argumentação consiste em uma ou mais sentenças nas quais várias premissas são apresentadas para sustentar uma conclusão. As sentenças que fazem parte da argumentação constituem uma expressão completa que visa convencer um interlocutor.

Em linguística, estuda-se a argumentação em textos em linguagem natural (Stab & Gurevych, 2017a). Na ciência da computação, a identificação ou avaliação automática da argumentação é estudada no campo da inteligência artificial (Bench-Capon & Dunne, 2007). Ao combinar essas duas áreas de pesquisa, em linguística computacional ou no processamento de linguagem natural (PLN), investigações têm sido realizadas para (i) identificar argumentos e suas unidades e (ii) gerar ou (iii) avaliar a qualidade de tais argumentos. Mais especificamente, as tarefas mais comumente investigadas são: a mineração de unidades de argumentação (Al-Khatib et al., 2016; Habernal & Gurevych, 2015, 2017), a detecção de evidências que apoiam reivindicações[1] (Rinott et al., 2015) e a identificação de relações argumentativas (Peldszus & Stede, 2015). Outros trabalhos classificaram esquemas de argumentação (Feng et al., 2014), realizam a análise de estruturas gerais de argumentação (Wachsmuth et al., 2015; Stab & Gurevych, 2017a) e geram reivindicações (Bilu & Slonim, 2016).

Algumas teorias ou dimensões da qualidade da argumentação foram avaliadas computacionalmente (Stab & Gurevych, 2017b; Wachsmuth et al., 2017d; Zhang et al., 2016). No entanto, de acordo com Wachsmuth et al. (2017b), ainda não se constituiu um conceito geral para a qualidade da argumentação ou uma definição clara de suas dimensões. Apesar da falta do conceito geral, tarefas relacionadas à argumentação computacional — como mineração, geração, identificação de argumentos e sua avaliação — têm se mostrado relevantes em atividades como apoio à escrita e assistência à discussão (Stab & Gurevych, 2017b; García-Gorrostieta et al., 2018).

No campo da comunicação, como bem pontuado por Lytos et al. (2019), a internet e as redes sociais são, hoje, o meio de comunicação mais utilizado. Consistem em espaços que permitem a emissão de opiniões sobre qualquer assunto e são fonte para a produção de um grande volume de textos com potencial argumentativo. De especial interesse para este trabalho é a rede social Twitter[2], bastante utilizada para troca de informações e opiniões sobre política no Brasil.

A argumentação no Twitter, além de envolver características específicas do gênero textual, também é permeada pelas necessidades de comunicação, pelo contexto histórico e pelo assunto da mensagem (Marcuschi et al., 2002), como exemplificado no exemplo (1).[3]

(1) @gleisi Prezada **coxa** vc não tem vergonha nessa cara reformada com dinheiro público**??** **A merda do seu partido** que se diz do povo não fez nada disso e ainda **enfiou dinheiro nosso no rabo** do **joesley** do **eike** e do **Marcelo**. **Esse discursinho vagabundo não cola mais**.

Entre os indícios linguísticos que impactam a qualidade da argumentação neste exemplo (1), podem ser citados: (i) referências pejorativas como "coxa"; (ii) a repetição de sinais de pontuação ("??") que indicam indignação; (iii) a presença de discurso de ódio contra um partido político em "a merda do seu partido" e contra a autora do *tweet* original (ao qual este é uma resposta) em "esse discursinho vagabundo"; (iv) expressões coloquiais como a expressão idiomática em "enfiou [...] no rabo" e gíria em "não cola mais"; (v) uma tentativa de tornar o argumento mais pessoal em "dinheiro nosso"; e (vi) contexto histórico citando "joesley" (Joesley Batista), "eike" (Eike Batista) e "Marcelo" (Marcelo Odebrecht), três empresários brasileiros que ficaram nacionalmente conhecidos em um dos escândalos políticos ocorridos no Brasil.

Ao lidar com textos do Twitter, escritos em português, esse trabalho enfrenta desafios não

[1] Reivindicações, no contexto dos trabalhos relacionados à mineração de unidades de argumentação, se referem a quaisquer declarações ou afirmações que possuam propósito argumentativo.

[2] https://twitter.com/

[3] Todos os exemplos de *tweets* apresentados neste artigo são transcritos exatamente como foram publicados pelos seus autores. O destaque (em negrito) de alguns trechos foi inserido pelos autores deste artigo para enfatizar.

presentes na maioria dos trabalhos da literatura, que focam na mineração de argumentos em textos formais em inglês. Em relação ao idioma, vale destacar que a língua portuguesa é uma das mais faladas no mundo, com mais de 280 milhões de falantes. Nesse sentido, são notáveis os avanços e a importância que tem sido dada atualmente ao desenvolvimento e aprimoramento dos recursos de PLN para o português. Isso pode estar relacionado, em grande parte, aos aspectos da globalização e da popularização do acesso ao ambiente online, que são fatores facilitadores do intercâmbio sociolinguístico dessa língua.[4]

Em relação ao gênero textual, é importante salientar que embora hajam recursos linguísticos valiosos para o português, a maioria foi desenvolvida para lidar com textos formais, bem escritos e autocontidos, como artigos de jornal. No entanto, o cenário atual da comunicação nas mídias sociais exige que os recursos de PLN sejam capazes de lidar com textos cheios de desafios, como a presença de gírias, linguagem figurada, erros de português e uso do "internetês" (vocabulário próprio composto por termos e abreviações usualmente encontrados em textos de mídias sociais). Nesse sentido, embora fenômenos linguísticos que trazem indícios de uma qualidade da argumentação boa ou ruim possam ser identificados nos *tweets*, não há garantia de que os recursos e as ferramentas disponíveis hoje para o processamento automático também sejam capazes de identificá-los.

No que diz respeito à avaliação da qualidade da argumentação, que é o foco deste artigo, desde o esquema argumentativo de Toulmin (2003), estudos têm sido realizados para simplificar a compreensão da estrutura e determinar a importância dos elementos argumentativos do texto. Wachsmuth et al. (2017b) propuseram uma taxonomia composta por três dimensões para avaliar a qualidade da argumentação: retórica, lógica e dialética. No entanto, desde então, poucos estudos (Potthast et al., 2019; Wachsmuth & Werner, 2020; Skitalinskaya et al., 2021) se dedicaram à avaliação automática da qualidade da argumentação, muito menos em gêneros textuais cujos textos apresentam conteúdos distantes da norma linguística padrão e longe da própria noção convencional de argumentação.

Entre os trabalhos mais recentes que exploraram essa temática pode-se citar o de Gretz et al. (2019), que aplicaram o modelo de dimensões de qualidade da argumentação proposto por Wachsmuth et al. (2017b) para avaliar a qualidade da argumentação no corpus IBM-Rank-30k, contendo 30.497 argumentos provenientes de clubes de debates, no idioma inglês, onde os participantes são incentivados a redigirem textos bem escritos e com argumentos de alta qualidade. Foram apresentadas avaliações com três métodos baseados em BERT: BERT-Vanilla, BERT-Finetune e BERT-FT$_{\text{TOPIC}}$.[5] Os experimentos evidenciaram um melhor desempenho com o BERT-FT$_{\text{TOPIC}}$, com um coeficiente de correlação de Pearson de 0,53, calculado em relação a uma fração superior e inferior dos argumentos que estão nos extremos da escala de qualidade da argumentação. De acordo com os autores, as dimensões de Relevância e Eficácia Globais são as mais indicativas para os índices gerais de qualidade. Outros trabalhos que usaram o BERT são os de (Fromm et al., 2019) e (Reimers et al., 2019), mas ambos para mineração de argumentos e não para a avaliação da qualidade da argumentação.

Considerando as lacunas identificadas nos estudos que focam na avaliação automática da qualidade da argumentação, esse trabalho é inédito no sentido de que é o primeiro a investigar métodos para avaliar automaticamente a qualidade da argumentação: (i) em postagens de redes sociais e (ii) em português. Assim, neste artigo investigou-se como utilizar o aprendizado de máquina para classificar a qualidade da argumentação em *tweets* de domínio político escritos em português.

As questões de pesquisa que se busca responder com este estudo são:

(Q1) É possível predizer automaticamente a qualidade da argumentação em *tweets* no domínio da política brasileira?

(Q2) Como identificar automaticamente os *tweets* no campo da política brasileira com bons argumentos?

Para tanto, foram usados algoritmos de Aprendizado de Máquina (AM) baseados em *features*, como regressão logística (LR), *K-Nearest Neighbor* (KNN), árvore de decisão (DT), máquinas de vetor de suporte (SVM), Flo-

[4]Para se ter uma ideia desse crescente interesse, o Centro Regional de Estudos para a Desenvolvimento da Sociedade da Informação (CETIC) observou o comportamento de brasileiros maiores de 16 anos na internet entre 23 de junho e 8 de julho de 2020. Seus achados mostraram que 49% dos internautas realizavam atividades laborais e 72% buscavam informações relacionadas à saúde na internet. Para mais informações, consulte: https://cetic.br/media/docs/publicacoes/2/20200817133735/painel_tic_covid19_1edicao_livro%20eletr%C3%B4nico.pdf

[5]Consiste em concatenar o tópico da discussão ao argumento, separando-os por um delimitador.

resta Aleatória (RF) e *Naive Bayes* (NB) — e também um ajuste fino de um modelo BERT e um RoBERTa. Os experimentos realizados com o BERT alcançaram uma precisão de 100% para as mensagens consideradas com Alta qualidade de argumentação. Enquanto os algoritmos baseados em *features* obtiveram precisões médias que variaram de 32% a 54%.

Este artigo está organizado em cinco seções, além desta introdução. Na seção 2, são apresentados os trabalhos mais relacionados a esta pesquisa. A seção 3 descreve brevemente o processo de construção do corpus e as pistas linguísticas propostas por Silva et al. (2021) e adotadas neste artigo para definir uma boa qualidade de argumentação em um *tweet* do domínio da política brasileira. Na seção 4, apresentam-se os experimentos realizados para avaliar a qualidade da argumentação. Os resultados de tais experimentos são descritos e analisados na seção 5. Por fim, na seção 6, são feitas algumas considerações finais, com a apresentação de limitações e apontamentos sobre trabalhos futuros.

2. Trabalhos relacionados

Avaliar a validade, a qualidade e a força dos argumentos representa um desafio inerente ao discurso argumentativo. Vale destacar que existem fundamentos teóricos e diversas teorias normativas para embasar a tarefa, tais como (i) o modelo argumentativo de Toulmin (2003); (ii) os esquemas e questões críticas de Walton & Walton (1989); (iii) o modelo ideal de argumentação crítica na abordagem pragma-dialética (Eemeren & Grootendorst, 1987), em que as falácias são consideradas movimentos incorretos em uma discussão cujo objetivo é a resolução bem sucedida de uma disputa; e (iv) o estudo das falácias (Boudry et al., 2015). No entanto, julgar os critérios qualitativos da argumentação cotidiana ainda representa um desafio para os estudiosos e profissionais da argumentação (Weltzer-Ward et al., 2009; Swanson et al., 2015; Rosenfeld & Kraus, 2016).

Apesar desses fundamentos teóricos, os métodos e técnicas já propostos para avaliar a qualidade dos argumentos não concordam sobre quais critérios devem ser considerados, nem mesmo sobre se a qualidade deve ser avaliada do ponto de vista teórico ou prático. Wachsmuth et al. (2017a) tentaram elucidar a questão de quão diferentes são as visões teóricas e práticas da qualidade da argumentação. Do ponto de vista teórico, apontam que a convicção é entendida como a principal qualidade lógica, e sustentam o fato de que a avaliação teórica da qualidade da argumentação permanece complexa. Eles também apontaram que as abordagens práticas indicam o que focar para simplificar a teoria, enquanto a teoria parece benéfica para orientar a avaliação da qualidade na prática.

Na mesma direção, outros estudos têm buscado avaliar a relevância dos argumentos por meio da identificação de sentenças argumentativas com a posterior avaliação da importância/relevância delas. Potthast et al. (2019) avaliaram o grau de relevância de um conjunto de argumentos no corpus args.me[6] (Wachsmuth et al., 2017c), constituído por textos escritos em inglês, provenientes de cinco portais de debates, com o objetivo de construir um motor de busca de argumentos na web. Quarenta anotadores avaliaram a relevância de cada um dos 437 argumentos relacionados a 40 tópicos selecionados, além de sua qualidade retórica, lógica e dialética. Dos 437 argumentos anotados, 208 foram marcados a favor e 195 foram marcados como contrários ao tópico em questão, além de 34 que foram anotados como não argumentativos. Pontuações de 1 (baixa) a 4 (alta) foram atribuídas às dimensões de qualidade de argumentação — retórica (*rethoric*), lógica (*logic*) e dialética (*dialectic*) — e à sua relevância (*relevance*). A distribuição das pontuações (de 1 a 4) produzidas nesse trabalho pode ser vista na Figura 1.

As pontuações de relevância indicam que muitos argumentos relevantes (classificados como 4) foram recuperados do corpus args.me. Outros trabalhos também investigaram o aspecto da relevância dos textos argumentativos (Wachsmuth et al., 2017d; Gleize et al., 2019).

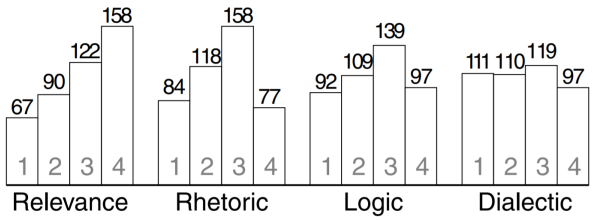

Figura 1: Distribuições de pontuação por dimensões de relevância e qualidade (Potthast et al., 2019).

Habernal & Gurevych (2016) sugeriram que a avaliação da qualidade da argumentação deve ser feita comparando argumentos, enquanto outros trabalhos (Persing & Ng, 2015; Wachsmuth et al., 2017b) relataram avaliações da qualidade dos argumentos individuais com resultados satisfatórios.

Vale mencionar que, neste trabalho, a quali-

[6]Disponível em: `https://www.args.me`

dade de argumentação é avaliada para um *tweet* isoladamente e não com base na comparação entre *tweets*.

Trabalhos mais recentes (Wachsmuth et al., 2017b; Lauscher et al., 2020; Wachsmuth & Werner, 2020; Skitalinskaya et al., 2021) têm utilizado uma taxonomia que visa avaliar aspectos individuais com base nas características da estrutura argumentativa, como o apelo emocional empregado, a organização da sentença e a credibilidade do autor da mensagem. A taxonomia escolhida para ser utilizada neste trabalho é apresentada na seção 2.3.

Embora passos importantes em AM tenham sido dados rumo ao processamento de *tweets*, como a detecção de argumentos, alegações e evidências, Schaefer & Stede (2021) demonstraram que várias áreas ainda estão em desenvolvimento, como o emprego de abordagens de AM utilizando aprendizado profundo e técnicas de redes neurais, além de pesquisas em outros idiomas diferentes do inglês.

Antes de detalhar a taxonomia adotada nesta pesquisa, as duas próximas subseções citam alguns trabalhos que investigaram a qualidade da argumentação em vários domínios, mais usuais do que o da política, aqui investigado.

2.1. Avaliação da qualidade da argumentação em redações de alunos

Fornecer aos estudantes *feedback* útil a respeito da persuasividade de seus argumentos tem sido objeto de estudos recentes em mineração de argumentos. Alguns corpora foram construídos, principalmente em inglês (Stab & Gurevych, 2014; Persing & Ng, 2015; Stab & Gurevych, 2017a; Carlile et al., 2018; Putra et al., 2021), e estudos têm sido realizados para avaliar a qualidade da argumentação em redações escritas por estudantes.

Stab & Gurevych (2017b) apontaram que as premissas de um argumento bem fundamentado devem fornecer evidências suficientes para aceitar ou rejeitar sua afirmação. Para chegar a essa conclusão, um corpus composto por 402 redações (Stab & Gurevych, 2017a) foi anotado com estruturas de argumentação. Os autores relataram uma pontuação de concordância de Fleiss entre os anotadores de 0,877, o que indicou que os anotadores foram capazes de identificar de forma confiável as principais reivindicações presentes em redações persuasivas. Para a construção do modelo computacional, os autores experimentaram máquinas de vetores de suporte (SVMs) e redes neurais convolucionais (CNNs) e alcançaram 84% de precisão na tarefa de identificar argumentos insuficientemente suportados. O corpus final e as diretrizes de anotação estão disponíveis.[7]

Carlile et al. (2018), com o objetivo de realizar futuras avaliações no nível de persuasão dos argumentos em redações de alunos, construíram um corpus composto por 102 redações selecionadas aleatoriamente do *Argument Annotated Essays* (Stab & Gurevych, 2014). O corpus foi anotado com árvores de argumentos, pontuações de persuasão e atributos de componentes de argumentos que, segundo os autores, impactam essas pontuações. Os autores relatam uma concordância que variou de 0,549 a 1,000 (valores de α de Krippendorff (2011)), de acordo com o atributo. O corpus anotado e as diretrizes de anotação estão disponíveis.[8] Na mesma linha de pesquisa, outros trabalhos avaliaram a qualidade da argumentação em teses e trabalhos acadêmicos (García-Gorrostieta et al., 2018; García-Gorrostieta & López-López, 2018).

Putra et al. (2021) criaram um corpus com anotação estrutural argumentativa para redações escritas em inglês como língua estrangeira e também definiram um esquema de anotações. O corpus anotado produzido como resultado desse esforço, o ICNALE-AS, inclui 434 redações enviadas por estudantes de inglês de várias nações asiáticas. A análise de concordância inter-anotador mostrou que o esquema de anotação proposto é estável, alcançando um coeficiente de Cohen de 0,66.

Apesar dos trabalhos apresentados nesta seção terem avaliado a qualidade da argumentação, eles o fizeram para um gênero textual formal e menos desafiador do que o Twitter. Os textos formais admitem forma e estilo e, como consequência, podemos presumir que apresentam uma boa argumentação com argumentos encadeados baseados em fatos reais e frases estruturadas. Todos esses aspectos podem não ocorrer em argumentos feitos em postagens no Twitter; e, mesmo quando ocorrem, devem ser ressignificados de acordo com esse gênero e com a maneira como os usuários das mídias sociais utilizam esses aspectos, o que resulta, por exemplo, em uma possível argumentação baseada em *fake news*.

[7] Disponível em: https://www.ukp.tu-darmstadt.de/data/

[8] Disponível em http://www.hlt.utdallas.edu/~zixuan/EssayScoring

2.2. Avaliação da qualidade da argumentação em mensagens de fóruns e portais de discussão

Para investigar a identificação de postagens persuasivas em fóruns de discussão (como Change My View[9]), Wei et al. (2016) criaram um corpus de mensagens argumentativas selecionando tópicos com mais de 100 comentários publicados entre janeiro de 2014 e janeiro de 2015, totalizando 1.785 tópicos com 374.472 comentários.

Experimentos foram realizados para encontrar quais *features* seriam as mais adequadas para predizer comentários persuasivos. Entre as *features* investigadas estavam: (i) o número de palavras e frases, (ii) a presença/ausência de pontuação, (iii) as *part-of-speech tags* (POS tags), entre outras. Depois disso, os autores calcularam a correlação entre a pontuação humana de um comentário argumentativo e o conjunto de atributos da reputação do autor da postagem. Então, para a tarefa de classificação de comentários, três conjuntos de *features* foram avaliados, incluindo as que consideravam apenas as características superficiais do texto, aquelas que consideravam a interação social e aquelas focadas na argumentação propriamente dita. Os resultados experimentais mostraram que as *features* baseadas em argumentação são mais informativas no estágio inicial da discussão e que a eficácia das *features* de interação social aumenta com o número de comentários na discussão.

Habernal & Gurevych (2016) investigaram a comparação qualitativa entre pares de argumentos: dados dois argumentos (como mostrado na Figura 2), um deles deve ser selecionado como o mais convincente. A pesquisa produziu os seguintes resultados: (i) um corpus anotado composto por 16.000 pares de argumentos, escritos em inglês, (ii) a análise dos dados anotados em relação às propriedades definidas como convincentes, e (iii) modelos computacionais gerados com SVM e uma arquitetura neural bidirecional de memória de longo prazo (BLSTM). O modelo SVM superou o modelo BLSTM (78% versus 76% de precisão, respectivamente) com uma diferença sutil, mas significativa, segundo os autores. Os—dados anotados e os códigos estão disponíveis publicamente[10].

Wachsmuth & Werner (2020) investigaram a avaliação automática de argumentos extraídos de portais de debate usando a taxonomia proposta por Wachsmuth et al. (2017b). Para tanto, com-

Argument 1	Argument 2
physical education should be mandatory cuhz 112,000 people have died in the year 2011 so far and it's because of the lack of physical activity and people are becoming obese!!!!	YES, because some children don't understand anything excect physical education especially rich children of rich parents.

Figura 2: Exemplo de argumentos sobre a obrigatoriedade da educação física (Habernal & Gurevych, 2016).

binaram *features* textuais (como mostrado nos exemplos da Figura 3) e SVM. Os exemplos na Figura 3 demonstram algumas *features* textuais[11] que, segundo os autores, podem ser preditivas de certas dimensões. Os autores relataram que o tamanho limitado do corpus dificultou a adoção de recursos mais complexos para a avaliação da qualidade. No entanto, eles destacaram que modelar a subjetividade por meio de *features* textuais pode ser propício para avaliar as dimensões lógica e dialética. Quanto à dimensão retórica, apontaram três aspectos difíceis: clareza, credibilidade e apelo emocional.

Estudos mais recentes (Fromm et al., 2022; Skitalinskaya et al., 2021; Toledo et al., 2019) usaram modelos BERT (Devlin et al., 2019) com ajuste fino (*fine-tuning*) em diferentes conjuntos de dados para avaliar a qualidade da argumentação. Skitalinskaya et al. (2021) investigaram a avaliação da qualidade da argumentação, independentemente dos aspectos discutidos. Para tanto, um corpus de 377.000 pares de comentário e argumento foi gerado a partir do fórum de discussão `kialo.com`,[12] abrangendo diversos temas de política, ética, entretenimento e outros. Duas tarefas foram realizadas: (i) avaliar qual afirmação de um par de comentários é melhor e (ii) classificar todas as versões de uma afirmação por qualidade. Os experimentos com regressão logística baseada em *embeddings* e redes neurais baseadas em *transformers* mostraram resultados promissores, sugerindo que os indicadores aprendidos generalizam bem entre os tópicos.

Para a primeira tarefa, os experimentos conduzidos com Sentence-BERT (SBERT) (Reimers & Gurevych, 2019) apresentaram uma precisão de até 77,7%. Na segunda tarefa, o modelo baseado em SBERT superou todas as abordagens

[9]Disponível em: `https://www.reddit.com/r/changemyview`

[10]Disponível em: `https://github.com/UKPLab/acl2016-convincing-arguments`

[11]As *features* textuais apresentadas na pesquisa de Wachsmuth & Werner (2020) são semelhantes às pistas linguísticas (seção 3.1) estudadas neste trabalho.

[12]Disponível em `https://kialo.com`

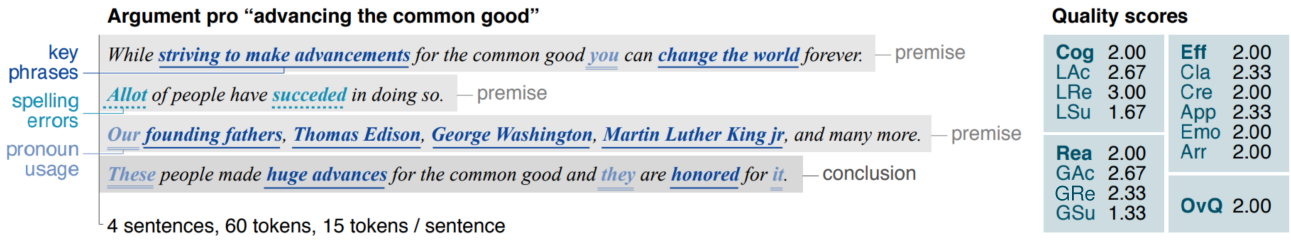

Figura 3: Exemplo de um argumento e os recursos linguísticos que afetam sua qualidade (Wachsmuth & Werner, 2020).

testadas alcançando até 0,73 em correlação de Pearson e 0,72 em correlação de Spearman.

Esses trabalhos investigaram a avaliação da argumentação em mensagens de fóruns de discussão e portais de debate (Wei et al., 2016; Habernal & Gurevych, 2016) e redações de alunos (Stab & Gurevych, 2017b; Carlile et al., 2018; Wachsmuth et al., 2016). No entanto, os *tweets* no domínio da política brasileira, nos dias atuais, possuem características mais desafiadoras do que as encontradas em fóruns e portais de debate, como: (i) um número muito limitado de caracteres, o que dificulta o uso de estratégias de argumentação linguística; e (ii) a presença de discurso incivil e intolerante (Rossini, 2019, 2022), decorrente da polarização e agressividade presentes no cenário atual da política brasileira, que traz a necessidade de estratégias para identificar discurso de ódio e polaridade, por exemplo.

2.3. Taxonomia de Wachsmuth et al.

Wachsmuth et al. (2017b) conduziram uma pesquisa sobre a qualidade da argumentação considerando tanto a teoria da argumentação quanto as perspectivas de mineração de argumentos. Com base nesse estudo, foi proposta a Taxonomia da Qualidade da Argumentação, cujas dimensões são utilizadas para definir a "qualidade". A Figura 4 ilustra esta taxonomia, com todas as dimensões dela.

De acordo com essa taxonomia, a qualidade da argumentação pode ser dividida nas dimensões lógica, retórica e dialética (Blair, 2012), descritas a seguir:

- A **dimensão lógica** refere-se à estrutura e composição de um argumento. Um argumento de alta qualidade lógica é baseado em premissas aceitáveis e as combina de forma convincente para apoiar a afirmação do argumento. Está relacionado com a irrefutabilidade lógica do argumento.
- A **dimensão retórica**, ao contrário, inclui noções de eficácia persuasiva, linguagem correta, precisão e estilo. Um argumento de alta qualidade retórica é bem escrito e atraente para o público e está relacionado à eficácia retórica do argumento. Especificamente, um argumento é retoricamente eficaz se for capaz de convencer o público-alvo (ou corroborar a concordância) que a posição do autor sobre o assunto é a correta.
- A **dimensão dialética** captura a contribuição de um argumento para o discurso. Um argumento de alta qualidade dialética é útil para apoiar a tomada de decisão cooperativa ou para resolver conflitos. O argumento é razoável se for capaz de contribuir para a resolução do problema de uma forma que seja suficientemente aceitável pelo público-alvo.

Wachsmuth et al. (2017b) testaram a taxonomia em um experimento de anotação com o Dagstuhl-15512-ArgQuality,[13] que contém 320 textos argumentativos com notas atribuídas por três anotadores que compõem os 15 aspectos da taxonomia. Nesse processo de anotação, cada texto foi primeiramente classificado como argumentativo ou não. Em seguida, para os textos argumentativos, todos os aspectos foram avaliados com notas 1 (baixo), 2 (médio) ou 3 (alto), além da opção "não posso julgar."

Na Figura 5, pode-se ver as pontuações atribuídas pelos três anotadores (A, B e C) em dois textos produzidos em resposta à pergunta "garrafas plásticas de água devem ser banidas?". O valor mais alto em cada coluna está marcado em negrito. A linha inferior representa a maioria dos votos dos três anotadores.[14]

A Figura 6 mostra os resultados deste experimento de anotação para os 304 textos do corpus

[13]Corpus Dagstuhl-15512-ArgQuality disponível em: http://arguana.com/

[14]A dimensão lógica mede a convicção (Co) e é composta por 3 aspectos: aceitabilidade local (LA), relevância local (LR) e suficiência local (LS). A dimensão retórica mede a eficácia (Ef) e é composta por 5 aspectos: credibilidade (Cr), apelo emocional (Em), clareza (Cl), adequação (Ap) e organização (Ar). Por fim, a dimensão dialética mede a razoabilidade (Re) e é composta por 3 aspectos: aceitabilidade global (GA), relevância global (GR) e suficiência global (GS).

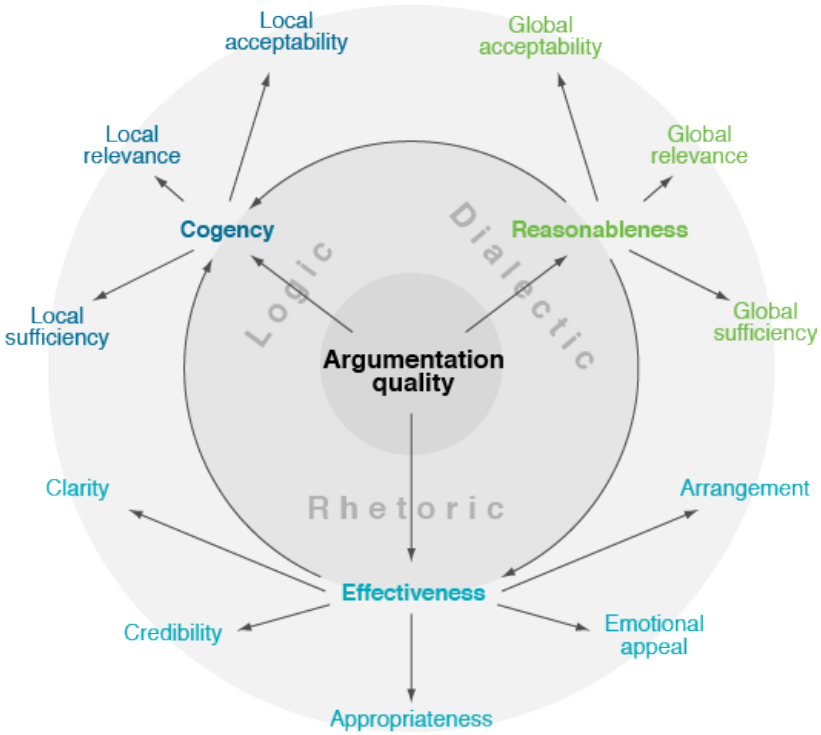

Figura 4: Taxonomia da Qualidade da Argumentação de Wachsmuth et al. (2017b).

Arguments	**Pro** Water bottles, good or bad? Many people believe plastic water bottles to be good. But the truth is water bottles are polluting land and unnecessary. Plastic water bottles should only be used in emergency purposes only. The water in those plastic are only filtered tap water. In an emergency situation like Katrina no one had access to tap water. In a situation like this water bottles are good because it provides the people in need. Other than that water bottles should not be legal because it pollutes the land and big companies get 1000% of the profit.	**Con** Americans spend billions on bottled water every year. Banning their sale would greatly hurt an already struggling economy. In addition to the actual sale of water bottles, the plastics that they are made out of, and the advertising on both the bottles and packaging are also big business. In addition to this, compostable waters bottle are also coming onto the market, these can be used instead of plastics to eliminate that detriment. Moreover, bottled water not only has a cleaner safety record than municipal water, but it easier to trace when a potential health risk does occur. (http://www.friendsjournal.org/bottled-water) (http://www.cdc.gov/healthywater/drinking/bottled/)

Scores	Co	LA	LR	LS	Ef	Cr	Em	Cl	Ap	Ar	Re	GA	GR	GS	Ov	Co	LA	LR	LS	Ef	Cr	Em	Cl	Ap	Ar	Re	GA	GR	GS	Ov
Annotator A	3	3	3	2	3	3	3	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	3	2	3	3	3	3	3	3	3	**3**
Annotator B	2	2	3	2	1	2	2	2	2	1	2	2	2	1	2	2	3	3	2	2	3	2	3	3	2	3	3	2	2	**3**
Annotator C	2	3	3	2	2	2	2	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	2	1	3	3	3	3	3	3	3	**3**
Majority score	2	3	3	2	2	2	2	3	3	3	3	3	3	2	**3**	**3**	3	3	3	3	3	2	3	3	3	3	3	3	3	**3**

Figura 5: Pontuações de cada anotador e pontuação majoritária para todas as dimensões de qualidade. Os argumentos são sobre o tópico "banir garrafas plásticas de água" (Wachsmuth et al., 2017b).

classificados como argumentativos por todos os anotadores: (a) a distribuição das pontuações majoritárias para cada dimensão; (b) o α de Krippendorff (2011) utilizado para medir a concordância entre anotadores; (c) a correlação para cada par de dimensões, calculada com base na média das correlações de todos os anotadores. O valor mais alto em cada coluna é destacado em negrito.

Nesta pesquisa foi escolhida a dimensão retórica da taxonomia de Wachsmuth et al. (2017b) para a avaliação da qualidade da argumentação em postagens do Twitter no domínio da política no Brasil. Assim como Gretz et al. (2019), foram selecionados os aspectos da taxonomia da qualidade da argumentação que possuíam características semelhantes às *features* linguísticas disponíveis. Desse modo, a escolha pela dimensão retórica se deu com base nas pistas linguísticas apontadas como relevantes pelos anotadores. Segundo Wachsmuth et al. (2017b), os aspectos que constituem a dimensão retórica estão relacionados ao apelo emocional aplicado na argumentação, ambiguidade, imprecisão, estilo de linguagem e organização da estrutura do texto. Portanto, entende-se que essas características podem ser, em certa medida, identificadas por meio de recursos linguísticos superficiais.

A dimensão retórica, segundo Wachsmuth et al. (2017b), possui cinco aspectos:

		(a) Maj. Scores			(b) Agreement			(c) Pearson Correlation Coefficients											
	Quality Dimension	**1**	**2**	**3**	α	**full**	**maj.**	**Co**	**LA**	**LR**	**LS**	**Ef**	**Cr**	**Em**	**Cl**	**Ap**	**Ar**	**Re**	**GA**	**GR**	**GS**
Co	**Cogency**	150	131	23	.44	40.1%	91.8%		.64	.61	**.84**	**.81**	.46	.27	.41	.32	.55	.78	.64	**.71**	.70
LA	Local acceptability	84	169	51	.46	27.0%	90.8%	.64		.51	.53	.60	**.54**	.30	.40	.54	.46	.68	.75	.46	.45
LR	Local relevance	25	155	**124**	.47	32.6%	92.4%	.61	.51		.56	.56	.39	.27	.46	.35	.50	.62	.58	.68	.45
LS	Local sufficiency	172	119	13	.44	37.2%	92.8%	**.84**	.53	.56		.73	.39	.25	.37	.23	.51	.67	.51	.68	**.74**
Ef	**Effectiveness**	184	111	9	.45	42.1%	94.4%	.81	.60	.56	.73		.48	.31	.35	.34	.54	.75	.58	.66	.71
Cr	Credibility	99	199	6	.37	37.8%	95.7%	.46	.54	.39	.39	.48		**.37**	.32	.49	.37	.52	.52	.36	.40
Em	Emotional appeal	48	**235**	21	.26	42.8%	94.4%	.27	.30	.27	.25	.31	.37		.14	.30	.20	.30	.26	.26	.22
Cl	Clarity	42	191	71	.35	29.3%	89.8%	.41	.40	.46	.37	.35	.32	.14		.45	.56	.44	.45	.38	.27
Ap	Appropriateness	43	196	65	.36	17.4%	87.5%	.32	.54	.35	.23	.34	.49	.30	.45		.48	.47	.59	.20	.20
Ar	Arrangement	91	189	24	.39	26.6%	93.4%	.55	.46	.50	.51	.54	.37	.20	**.56**	.48		.55	.51	.49	.48
Re	**Reasonableness**	126	159	19	.50	41.4%	95.7%	.78	.68	.62	.67	.75	.52	.30	.44	.47	.55		**.78**	.65	.61
GA	Global acceptability	88	161	55	.44	31.6%	95.4%	.64	**.75**	.58	.51	.58	.52	.26	.45	**.59**	.51	.78		.46	.43
GR	Global relevance	69	167	68	.42	21.7%	90.1%	.71	.46	**.68**	.68	.66	.36	.26	.38	.20	.49	.65	.46		.61
GS	Global sufficiency	**231**	72	1	.27	**44.7%**	**98.0%**	.70	.45	.45	.74	.71	.40	.22	.27	.20	.48	.61	.43	.61	
Ov	**Overall quality**	152	128	24	**.51**	44.1%	94.4%	**.84**	.66	.61	.74	**.81**	.52	.30	.45	.42	**.59**	**.86**	.71	.70	.68

Figura 6: Resultados para os 304 textos do corpus classificados como argumentativos por todos os anotadores (Wachsmuth et al., 2017b).

1. **Credibilidade** (Cr): Credibilidade refere-se a como o autor transmite seus argumentos e os torna críveis. Segundo Wachsmuth et al. (2017b), um estilo apropriado em termos de escolha de palavras suporta a credibilidade. Além disso, de acordo com esses autores, aspectos que podem ser considerados para avaliar a credibilidade são: a honestidade do autor da mensagem, a polidez da linguagem utilizada ou o conhecimento e experiência do autor sobre os assuntos discutidos.
2. **Apelo emocional** (Em): O apelo emocional é considerado bem-sucedido em um argumento se ele cria emoções de tal forma que torna o público-alvo mais receptivo aos argumentos do autor.
3. **Clareza** (Cl): Clareza refere-se ao uso de uma linguagem gramaticalmente correta e amplamente inequívoca e que evita complexidade desnecessária e desvio do assunto discutido. A linguagem utilizada deve facilitar a compreensão e não deixar dúvidas sobre a posição do autor e a forma como ele a defende.
4. **Adequação** (Ap): A adequação de um argumento refere-se à linguagem (forma e conteúdo) utilizada para apoiar a criação de credibilidade e emoções, bem como a adequação ao assunto discutido.
5. **Organização** (Ar): Uma argumentação é considerada adequadamente organizada se apresentar a pergunta, os argumentos e a conclusão na ordem correta.

É importante destacar que o corpus utilizado no estudo de Wachsmuth et al. (2017b) é composto por mensagens que têm características diferentes do corpus adotado nesta pesquisa. Primeiramente, as mensagens de fóruns de discussão se caracterizam por serem mais longas do que as mensagens do Twitter, que têm um limite de 280 caracteres. O tamanho limitado das mensagens do Twitter pode impactar aspectos como Clareza e Organização. Outra diferença está relacionada ao contexto no qual as mensagens foram produzidas. No corpus de Wachsmuth et al. (2017b), as mensagens eram sobre tópicos gerais, às vezes controversos, mas sem a polarização política existente atualmente no Brasil, a qual pode impactar aspectos como Credibilidade, Apelo emocional e Adequação. Assim, embora a mesma taxonomia de Wachsmuth et al. (2017b) tenha sido adotada neste trabalho para embasar as medidas de qualidade da argumentação, o gênero textual e o domínio do corpus utilizado nesta pesquisa podem levar a resultados diferentes daqueles encontrados em (Wachsmuth et al., 2017b).

3. Corpus e anotação

Embora existam corpora abrangendo vários aspectos da análise argumentativa, alguns deles descritos na seção 2, até onde se sabe o corpus desenvolvido no projeto Arg Q!,[15] e descrito em (Silva et al., 2021), é o primeiro construído especificamente para a análise da qualidade da argumentação no domínio da política brasileira. Portanto, neste trabalho, foi utilizado o corpus construído e anotado conforme descrito por Silva et al. (2021). Este corpus é composto por *tweets* coletados como respostas a mensagens de parlamentares brasileiros postadas de 6 de março a 6 de abril de 2021.[16]

[15] Disponível em: `https://argq.org/`

[16] Embora os *tweets* dos parlamentares tenham sido considerados como semente para recuperar as respostas dos

Após a coleta dos *tweets*, cerca de 400 deles foram anotados com base nas pistas linguísticas descritas na Seção 3.1, conforme detalhado na Seção 3.2. Ressalta-se que apenas dados públicos foram coletados do Twitter e, embora os usuários não sejam identificados nos *tweets* do corpus, não podemos garantir que não seja possível rastrear a identidade do autor da mensagem.

3.1. Pistas linguísticas

Para a anotação do corpus foram utilizadas 30 pistas linguísticas definidas em (Silva et al., 2021): 4 para o aspecto de Clareza, 7 para o aspecto de Organização, 6 para o aspecto Credibilidade e 13 para o aspecto de Apelo emocional (6 para polaridade e 7 para intensidade). A Adequação não foi considerada em Silva et al. (2021) uma vez que se mostrou não relevante para a qualidade da argumentação em *tweets*, e também porque os anotadores apontaram que os critérios referentes à Adequação já eram contemplados pelos outros quatro aspectos. Portanto, o aspecto Adequação também foi desconsiderado neste trabalho. Nas seções seguintes, citamos as pistas linguísticas definidas para cada aspecto.

3.1.1. Clareza

Wachsmuth et al. (2017b) consideram um argumento claro se ele usa uma linguagem gramaticalmente correta e amplamente inequívoca e evita complexidade e desvios desnecessários da questão discutida. Além disso, a linguagem utilizada deve facilitar a compreensão e não deixar dúvidas sobre a posição do autor e a forma como ele a defende.

Nessa perspectiva, Silva et al. (2021) assumiram que todo argumento escrito em português tem o potencial de ser naturalmente claro, a menos que haja certos fenômenos linguísticos que interfiram negativamente na clareza. Dessa forma, a pontuação para o aspecto Clareza diminui com a presença de uma ou mais pistas linguísticas que prejudicam a clareza da argumentação, a saber: (i) questão que leva à dúvida, (ii) linguagem complexa desnecessária, (iii) presença de erros de língua portuguesa e (iv) desvio desnecessário do assunto principal. O aspecto Clareza foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (1).

$$\text{Clareza} = \begin{cases} \text{Baixa}, & \text{se } |Pistas| \geq 3 \\ \text{Média}, & \text{se } |Pistas| = 2 \\ \text{Alta}, & \text{caso contrário} \end{cases} \quad (1)$$

3.1.2. Organização

A definição de organização proposta por Wachsmuth et al. (2017b) considera que um texto deve ser composto de três partes, na seguinte ordem: (i) introdução do assunto, (ii) alguns argumentos para sustentar a conclusão e (iii) a conclusão. Essa definição, no entanto, não se aplica quando analisam-se textos do Twitter, pois os usuários não têm espaço suficiente para desenvolver essas 3 partes devido à limitação de caracteres máximos que uma mensagem pode conter (que é de 280). Assim, Silva et al. (2021) redefiniram a noção de organização para este gênero textual considerando a presença de pistas linguísticas que pontuam positivamente para a organização de um argumento.

As sete pistas linguísticas utilizadas para avaliar o aspecto Organização estão relacionadas à presença de relações linguísticas bem conhecidas: (i) relação condicional; (ii) relação de concessão; (iii) oposição ou contraste; (iv) comparação entre duas ideias; (v) relação de causa e efeito, explicação ou propósito; (vi) encadeamento cronológico ou enumeração; e (vii) exemplificação ou interligação lógica.

Diferentemente do aspecto da Clareza, que se ancora no pressuposto de que todo argumento no Twitter tem o potencial de ser inerentemente claro, no aspecto Organização, considera-se que esse gênero textual tem o potencial intrínseco de ser desorganizado, a menos que ocorram determinados fenômenos linguísticos que contribuam positivamente para a organização. Dessa forma, o aspecto Organização foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (2).

$$\text{Organização} = \begin{cases} \text{Alta}, & \text{se } |Pistas| \geq 2 \\ \text{Média}, & \text{se } |Pistas| = 1 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (2)$$

3.1.3. Credibilidade

De acordo com Wachsmuth et al. (2017b), um argumento deve ser avaliado como bem-sucedido em criar credibilidade se transmitir argumentos e outras informações de uma forma que torne o

seguidores, vale ressaltar que a avaliação da qualidade da argumentação foi feita apenas nas respostas dos seguidores.

autor crível, por exemplo, indicando a honestidade do autor, a polidez da linguagem utilizada ou revelando o conhecimento do autor ou a experiência em relação aos assuntos abordados.

Para a avaliação do aspecto Credibilidade, Silva et al. (2021) consideraram que um argumento escrito em português é verossímil se algumas pistas linguísticas estiverem presentes na superfície textual; ou seja, não levaram em consideração nenhum critério ou dado externo como a aceitação ou engajamento do autor nas redes sociais. Vale ressaltar que o Twitter é uma plataforma aberta e, portanto, qualquer usuário cadastrado pode postar mensagens diversas nesta plataforma de mídia social. Como não há pré-análise do perfil do usuário ou do conteúdo postado por ele, a dúvida sobre a credibilidade do que é postado é inerente à plataforma.

Assim, a pontuação do aspecto Credibilidade aumenta se as seguintes pistas estiverem presentes: (i) menção a uma data específica; (ii) menção a um fato midiático, histórico ou enciclopédico; (iii) menção a uma autoridade pública; (iv) presença de uma *hashtag* que reforça uma posição; (v) presença de um termo especializado; e (vi) um relato de alguma experiência pessoal ou individual. O aspecto Credibilidade foi classificado com base na quantidade (cardinalidade) de pistas identificadas (Pistas), como apresentado na equação (3).

$$\text{Credibilidade} = \begin{cases} \text{Alta}, & \text{se } |Pistas| \geq 3 \\ \text{Média}, & \text{se } |Pistas| = 2 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (3)$$

3.1.4. Apelo emocional

Segundo Wachsmuth et al. (2017b), um argumento tem apelo emocional quando cria emoções no interlocutor. Silva et al. (2021) decidiram dividir as pistas linguísticas para o aspecto apelo emocional em polaridade e intensidade.

Polaridade A polaridade de um argumento é considerada positiva ou negativa se contribui para criar emoções boas ou ruins no leitor, respectivamente. Se o seu conteúdo não causar nenhuma emoção ou se o apelo emocional estiver bem equilibrado entre positivo e negativo, então a polaridade do argumento é considerada neutra.

Assim, as pistas linguísticas de Silva et al. (2021) para uma polaridade positiva são: (i) a presença de uma referência cordial a uma pessoa/organização e (ii) o uso de linguagem polida. Por outro lado, as pistas linguísticas para uma polaridade negativa são: (i) a presença de uma referência pejorativa a uma pessoa/organização; (ii) o uso de xingamento ou palavra de baixo calão; (iii) a presença de discurso de ódio ou ameaça; e (iv) o uso de expressões que denotam especulação. Não há nenhuma pista linguística específica para uma polaridade neutra, pois é o meio termo entre os dois extremos. A Polaridade do aspecto Apelo emocional foi classificada com base na diferença entre a quantidade de pistas positivas (Pistas$^+$) e negativas (Pistas$^-$), como apresentado na equação (4).

$$\text{AE}_{pol} = \begin{cases} \text{Negativa}, & \text{se } \left|Pistas^+\right| - \left|Pistas^-\right| < 0 \\ \text{Positiva}, & \text{se } \left|Pistas^+\right| - \left|Pistas^-\right| > 0 \\ \text{Neutra}, & \text{caso contrário} \end{cases} \quad (4)$$

Intensidade O apelo emocional também é avaliado de acordo com sua intensidade, que é determinada pela quantidade de pistas linguísticas que potencializam a criação de emoções. Quanto maior o número de pistas linguísticas, maior a intensidade do apelo emocional. A intensidade é avaliada de acordo com a presença de (i) um pronome ou verbo na primeira pessoa; (ii) a repetição de sinais de pontuação; (iii) estrutura enfática (por exemplo, palavra em maiúscula); (iv) uma frase imperativa ou palavra de ordem; (v) uma expressão que denota exagero (por exemplo, "sempre", "nunca", "todo mundo"); (vi) linguagem não verbal (por exemplo, *emoticons*); e (vii) a presença de uma expressão idiomática, provérbio ou metáfora. A Intensidade do aspecto Apelo emocional foi classificada com base na quantidade de pistas de intensidade identificadas (Pistas), mas também em relação à quantidade de pistas positivas (Pistas$^+$) e negativas (Pistas$^-$) de polaridade identificadas, como apresentado na equação (5).

$$\text{AE}_{int} = \begin{cases} \text{Alta}, & \text{se } \left|Pistas^-\right| \geq 3 \\ & \text{ou } \left|Pistas^+\right| \geq 2 \\ & \text{ou } |Pistas| \geq 4 \\ \text{Média}, & \text{se } \left|Pistas^-\right| = 2 \\ & \text{ou } \left|Pistas^+\right| = 1 \\ & \text{ou } 2 \leq |Pistas| \leq 3 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (5)$$

3.1.5. Qualidade geral da argumentação

Após a anotação das pistas linguísticas associadas a cada aspecto, os anotadores em (Silva et al., 2021) definiram como eles seriam combinados para resultar em um valor para a Qualidade Geral da argumentação. Para tanto, o mapeamento de categorias para valores foi realizado como apresentado na tabela 1.

f		g	
Categoria	**Valor**	**Categoria**	**Valor**
Baixa	1	Negativa	-1
Média	2	Neutra	0
Alta	3	Positiva	1

Tabela 1: Funções f e g de mapeamento das categorias atribuídas pelos anotadores para valores.

A nota do Apelo emocional (N_{AE}) foi definida tal como se apresenta algoritmicamente de seguida.

if AE_{pol} =Neutra **then**
  $N_{AE} \leftarrow \frac{f(AE_{int})}{2}$
else
  $N_{AE} \leftarrow g(AE_{pol}) \times f(AE_{int})$
end if

E as notas dos demais aspectos foram definidas com base na aplicação da função f:

- $N_{Cla} = f(\text{Clareza})$
- $N_{Org} = f(\text{Organização})$
- $N_{Cred} = f(\text{Credibilidade})$

A partir das notas associadas a cada aspecto, a nota final foi dada pela soma das notas dos aspectos. Por fim, definiu-se que a nota para a qualidade geral da argumentação (N_{QG}) seria a média das notas dos anotadores e a categoria associada à Qualidade Geral foi definida como apresentado na equação (6).

$$QG = \begin{cases} \text{Alta}, & \text{se } N_{QG} \geq 7 \\ \text{Média}, & \text{se } 5 \leq N_{QG} < 7 \\ \text{Baixa}, & \text{caso contrário} \end{cases} \quad (6)$$

3.2. Anotação do corpus

Com base nas pistas linguísticas apresentadas anteriormente e seguindo as diretrizes estabelecidas em (Silva et al., 2021),[17] 400 *tweets* foram analisados por quatro pesquisadores em linguística computacional (3 com formação em linguística e 1 em computação). Desses, 48 *tweets* foram descartados por não terem sido considerados argumentativos[18] por todos os quatro anotadores. Assim, o corpus final é composto por 352 *tweets*.

Na Figura 7 é exibido um gráfico de distribuição de pontuação para cada aspecto dos 352 *tweets* argumentativos. A maioria deles tem alta Clareza e Organização, mas baixa Credibilidade. De fato, apenas 3% deles foram avaliados como de alta credibilidade. Com relação ao apelo emocional, 54% dos *tweets* foram avaliados como de polaridade negativa e 75% de média intensidade.

Na Tabela 2 (a) são exibidas as pontuações finais de cada aspecto para os 352 *tweets*. O grau de concordância entre anotadores foi calculado e é exibido na Tabela 2 (b) e expresso como o intervalo de α de Krippendorff (2011) (valor mais baixo - valor mais alto) que apresentaram os trios menos concordantes e mais concordantes de anotadores[19], e tanto os acordos totais quanto os majoritários.

A "Concordância total" é alcançada quando todos os anotadores concordam com a mesma pontuação, e "maioria" indica que pelo menos três anotadores concordaram. A concordância total encontrada ficou entre 27,84% e 57,67%, e a concordância majoritária dos anotadores ficou entre 69,89% e 86,93%. Com exceção do aspecto Clareza, todos os aspectos tiveram valores máximos de concordância acima de 0,40, diferentemente de Wachsmuth et al. (2017b) (veja Figura 6) onde os valores de concordância α para todos os aspectos ficaram abaixo de 0,40.

4. Experimentos

Para responder as questões de pesquisa definidas para este trabalho, foram realizados experimentos utilizando algoritmos de AM baseados em *features* e, também, uma abordagem neural baseada em transformers (BERT e RoBERTa).

[17] As diretrizes utilizadas para a anotação estão disponíveis no site do projeto Arg Q!: https://argq.org/

[18] Tradicionalmente, um texto é considerado argumentativo se contém argumentos organizados e estruturados em uma sequência lógica. No entanto, no que diz respeito ao gênero textual e domínio deste corpus, este conceito foi adaptado para abranger quaisquer *tweets* em que a posição/opinião do autor pudesse ser determinada.

[19] Decidimos relatar a concordância alcançada entre os trios para que nossos resultados pudessem ser comparados com os de Wachsmuth et al. (2017b) já que em seu trabalho havia três anotadores.

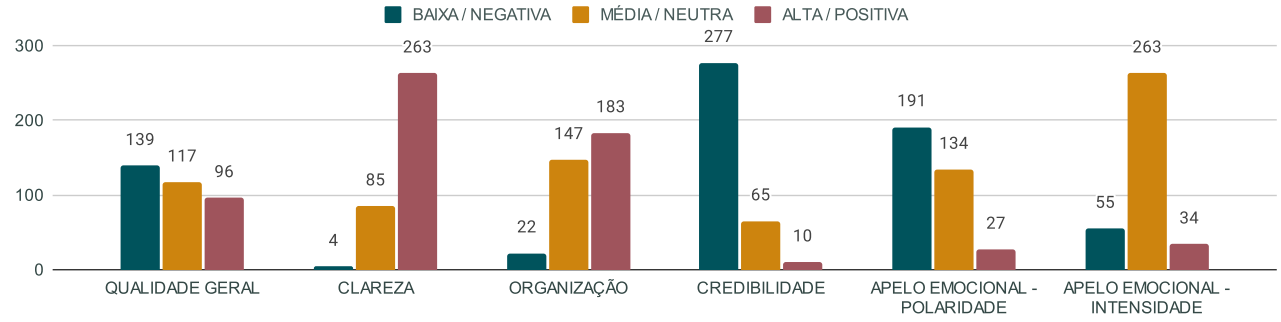

Figura 7: Distribuições de pontuação por aspectos de qualidade da argumentação no corpus de Silva et al. (2021) (tradução nossa).

Aspecto da qualidade	(a) Pontuação final			(b) Concordância		
	Baixo/ Negativo	Médio/ Neutro	Alto/ Positivo	α trios	total (4/4)	maioria (3-4/4)
Clareza	4	85	**263**	0,26 - 0,30	48,58%	79,26%
Organização	22	**147**	183	0,51 - 0,71	50,57%	82,67%
Credibilidade	**277**	65	10	0,36 - 0,48	57,67%	86,93%
Apelo emocional - Polaridade	**191**	134	27	0,60 - 0,66	51,99%	82,67%
Apelo emocional - Intensidade	55	**263**	34	0,48 - 0,55	40,63%	82,39%
Qualidade geral	139	117	96	0,50 - 0,54	27,84%	69,89%

Tabela 2: Pontuações para cada aspecto e a concordância entre os juízes humanos do corpus de Silva et al. (2021) (tradução nossa).

Para a abordagem neural, foram utilizadas apenas os *tweets* presentes no corpus, como descrito na seção 4.2. Entretanto, para a avaliação com os algoritmos de AM tradicionais, foi necessária a construção de um conjunto de *features* linguísticas para a tarefa de classificação, como detalhado na seção 4.1.

4.1. Experimentos com AM baseado em *features*

Para realizar os experimentos com AM tradicional foi necessário definir as *features* linguístico-computacionais (como descrito na seção 4.1.1) e treinar os modelos computacionais como descrito na seção 4.1.2.

4.1.1. Definição das features computacionais

Com o objetivo de encontrar as *features* linguístico-computacionais que melhor se correlacionam com as pistas linguísticas utilizadas para medir a qualidade da argumentação, foi definido e gerado um conjunto de 337 *features*.

Essas *features* foram categorizadas em 14 grupos, a maioria delas geradas pelo NILC-Metrix[20] (Leal et al., 2022):

1. **Medidas Psicolinguísticas** extraem características subjetivas do texto, tais como: imageabilidade, concretude, familiaridade e idade de aquisição. 44 *features* deste tipo foram geradas pelo NILC-Metrix.

   Imageabilidade envolve a facilidade e rapidez de evocar uma imagem mental associada a uma palavra;

   Concretude diz respeito ao grau em que uma palavra se refere a objetos, pessoas, lugares ou coisas que podem ser percebidas pelos sentidos;

   Familiaridade é o grau em que as pessoas conhecem e usam palavras em suas vidas cotidianas;

   Idade de Aquisição é uma estimativa da idade que a pessoa tinha pessoa quando uma palavra foi aprendida.

[20] NILC-Metrix (Leal et al., 2022) composto por 200 métricas linguísticas e psicolinguísticas utilizadas na avaliação de métodos de predição de complexidade textual. Disponível em: `http://fw.nilc.icmc.usp.br:23380/nilcmetrix`

2. **Informação morfossintática** engloba *features* relacionadas à classe gramatical de uma palavra e a sua função sintática no texto, como a proporção de adjetivos, advérbios, pronomes, substantivos, verbos e preposições, em relação à quantidade total de palavras no texto. Neste trabalho, para extrair essas *features* foram usados Apertium[21] (Armentano-Oller et al., 2006), NLPNet[22] (Fonseca & Rosa, 2013) e NILCMetrix. Foram geradas 94 *features* deste tipo.
3. **Complexidade Sintática** está relacionada à dificuldade de processamento de alguns tipos de estruturas de frases sofisticadas, por exemplo, a proporção de orações subordinadas reduzidas pela quantidade de orações do texto ou a proporção de orações na voz passiva analítica em relação à quantidade de orações do texto. 37 *features* deste tipo foram geradas pelo NILC-Metrix.
4. **Frequências de Palavras** mostram os valores das frequências absolutas e relativas das palavras no texto. 20 *features* deste tipo foram geradas pelo NILC-Metrix.
5. **Conectivos** incluem métricas relacionadas à quantidade de conectivos, operadores lógicos ou palavras que denotam negação em relação às palavras do texto. 18 *features* deste tipo foram geradas pelo NILC-Metrix.
6. **Simplicidade Textual** fornece métricas que medem o nível de complexidade em um texto com relação à dificuldade de compreensão de leitura. 9 *features* deste tipo foram geradas pelo NILC-Metrix.
7. **Índices de Legibilidade** incluem métricas que medem a legibilidade de um texto, como o índice Brunet e Flesch. 10 *features* deste tipo foram geradas pelo NILC-Metrix.
8. **Indicadores Temporais de um Léxico** incluem índices relacionados à diversidade de tempos verbais que ocorrem no texto. 14 *features* deste tipo foram geradas pelo NILC-Metrix.
9. **Informações Semânticas** referem-se a várias métricas que fornecem informações sobre o significado das palavras no texto, como a quantidade média de hiperônimos por verbo nas sentenças, ou a proporção de substantivos abstratos em relação à quantidade de palavras do texto. 18 *features* deste tipo foram geradas pelo NILC-Metrix.
10. **Medidas Descritivas** referem-se a métricas como a quantidade de parágrafos, frases e palavras em um texto e de sílabas por palavra. 16 *features* deste tipo foram geradas pelo NILC-Metrix.
11. **Coesão Semântica** é expressa por métricas que calculam as relações semânticas entre palavras em um texto, por exemplo a média de similaridade entre os pares de sentenças no texto, ou a média da entropia cruzadas das sentenças do texto. 19 *features* deste tipo foram geradas pelo NILC-Metrix.
12. **Polaridade do sentimento** mede a frequência de palavras com emoções positivas, negativas e neutras no texto. Neste trabalho, realizamos análise de polaridade com base em um modelo treinado usando o algoritmo de Floresta Aleatória e o corpus TweetSentBR[23] (Brum & Nunes, 2018) e também um modelo baseado em BERT de (Capellaro & Caseli, 2021). Foram geradas 4 *features* deste tipo.
13. **Linguagem Tóxica** indica se há ou não linguagem tóxica no texto. 7 *features* desse tipo foram geradas com um modelo baseado em BERT treinado no corpus ToLD-Br[24] (Leite et al., 2020), todas elas *features* binárias: não tóxico, LGBTQ+fobia, obsceno, insultuoso, racismo, misoginia e xenofobia.
14. A frequência de uso para **diferentes categorias de palavras** gerado com base na versão em português do LIWC[25] (Balage Filho et al., 2013). Foram geradas 62 *features* deste tipo.

4.1.2. Treinamento dos modelos de AM baseado em features

Este processo consiste em três etapas principais: **(i)** o pré-processamento, criação do conjunto de *features* e seleção das melhores *features*; **(ii)** ajuste de hiperparâmetros com validação cruzada aninhada; e **(iii)** melhor seleção de modelo e geração de métricas. Esta configuração experimental é ilustrada na Figura 8.

Para realizar o **primeiro passo** (i), cada *tweet* foi pré-processado pelo Enelvo[26] para remover informações desnecessárias (como o identificador do usuário do Twitter ao qual a men-

[21] Disponível em: https://www.apertium.org/
[22] Disponível em: http://nilc.icmc.usp.br/nlpnet/
[23] Disponível em: https://bitbucket.org/HBrum/tweetsentbr/src/master/
[24] Disponível em: https://github.com/JAugusto97/ToLD-Br
[25] Disponível em: http://143.107.183.175:21380/portlex/index.php/en/liwc
[26] Disponível em: https://github.com/thalesbertaglia/enelvo

sagem se referia como resposta) e normalizar palavras ruidosas em conteúdo gerado pelo usuário (por exemplo, substituir abreviações como "vc" por sua versão utilizada na norma culta "você"). Em seguida, os *tweets* normalizados foram processados pelas ferramentas linguístico-computacionais a fim de extrair as *features* descritas na seção 4.1.1.

Após a geração das 337 *features*, foi aplicado um método de seleção de *features* para determinar quais delas se correlacionam melhor com a qualidade geral da argumentação. O objetivo da seleção de *features* é encontrar aquelas que possivelmente são as melhores preditoras para tarefas de AM. Esse procedimento é importante, pois algumas *features* podem ser irrelevantes para a tarefa de AM e, desse modo, adicionar ruído ao modelo treinado. Adi et al. (2019) apontam que entre os problemas causados por conjuntos de dados de alta dimensão estão o alto custo computacional e a baixa precisão do modelo gerado. Para contornar esses problemas, sugere-se selecionar apenas as *features* mais relevantes que tenham uma alta correlação com a classe.

Alguns métodos de seleção de *features* podem ser aplicados para reduzir a alta dimensionalidade do conjunto de *features*, como Eliminação Recursiva de *Features* com Validação Cruzada (RFECV) (Misra & Yadav, 2020), Ganho de Informação (IG), Taxa de Ganho de Informação (Taxa de Ganho) (Adi et al., 2019) e Análise de Componentes Principais (PCA) (Maćkiewicz & Ratajczak, 1993). Neste trabalho, o método escolhido para a fase de seleção de *features* foi o RFECV dado que a literatura aponta a efetividade dele na eliminação de *features* irrelevantes por se tratar de um método que busca evitar o problema de *overfiting* com Eliminação Recursiva de *Features* (RFE), aplicando a validação cruzada estratificada (Misra & Yadav, 2020). O RFECV classifica as *features* com eliminação recursiva de *features* e validação cruzada de 10 vezes e, assim, seleciona o número ideal de *features* para construção de modelo.

Assim, as melhores *features* foram selecionadas com o auxílio da biblioteca scikit-learn usando RFE com Classificador de Floresta Aleatória e validação cruzada de 10 vezes. O conjunto final de *features* selecionadas para predizer a qualidade geral da argumentação contém 290 *features*. As 10 *features* mais significativas para a qualidade geral, de acordo com essa seleção automática, são apresentadas na Tabela 3. Contudo, como alguns dos algoritmos investigados nesta pesquisa são conhecidos por lidarem bem com um número elevado de *features*, como o SVM, também foram realizados experimentos sem esta etapa de seleção de *features*.

A **segunda etapa** (ii) consiste em refinar os parâmetros dos algoritmos de AM. Para isso, foi utilizado o GridSearchCV,[27] uma pesquisa exaustiva sobre valores de parâmetros especificados para um estimador. Conforme mostrado na Figura 8, as dobras de treinamento do laço externo são usadas no laço interno para ajustar os hiperparâmetros. O laço interno seleciona a melhor configuração de hiperparâmetros.

Cinco dobras estratificadas foram utilizadas em ambas os laços. O laço interno faz uma pesquisa de grade no espaço de hiperparâmetros que é validado de forma cruzada em relação aos conjuntos de treinamento e validação adquiridos pelo laço externo. A configuração do hiperparâmetro que maximiza a pontuação de precisão é retornada para cada pesquisa de grade. A generalização da configuração do modelo selecionado é então validada usando as métricas padrão de acurácia, precisão, cobertura e medida F nos conjuntos de teste criados pelo laço externo.

A **terceira etapa** (iii) consiste em gerar os melhores modelos de AM para cada algoritmo testado e classificar os *tweets* de teste usando os modelos treinados. Testamos seis modelos classificadores em nosso conjunto de dados: Regressão Logística (*Logistic Regression*, LR), *K-Nearest Neighbors* (KNN), Árvores de Decisão (*Decision Tree*, DT), Máquinas de Vetores de Suporte (*Support Vector Machines*, SVM), Floresta Aleatória (*Random Forest*, RF) e *Naive Bayes* (NB).[28]

4.2. Experimentos com BERTimbau e RoBERTaTwitterBR

Além dos experimentos descritos utilizando os algoritmos de AM baseado em *features*, foram realizados experimentos com o BERTimbau[29] (Souza et al., 2020) e o RobertaTwitterBR,[30], modelos

[27] Disponível em:https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html

[28] Os hiperparâmetros e respectivos valores explorados foram os seguintes: LR: *penalty*=l2, C=*np.power*(10., *np.arange*(-4, 4)), KNN: *n neighbors*=*list*(*range*(1, 10)), p=[1, 2], DT: *max depth*=*list*(*range*(1, 10), *criterion*=[*gini*, *entropy*], SVM: *kernel*=*rbf*, C=*np.power*(10., *np.arange*(-4, 4)), *gamma*=*np.power*(10., *np.arange*(-5, 0)); *kernel*=*linear*, C=*np.power*(10., *np.arange*(-4, 4)), RF: *n estimators*=[10, 100, 500, 1000, 10000], *criterion*=[*gini,entropy*], *max depth*=[10,11,12,13,14], NB: *var smoothing*= *np.logspace*(0,-9, num=100).

[29] Disponível em: https://github.com/neuralmind-ai/portuguese-bert

[30] https://huggingface.co/verissimomanoel/RobertaTwitterBR

ID	Grupo	Métrica	Descrição
1	Linguagem tóxica	toxiclang	Identifica a presença de linguagem tóxica no texto
2	NILCMetrix-Medidas Psicolinguísticas	idade_de_aquisição	Valores de idade média de aquisição de palavras de conteúdo de texto
3	NILCMetrix-Frequência de palavras	freq_bra	Média dos valores de frequência das palavras no texto na escala logarítmica Zipf via Corpus Brasileiro
4	NILCMetrix-Medidas Psicolinguísticas	imageabilidade_25_4_ratio	Proporção de palavras com valor de imageabilidade entre 2,5 e 4 em relação a todas as palavras de conteúdo do texto
5	NILCMetrix-Medidas psicolinguísticas	imageabilidade_mean	Imageabilidade média das palavras de conteúdo no texto
6	NILCMetrix - Medidas Descritivas	syllabes_per_content_word	Número médio de sílabas por palavra de conteúdo no texto
7	TweetSentBR- Sentimento neutro	sent_neu	Proporção de palavras com emoção neutra em relação a todas as palavras do texto
8	NILCMetrix-Frequência de palavras	freq_brwac	Média dos valores das frequências das palavras do texto na escala logarítmica Zipf via BrWac
9	NILCMetrix-Medidas Psicolinguísticas	imageabilidade_4_55_ratio	Proporção de palavras com valor de imageabilidade entre 4 e 5,5 em relação a todas as palavras de conteúdo do texto
10	NILCMetrix-Medidas descritivas	sentence_length_standard_deviation	Desvio padrão do número de palavras por frase

Tabela 3: Top-10 *features* com melhor correlação com a qualidade geral da argumentação.

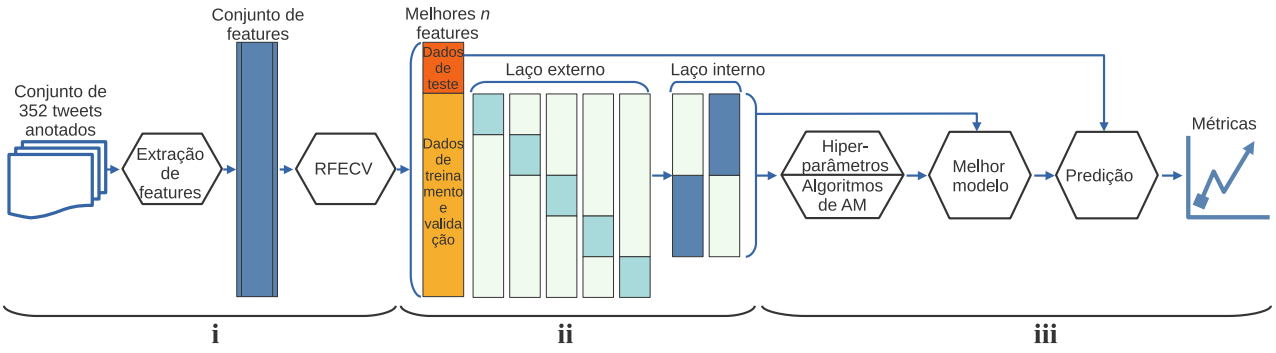

Figura 8: Configuração experimental adotada para a geração de modelos de AM baseados em *features*, dividida em: i. extração e seleção de *features*, ii. otimização com validação cruzada aninhada; e iii. treinamento e teste.

neurais treinados em português. O BERTimbau foi treinado no BrWaC (*Brazilian Web as Corpus*), um grande corpus em português, por 1 milhão de passos, usando *full word mask*. O RobertaTwitterBR foi treinado com aproximadamente 7 milhões de *tweets* em português.

Nestes experimentos, somente o texto das mensagens foi utilizado, ao contrário dos experimentos com os algoritmos de AM tradicional que utilizaram as *features* linguísticas. O único processamento realizado com as mensagens para estes experimentos foi a remoção dos nomes dos usuários. Um script Python foi usado para realizar essa tarefa.

A Figura 9 ilustra esse processo que foi realizado em duas etapas: (i) ajuste fino do modelo neural; e (ii) avaliação do modelo no conjunto de teste e geração das métricas. Vale ressaltar que as mesmas partições de treinamento e teste foram usadas tanto nos experimentos com AM baseado em *features* quanto nos experimentos com os modelos neurais, de forma estratificada.

5. Resultados

Neste artigo, investigou-se como o AM pode ser aplicado para classificar a qualidade da argumentação em *tweets* do domínio da política brasileira. Para isso, foram testados diferentes classi-

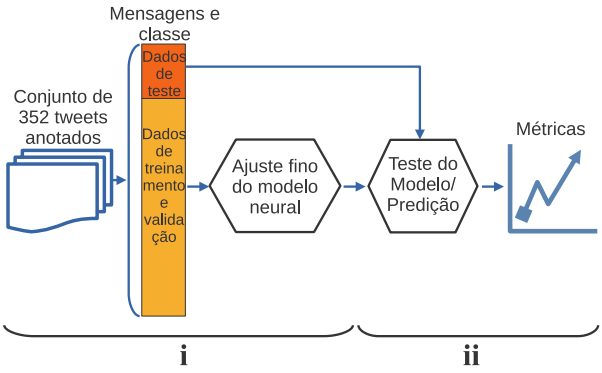

Figura 9: A configuração experimental proposta foi dividida em duas etapas: i. ajuste fino dos modelos neurais e ii. avaliação do modelo.

ficadores com diferentes hiperparâmetros em um corpus de 352 *tweets* do domínio da política brasileira. Desses, 281 *tweets* foram usados para treinamento ou ajuste fino dos modelos e os restantes 71 *tweets* foram usados para teste.

As seções a seguir apresentam as análises quantitativas e qualitativas dos resultados obtidos nos experimentos.

5.1. Análise quantitativa

Em termos das medidas quantitativas usualmente aplicadas na avaliação de modelos computacionais, apresentadas na Tabela 4, concluímos que o modelo neural obtido com o ajuste fino do BERTimbau obteve os melhores valores: 63,38% de acurácia, 69,65% de precisão, 63,38% de cobertura e uma medida F de 63,61%.[31]

Entre os algoritmos de AM baseado em *features*, o desempenho variou de 35% a 54% de medida F, ficando a pelo menos cerca de 9 pontos percentuais abaixo do melhor modelo neural. Quanto ao baixo desempenho desses modelos, vale mencionar que a alta quantidade de *features* usadas no treinamento pode ter influenciado esses resultados. Mesmo com a etapa de seleção de *features*, realizada como descrito na Figura 8, os modelos tradicionais foram treinados com 290 *features* para 281 instâncias e essa alta dimensionalidade pode ter confundido os algoritmos no momento de gerar os modelos. Contudo, a configuração experimental do AM baseado em *features* foi proposta para dar autonomia ao processo de AM na definição das *features* mais relevantes (sem a interferência do especialista humano), de

[31] Os melhores resultados foram alcançados com os seguintes hiperparâmetros otimizados: número de épocas= 20; *batch size*= 8; *early stop*= 2; *learning rate*=1e-5. Os mesmos hiperparâmetros foram utilizados para o BERTimbau e RoBERTaTwitterBR.

Classificador	Acurácia	Precisão	Cobertura	Medida F
LR	45,07%	45,34%	45,07%	44,56%
LR + RFECV	46,48%	46,53%	46,48%	45,75%
K-NN	47,89%	49,20%	47,89%	46,86%
K-NN + RFECV	39,44%	41,68%	39,44%	37,80%
DT	46,48%	32,57%	46,48%	37,71%
DT + RFECV	46,48%	32,57%	46,48%	37,71%
SVM	43,66%	42,99%	43,66%	41,10%
SVM + RFECV	35,21%	35,95%	35,21%	35,39%
RF	54,93%	54,90%	54,93%	54,44%
RF + RFECV	45,07%	46,53%	45,07%	43,24%
NB	45,07%	45,06%	45,07%	44,29%
NB + RFECV	50,70%	50,89%	50,70%	50,49%
RobertaTwitterBR	49,30%	48,41%	49,30%	48,11%
BERTimbau	63,38%	69,65%	63,38%	63,61%

Tabela 4: Valores das medidas de avaliação obtidas para a qualidade geral da argumentação nos modelos baseados em *features* (LR, K-NN, DT, RF e NB) com (+ RFECV) e sem a etapa de seleção de *features* e nos modelos neurais (RobertaTwitterBR e BERTimbau).

modo semelhante ao que fazem os modelos neurais durante o ajuste fino.

É importante destacar que os experimentos conduzidos com DT apresentaram resultados idênticos com e sem a etapa de seleção de *features*. Isso se deve às características do próprio algoritmo, que já faz a seleção de *features* durante o treinamento para definir qual *feature* vai em cada nó da árvore. O DT trabalha dividindo recursivamente o conjunto de *features* em subconjuntos homogêneos, até que um critério de parada seja atingido. Durante o processo de construção da árvore, o algoritmo avalia diferentes *features* e escolhe aquela que melhor separa as classes ou minimiza o erro.

Nota-se, também, que os experimentos conduzidos com SVM sem a etapa de seleção de *features* apresentaram melhores resultados se comparados com o experimento com a seleção de *features*. Isso também se deve às características do algoritmo, que trabalha transformando as *features* em um espaço de alta dimensão, no qual é mais provável que as classes sejam linearmente separáveis: quanto mais pontos, melhor o detalhamento na definição das bordas. Em seguida, o SVM encontra o hiperplano que separa as classes, minimizando o erro ou a perda (Vapnik, 1999).

A partir dos valores das medidas de avaliação apresentados nesta seção é possível concluir que os modelos de AM basedos em *features* não são os mais indicados para identificar automaticamente os *tweets* do campo da política brasileira que tenham bons argumentos. Embora os experimentos tenham sido realizados de maneira bastante abrangente em termos de recursos linguístico-computacionais usados para o português, *features*

e algoritmos, não é possível apontar claramente quais desses fatores levaram ao baixo desempenho.

Em relação aos resultados dos modelos neurais, a Tabela 5 e a Figura 10 apresentam os resultados detalhados do modelo de melhor desempenho obtido com o ajuste fino do BERTimbau. Esse modelo teve uma excelente precisão para a classe Alta. Das 19 instâncias da classe Alta presentes no corpus de teste, o modelo foi capaz de predizer 9 delas corretamente (resultando em uma cobertura de 47,37%) mas com uma precisão de 100%. Assim, embora a amostra do corpus de teste seja pequena, esses resultados trazem fortes indícios de uma excelente assertividade do modelo para prever *tweets* com alta qualidade da argumentação. Esses *tweets* são apresentados na Tabela 6.

	Precisão	**Cobertura**	**Medida F**
Baixa	64,52%	71,43%	67,80%
Média	51,61%	66,67%	58,18%
Alta	100,00%	47,37%	64,29%

Tabela 5: Resultados detalhados para a predição da qualidade geral da argumentação retornados pelo modelo obtido com o ajuste fino do BERTimbau.

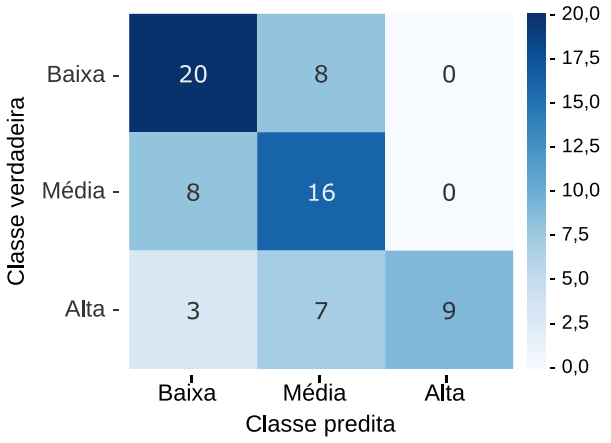

Figura 10: Matriz de confusão do modelo obtido com o ajuste fino do BERTimbau.

5.2. Análise qualitativa

Para complementar a análise quantitativa, esta seção traz uma análise qualitativa das classificações do modelo de melhor desempenho para a predição da qualidade da argumentação nos *tweets* do domínio da política brasileira.

5.2.1. Análise da classe de Alta qualidade de argumentação

Como mencionado anteriormente, o objetivo dos experimentos apresentados neste artigo foi gerar um modelo capaz de predizer, com um bom desempenho, aqueles *tweets* que tenham alta qualidade de argumentação. Assim, a Tabela 6 apresenta os 9 *tweets* do corpus de teste que o modelo ajustado a partir do BERTimbau apontou corretamente como sendo de alta qualidade de argumentação.

Vale mencionar que todos foram considerados pelos anotadores humanos como sendo claros, organizados e com nenhum ou poucos erros de português. Além dessas características gerais que levaram os anotadores a considerar tais *tweets* como de alta qualidade da argumentação, outras pistas linguísticas (destacadas em negrito) incluem: uso de tratamento cordial (ex: Desculpe, Caro, senhora, Sr.), fato histórico (ex: como foi feito na Alemanha), numérico (ex: 56 milhões de eleitores) ou citação de figura/entidade pública que dá respaldo a alguma informação (ex: Arthur Lira, conselho de ética). A presença de marcadores discursivos (ex: como, porque, ainda mais) também foi apontada como positiva para melhorar a qualidade argumentativa porque eles unem, contrapõem ou interligam ideias.

Como exemplo, na mensagem 6.1, da Tabela 6, dentre as principais pistas linguísticas que fizeram com que os humanos considerassem esse *tweet* como de alta qualidade de argumentação, destacam-se a presença de: (i) termo especializado "Impeachment" (apontado por todos os anotadores); (ii) encadeamento cronológico expresso pela presença do "quando", usado com sentido temporal (apontado por todos os anotadores), e (iii) fato midiático (pista anotada pela maioria dos anotadores).

Por outro lado, outros 10 *tweets*, que correspondem às mensagens 7.1 até 7.10, da Tabela 7, considerados de alta qualidade argumentativa pelos anotadores, foram classificados pelo modelo BERTimbau de modo diferente: 70% deles como Média e 30% deles como Baixa.

Comparando os *tweets* da Tabela 7 com os da Tabela 6 podemos notar uma maior presença de abreviações (ex: n, q, ñ, tds) e grafias diferentes para palavras (ex: "K0V1D" para representar Covid, ou "vac1na" para vacina), além de uma ausência notável de pontuação (como o uso da vírgula). Essas características retratam bem as peculiaridades do gênero textual e aparentemente tiveram bastante peso na classificação do modelo BERTimbau.

Ex.	Texto
6.1	Rodrigo Maia, você hoje já falou que se arrepende do apoio a Bolsonaro no segundo turno. Parabéns por admitir isto. Agora... **quando** virá o arrependimento de não ter ao menos colocado para a frente algum dos pedidos de **Impeachment**?
6.2	Vc propôs essa emenda, esperando que passe ou apenas para constar? Com a **postagem do** seu presidente da câmara, que até já considerou que o Dep. Daniel Silveira contrapôs à democracia, **mesmo não tendo** sido julgado e condenado pelo STF, espera que essa sua proposta tenha sucesso? https://t.co/uJjvgcwqEt
6.3	**Desculpe senhora** deputada, **cansei** de vcs ! Ninguém faz nada, ninguém! Vcs brincam com o povo! **Se** hoje um governador maluco fizer um forno, **como foi feito na Alemanha** e começar a matar as pessoas,tudo bem , os caras que jamais devem ser citados, deram o direito !
6.4	**Caro** Deputado, não sei se irá ler **meu** posicionamento. **Mas**, **calaram a voz** de uma Deputado q foi eleito para **PODER FALAR POR NÓS!** Um PODER, calou a não a voz do Daniel, calou foi a NOSSA! Ontem foi deputado pondo mordaça da boca de outro deputado e traçando o fim do CONGRESSO.
6.5	Está na hora de exigir o respeito com seriedade, **impeachment** se faz mais que necessário, ele está tentando rebaixar a **Câmara dos Deputados** a seu serviço, uma ação judicial enérgica imediata. Ação do **Arthur Lira** agora, **se** deixar passar perderá a força 💪 👊 🇧🇷 🇧🇷 🇧🇷 🇧🇷
6.6	**Ao Sr**. Apresentar esse material ao **conselho de ética**, vamos ver **se** esse conselho de ética e justo, **ou** hipócrita **ou** incoerente. **Se** esse conselho disser que esses deputados e senadores que cometeram crime de ofensa, no cometeram crime **porque** tem **imunidade parlamentar**. Aí tem!!!
6.7	A prisão é ilegal já no momento que ele usa um **inquérito que** serve pra **tudo**, sem requerimento da polícia ou do **MP** o resto só invalida **ainda mais** a prisão, **o que esse ministro fez é caso de impeachment e prisão**
6.8	Mas também deputada, com essa oposição que **tudo** que o governo federal faz vocês acham que está errado. Imagine **se** o povo estivesse **todos** seguindo o **FIQUE EM CASA, A ECONOMIA A GENTE VER DEPOIS**. **Sou** a favor que sejam seguidos os protocolos: máscara, lavar as mãos e não aglomerar.
6.9	Deveria abrir uma **CPI para investigar** as **sabotagens** que o ex-presidente da câmara comandou durante seu mandato, **causando prejuízos incalculáveis** à **toda nação** e desrespeitando a vontade de mais de **56 milhões de eleitores**!

Tabela 6: 9 *tweets* de alta qualidade presentes no corpus de teste corretamente classificados pelo modelo ajustado a partir do BERTimbau.

5.2.2. Análise dos erros

A seguir, apresenta-se uma breve discussão das principais disparidades observadas entre a anotação humana e a classificação com o modelo ajustado a partir do BERTimbau. Na Tabela 7, são apresentadas algumas mensagens do conjunto de testes que demonstraram divergência entre as avaliações humanas e automáticas.

Os exemplos 7.8, 7.9 e 7.10, da Tabela 7, são alguns dos casos nos quais foram observadas as principais divergências, uma vez que o modelo os classificou como Baixa argumentatividade e os anotadores os avaliaram como Alta argumentatividade. Características como ausência de vírgulas no exemplo 7.8, presença de maiúsculas no exemplo 7.9 e de *hashtags* no exemplo 7.10 podem ter influenciado o modelo a classificá-los como de Baixa qualidade. Por outro lado, os anotadores humanos provavelmente os classificaram como de Alta qualidade com base no conteúdo e apelo emocional e não na forma superficial do texto.[32] No restante dos exemplos (7.11, 7.12, 7.13 e 7.14) houve diferenças entre Baixa/Média e Média/Alta argumentatividade. Essas diferenças foram vistas como plausíveis, considerando o contexto semântico do argumento.

[32]Os exemplos 7.8 e 7.9 foram classificados como de Alta qualidade pela maioria dos anotadores, e o exemplo 7.10 foi classificado como de Alta qualidade por todos os anotadores.

Além dos exemplos de erros apresentados na Tabela 7, outras 12 mensagens foram classificadas de forma equivocada pelo modelo. Destas, 6 de Baixa argumentatividade foram classificadas como Média e 6 de Média argumentatividade foram classificadas como Baixa. Importante ressaltar que todas as mensagens classificadas como de Alta qualidade da argumentação pelo modelo ajustado a partir do BERTimbau foram avaliadas da mesma forma pelos anotadores, isso indica uma precisão absoluta em classificar mensagens de alta argumentatividade.

6. Conclusões, limitações e trabalhos futuros

Este artigo apresentou experimentos para a predição da qualidade da argumentação em *tweets* do domínio da política brasileira. Os experimentos foram realizados com algoritmos de AM baseado em *features* e em redes neurais. O modelo que apresentou o melhor desempenho foi o obtido a partir do ajuste fino do BERTimbau.

Em relação às questões de pesquisa inicialmente propostas para este trabalho, a partir dos experimentos aqui apresentados pode-se concluir que: (Q1) é possível predizer automaticamente a qualidade da argumentação em *tweets* do domínio da política brasileira e (Q2) a maneira que se mostrou mais adequada para fazê-lo foi usando o modelo neural gerado a partir do

Ex.	Texto	Real	BERTimbau
7.1	**Acho que** o foco deveria ser nas desastrosas intervenções nas estatais (**que já** n deveriam ser mais estatais) e nos preços de energia, que logo se avizinham. N adianta ser **um leão contra** a prisão (q deveria ter discurso técnico), e **um gato contra** o desastre das políticas do gov.!	Alta	Média
7.2	**Precisamos tds** deixar bem claro que essa ñ representa o povo e que esses parlamentares são responsáveis pelas mortes diárias de brasileiros. É imoral que estejam votando um projeto de lei para se protegerem da lei **em vez de** salvar a vida de brasileiros!	Alta	Média
7.3	Quem sabe né o seu Jair da uma dentro. O dublê de ministro da saúde tá sendo um fiasco, um fracasso, uma vergonha nacional. Uma coisa é usina hidroelétrica outra coisa é a Petrobrás. Veremos.	Alta	Média
7.4	Hoje fui comprar 1kg de carne moída para o almoço e deu R$ 43,00 achei um absurdo! Em que mundo estamos com um custo tão alto de carne assim?! Mas de boa estamos comprando carne de primeira para nossos representantes políticos, então pq reclamar?! 😳🧐	Alta	Média
7.5	Cadê a dengue, cadê a gripe comum, a H1N1, cadê as demais doenças respiratórias, tão comuns em nosso país nessa época do ano? Sumiram ou estão se somando aos números do K0V1D? Tudo pela ciência e pela saúde! Principalmente agora que governadores e prefeitos podem comprar vac1na.	Alta	Média
7.6	O que o Brasil precisa é um sistema bancário que ofereca juros selic 2% ao ano para qualquer um empreender. Reduzir esse spread avassalador. Infelizmente politico so pensa em imposto e nao em melhorar a eficiencia.	Alta	Média
7.7	Esse é a teoria, . Na vida real, na prática brasileira, vendem água mineral a 30 reais, quando tem desastre natural... Privatizem mesmo! Abram o mercado sim! Quebrem os monopólios! Mas, acima de tudo, FISCALIZEM os espertinhos, senão eles deitam e rolam!	Alta	Média
7.8	Ah pelo amor Coco Bambú vive lotado se não tem grana para manter os funcionários durante o Lockdown tem péssima administração! Empatia zero e o governo federal precisa ajudar as pessoas essa que é a realidade!	Alta	Baixa
7.9	NAO É CRIME DE OPINIAO. VOU REPETIR PRO SARGENTO GARCIA ENTENDER. NAO É CRIME DE OPINIAO, e nao está coberto peka imunidade material so parlamentar, basta saber ler o art. 53. Mas é esperar muito de bolsonaristas. Que vergonha pro meu Parana ter deputado assim.	Alta	Baixa
7.10	Pode jogar a nossa Constituição no lixo, pois a acabam de transformar os 11 ministros do STF nos novos IMPERADORES do Brasil. #CamaraDosDeputadosVergonhaNacional #CâmaraDosDeputadosRasgandoAConstituição #DanielSilveiraFalouPeloPovo #STFVergonhaMundial	Alta	Baixa
7.11	Você pode ter 6 seguranças armados né deputado, e defender bandido, porque não pede ao STF para mandar investigar todo o governo do Rio de janeiro incluindo vocês de Dourados, vereadores e senadores, talvez assim consigam desarmar os traficantes dos morros	Média	Baixa
7.12	Os e o Sr perderam o meu respeito não acredito mais nessa turma de cristão não tem nada lembrei de Bararabas e dos lideres do sinedrio foi igual a passagem veio a minha mente no voto de muitos da AD e da IURD	Média	Baixa
7.13	Do jeito que esses deputados são, depois que vimos semana passada, acho difícil conseguir as assinaturas necessárias. São um bando de covarde, só pensam neles, no próprio ego. Nem a voz do povo eles escutam. 2022 é logo ali. #TodoPoderEmanaDoPovo #Democracia	Baixa	Média
7.14	Nada melhora pro povo já que não há instituições neste país apodreceu tdo e fede faz tempo, mas pelo menos é uma dívida a menos nas costas dos pagadores de impostos, mais conhecidos pela elite como trouxas	Baixa	Média

Tabela 7: Exemplos de erros de predição do modelo neural.

ajuste fino do BERTimbau. Tal modelo foi capaz de predizer com 100% de precisão instâncias da classe de Alta qualidade da argumentação, satisfazendo o objetivo deste trabalho que é o de encontrar *tweets* com boa capacidade argumentativa.

Embora o objetivo do trabalho tenha sido alcançado, algumas limitações são apresentadas na próxima subseção, seguida de algumas propostas de trabalhos futuros.

6.1. Limitações

Uma das principais limitações deste trabalho é o tamanho do corpus. Embora seja compatível com o corpus desenvolvido por Wachsmuth et al. (2017b) quanto ao número de instâncias, os resultados aqui apresentados têm seus impactos limitados pela pouca quantidade de instâncias usadas na geração e na avaliação dos modelos.

Outra limitação está relacionada à decisão de projeto adotada no momento da anotação do corpus que foi a de realizar a anotação da qualidade da argumentação de forma isolada, ou seja, desconsiderando o *tweet* semente. Diante disso, assumiu-se que o argumento pode ser classificado independentemente do tópico. Trabalhos da literatura (Fromm et al., 2019; Hidayaturrahman et al., 2021) apontam que, muitas vezes, frases contendo argumentos são estruturalmente semelhantes a frases puramente informativas sem qualquer posicionamento sobre o tópico e que considerar a informação do tópico é crucial para a tarefa, pois ele define o contexto semântico de um argumento. Para lidar com essa limitação, uma possível estratégia é concatenar as mensagens avaliadas com o *tweet* semente de modo a incorporar contexto e, consequentemente, melhorar o desempenho do modelo.

Outra decisão de projeto que pode ter impactado o desempenho do modelo está relacionada ao modo como os aspectos foram combinados para definir a Qualidade Geral da argumentação. Conforme visto na Tabela 7, algumas instâncias de teste anotadas como de Alta qualidade foram classificadas pelo modelo BERTimbau como de Média ou Baixa qualidade o que, analisando o conteúdo de tais *tweets* pode ter fundamento. Assim, uma proposta de trabalho futuro é fazer uma revisão dos critérios e do modo como os aspectos são combinados para definir a Qualidade Geral.

É importante considerar as características específicas do Twitter. Diferentemente de outras mídias sociais, o Twitter não possui uma política das mais rígidas para restringir ou filtrar o conteúdo das postagens ou o comportamento abusivo dos usuários. Por causa disso, postagens que contêm palavrões e discurso de ódio são comuns. No campo da política brasileira atual, essas características são ainda mais acentuadas, com postagens contendo *fake news*, ataques pessoais a políticos ou às famílias deles, ideologia política, entre outros. Assim, esses textos (*tweets*) costumam ter várias marcas que impactam negativamente a qualidade deles e, consequentemente, reduzem a qualidade geral da argumentação.

Por fim, outra limitação relacionada ao Twitter é o número muito limitado de caracteres (280) permitido para cada mensagem, o que dificulta o uso de estratégias mais elaboradas de argumentação linguística pelos autores das postagens.

6.2. Trabalhos futuros

Como trabalhos futuros, destacam-se três caminhos que trariam maior benefício às propostas aqui apresentadas: (i) como apresentado na Seção 6.1, concatenar as mensagens avaliadas com o *tweet* semente de modo a incorporar contexto e, consequentemente, melhorar o desempenho do modelo; (ii) testar uma combinação (*ensemble*) de classificadores, que constitui-se em múltiplos classificadores, treinados de forma individual, cujos resultados são combinados, também com o objetivo de buscar uma melhora no desempenho do modelo; (iii) aumentar o corpus com a finalidade de capturar novos contextos, além de um acréscimo da sua própria dimensão.

Além dessas, outra possibilidade seria gerar modelos específicos para cada um dos aspectos da qualidade da argumentação (Clareza, Organização, Credibilidade e Polaridade e Intensidade do Apelo Emocional) com o intuito de combiná-los para definir automaticamente a qualidade geral da argumentação, como foi feito de modo manual, pelos anotadores, no momento da geração do corpus (Silva et al., 2021).

Agradecimentos

Os autores agradecem aos linguistas e coautores em (Silva et al., 2021), que anotaram o corpus e contribuíram para a elaboração das diretrizes de anotação, utilizados neste trabalho: Amanda Pontes Rassi, Jackson Wilke da Cruz Souza, Renata Ramisch e Roger Alfredo de Marci Rodrigues Antunes. Agradecem, também, ao Sidney Evaldo Leal e ao Núcleo Interinstitucional de Linguística Computacional (NILC) pelos recursos linguístico-computacionais disponibilizados para o desenvolvimento desta pesquisa. Por fim, os autores agradecem ao Programa de Pós-Graduação em Ciência da Computação (PPGCC) da Universidade Federal de São Carlos (UFSCar) e à Rede Gonzaga de Ensino Superior (REGES), pelo apoio a este trabalho.

Referências

Adi, Sumarni, Yoga Pristyanto & Andi Sunyoto. 2019. The best features selection method and relevance variable for web phishing classification. Em *International Conference on Information and Communications Technology (ICOIACT)*, 578–583. 10.1109/ICOIACT46704.2019.8938566.

Al-Khatib, Khalid, Henning Wachsmuth, Matthias Hagen, Jonas Köhler & Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 1395–1404. 10.18653/v1/N16-1165.

Armentano-Oller, Carme, Rafael C. Carrasco, Antonio M. Corbí-Bellot, Mikel. L. Forcada, Mireia Ginestí-Rosell, Sergio Ortiz-Rojas, Juan. Antonio Pérez-Ortiz, Gema Ramírez-Sánchez, Felipe Sánchez–Martínez & Miriam. A. Scalco. 2006. Open-source Portuguese–Spanish machine translation. Em *VII Encontro para o Processamento Computacional da Língua Portuguesa (PROPOR)*, 50–59.

Balage Filho, Pedro P., Thiago Alexandre Salgueiro Pardo & Sandra M. Aluísio. 2013. An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 215–219.

Bench-Capon, Trevor JM & Paul E Dunne. 2007. Argumentation in artificial intelligence. *Artificial intelligence* 171(10-15). 619–641. 10.1016/j.artint.2007.05.001.

Bilu, Yonatan & Noam Slonim. 2016. Claim synthesis via predicate recycling. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 525–530. 10.18653/v1/P16-2085.

Blair, J. Anthony. 2012. Rhetoric, dialectic, and logic as related to argument. *Philosophy & Rhetoric* 45(2). 148–164. 10.5325/philrhet.45.2.0148.

Boudry, Maarten, Fabio Paglieri & Massimo Pigliucci. 2015. The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation* 29(4). 431–456. 10.1007/s10503-015-9359-1.

Brum, Henrico & Maria Graças Volpe Nunes. 2018. Building a sentiment corpus of tweets in Brazilian Portuguese. Em *11th International Conference on Language Resources and Evaluation (LREC)*, 4167–4172.

Capellaro, Leonardo & Helena Caseli. 2021. Análise de polaridade e de tópicos em tweets no domínio da política no Brasil. Em *XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, 47–55. 10.5753/stil.2021.17783.

Carlile, Winston, Nishant Gurrapadi, Zixuan Ke & Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. Em *56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 621–631. 10.18653/v1/P18-1058.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 4171–4186. 10.18653/v1/N19-1423.

Eemeren, Frans H & Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation* 1(3). 283–301. 10.1007/BF00136779.

Eemeren, Frans H. van & Rob Grootendorst. 2003. *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press. 10.1017/CBO9780511616389.

Feng, Vanessa Wei, Ziheng Lin & Graeme Hirst. 2014. The impact of deep hierarchical discourse structures in the evaluation of text coherence. Em *25th International Conference on Computational Linguistics (COLing)*, 940–949.

Fonseca, Erick. Rocha & João Luís G. Rosa. 2013. Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 98–107.

Fromm, Michael, Max Berrendorf, Johanna Reiml, Isabelle Mayerhofer, Siddharth Bhargava, Evgeniy Faerman & Thomas Seidl. 2022. Towards a holistic view on argument quality prediction. 10.48550/ARXIV.2205.09803. ArXiv cs.CL.

Fromm, Michael, Evgeniy Faerman & Thomas Seidl. 2019. TACAM: Topic and context aware argument mining. Em *IEEE/WIC/ACM International Conference on Web Intelligence*, 99–106. 10.1145/3350546.3352506.

García-Gorrostieta, Jesús Miguel & Aurelio López-López. 2018. Identifying argumentative paragraphs: Towards automatic assessment of argumentation in theses. Em *International Conference on Applications of Natural Language to Information Systems*, 83–90. 10.1007/978-3-319-91947-8_9.

García-Gorrostieta, Jesús M., Aurelio López-López & Samuel González-López. 2018. Automatic argument assessment of final project reports of computer engineering students. *Computer Applications in Engineering Education* 26(5). 1217–1226. 10.1002/cae.21996.

Gleize, Martin, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov & Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. Em *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 967–976. 10.18653/v1/P19-1093.

Gretz, Shai, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov & Noam Slonim. 2019. A large-scale dataset for argument quality ranking: Construction and analysis. 10.48550/ARXIV.1911.11408. ArXiv cs.CL.

Habernal, Ivan & Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2127–2137. 10.18653/v1/D15-1255.

Habernal, Ivan & Iryna Gurevych. 2016. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional LSTM. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1589–1599. 10.18653/v1/P16-1150.

Habernal, Ivan & Iryna Gurevych. 2017. Argumentation mining in user-generated web discourse. *Computational Linguistics* 43(1). 125–179. 10.1162/COLI_a_00276.

Hidayaturrahman, Emmanuel Dave, Derwin Suhartono & Aniati Murni Arymurthy. 2021. Enhancing argumentation component classification using contextual language model. 10.1186/s40537-021-00490-2.

Krippendorff, Klaus. 2011. Computing krippendorff 's alpha-reliability. Relatório técnico. University of Pennsylvania. http://repository.upenn.edu/asc_papers/43/.

Lauscher, Anne, Lily Ng, Courtney Napoles & Joel Tetreault. 2020. Rhetoric, logic, and dialectic: Advancing theory-based argument quality assessment in natural language processing. Em *27th International Conference on Computational Linguistics (COLing)*, 4563–4574. 10.18653/v1/2020.coling-main.402.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2022. NILC-Metrix: assessing the complexity of written and spoken language in Brazilian Portuguese. 10.48550/ARXIV.2201.03445. ArXiv cs.CL.

Leite, João Augusto, Diego Silva, Kalina Bontcheva & Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. Em *1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 914–924.

Lytos, Anastasios, Thomas Lagkas, Panagiotis Sarigiannidis & Kalina Bontcheva. 2019. The evolution of argumentation mining: From models to social media and emerging tools. *Information Processing & Management* 56(6). 102055. 10.1016/j.ipm.2019.102055.

Marcuschi, Luiz Antônio et al. 2002. Gêneros textuais: definição e funcionalidade. Em *Gêneros textuais e ensino*, Lucerna.

Maćkiewicz, Andrzej & Waldemar Ratajczak. 1993. Principal components analysis (PCA). *Computers & Geosciences* 19(3). 303–342. 10.1016/0098-3004(93)90090-R.

Misra, Puneet & Arun Singh Yadav. 2020. Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal of Emerging Technologies* 11(3). 659–665.

Peldszus, Andreas & Manfred Stede. 2015. Joint prediction in MST-style discourse parsing for argumentation mining. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 938–948. 10.18653/v1/D15-1110.

Persing, Isaac & Vincent Ng. 2015. Modeling argument strength in student essays. Em *53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 543–552. 10.3115/v1/P15-1053.

Potthast, Martin, Lukas Gienapp, Florian Euchner, Nick Heilenkötter, Nico Weidmann, Henning Wachsmuth, Benno Stein & Matthias Hagen. 2019. Argument search: Assessing argument relevance. Em *42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1117–1120. 10.1145/3331184.3331327.

Putra, Jan Wira Gotama, Simone Teufel & Takenobu Tokunaga. 2021. Annotating argumentative structure in English-as-a-foreign-language learner essays. *Natural Language Engineering* 28(6). 797–823. 10.1017/S1351324921000218.

Reimers, Nils & Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)*, 3982–3992. 10.18653/v1/D19-1410.

Reimers, Nils, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab & Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. Em *57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 567–578. 10.18653/v1/P19-1054.

Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M. Khapra, Ehud Aharoni &

Noam Slonim. 2015. Show me your evidence - an automatic method for context dependent evidence detection. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 440–450. 10.18653/v1/D15-1050.

Rosenfeld, Ariel & Sarit Kraus. 2016. Providing arguments in discussions on the basis of the prediction of human argumentative behavior. *ACM Transactions on Interactive Intelligent Systems* 6(4). 1–33. 10.1145/2983925.

Rossini, Patrícia. 2019. Disentangling uncivil and intolerant discourse in online political talk. Em *A Crisis of Civility?*, 142–157. Routledge. 10.4324/9781351051989-9.

Rossini, Patrícia. 2022. Beyond incivility: Understanding patterns of uncivil and intolerant discourse in online political talk. *Communication Research* 49(3). 399–425. 10.1177/0093650220921314.

Schaefer, Robin & Manfred Stede. 2021. Argument mining on twitter: A survey. *it – Information Technology* 63(1). 45–58. 10.1515/itit-2020-0053.

Silva, Cássio, Amanda Rassi, Jackson Souza, Renata Ramisch, Roger Antunes & Helena Caseli. 2021. Quality of argumentation in political tweets: what is and how to measure it / qualidade da argumentação em tweets de política: o que e como avaliar. *Estudos da Linguagem* 29(4). 2537–2586. 10.17851/2237-2083.29.4.2537-2586.

Skitalinskaya, Gabriella, Jonas Klaff & Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. Em *16th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1718–1729. 10.18653/v1/2021.eacl-main.147.

Souza, Fábio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *Brazilian Conference on Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28.

Stab, Christian & Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. Em *25th International Conference on Computational Linguistics (COLing)*, 1501–1510.

Stab, Christian & Iryna Gurevych. 2017a. Parsing argumentation structures in persuasive essays. *Computational Linguistics* 43(3). 619–659. 10.1162/COLI_a_00295.

Stab, Christian & Iryna Gurevych. 2017b. Recognizing insufficiently supported arguments in argumentative essays. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 980–990.

Swanson, Reid, Brian Ecker & Marilyn Walker. 2015. Argument mining: Extracting arguments from online dialogue. Em *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 217–226. 10.18653/v1/W15-4631.

Toledo, Assaf, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov & Noam Slonim. 2019. Automatic argument quality assessment: New datasets and methods. Em *Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP)*, 5625–5635. 10.18653/v1/D19-1564.

Toulmin, Stephen E. 2003. *The uses of argument*. Cambridge university press.

Vapnik, Vladimir. 1999. *The nature of statistical learning theory*. Springer.

Wachsmuth, Henning, Khalid Al-Khatib & Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. Em *26th International Conference on Computational Linguistics (COLing)*, 1680–1691.

Wachsmuth, Henning, Johannes Kiesel & Benno Stein. 2015. Sentiment flow - a general model of web review argumentation. Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 601–611. 10.18653/v1/D15-1072.

Wachsmuth, Henning, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych & Benno Stein. 2017a. Argumentation quality assessment: Theory vs. practice. Em *55th Annual Meeting of the Association for Computational Linguistics (ACL)*, 250–255. 10.18653/v1/P17-2039.

Wachsmuth, Henning, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst & Benno Stein. 2017b. Computational argumentation quality assessment in natural language. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 176–187.

Wachsmuth, Henning, Martin Potthast, Khalid Al-Khatib, Yamen Ajjour, Jana Puschmann,

Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff & Benno Stein. 2017c. Building an argument search engine for the web. Em *4th Workshop on Argument Mining (ArgMining 2017)*, 49–59.

Wachsmuth, Henning, Benno Stein & Yamen Ajjour. 2017d. "PageRank" for argument relevance. Em *15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 1117–1127.

Wachsmuth, Henning & Till Werner. 2020. Intrinsic quality assessment of arguments. Em *28th International Conference on Computational Linguistics*, 6739–6745. 10.18653/v1/2020.coling-main.592.

Walton, Douglas N & David N Walton. 1989. *Informal logic: A handbook for critical argument.* Cambridge University Press.

Wei, Zhongyu, Yang Liu & Yi Li. 2016. Is this post persuasive? ranking argumentative comments in online forum. Em *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 195–200. 10.18653/v1/P16-2032.

Weltzer-Ward, Lisa, Beate Baltes & Laura Knight Lynn. 2009. Assessing quality of critical thought in online discussion. *Campus-Wide Information Systems* 26(3). 168–177. 10.1108/10650740910967357.

Zhang, Justine, Ravi Kumar, Sujith Ravi & Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in Oxford-style debates. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NACL)*, 136–141. 10.18653/v1/N16-1017.

Novas Perspectivas

Proposta recebida em Fevereiro 2023 e aceite para publicação em Maio 2023 .

A compilação e a análise de métricas textuais de um corpus de redações

Compilation and analysis of textual metrics of an essay's corpus

Átila Augusto Soares Vital
Faculdade de Letras da Universidade Federal de Minas Gerais

Resumo

A prova de redação do Exame Nacional do Ensino Médio (Enem) é decisiva para a garantia da vaga em instituições de ensino superior no Brasil. De 2010 a 2020, foi observado que a quantidade de redações avaliadas em nota máxima (mil pontos) caiu de maneira drástica e abrupta: de 3.694 redações nota máxima em 2011 para apenas 28 em 2020. O objetivo deste trabalho é apresentar um corpus de redações nota máxima avaliadas pela banca do Enem, descrevê-las e tecer breves considerações a partir da análise de métricas textuais na série histórica de 2010 a 2020. A compilação foi feita de forma manual, pela internet. Para as descrições, foram utilizados o programa Orange: Data Mining e o analisador de complexidade textual NILC-Metrix (Leal et al., 2022). Os resultados sugerem que houve aumento expressivo no número de palavras e diminuição da razão type/token ao longo dos anos. Além disso, foram feitas medidas sintáticas que constataram o aumento da complexidade dos textos.

Palavras chave

redações; linguística de corpus; complexidade textual

Abstract

The writing test of the National High School Exam (Enem) is very important to guarantee a place for students in undergraduate institutions in Brazil. From 2010 to 2020, the number of texts evaluated in maximum grade (one thousand points) dropped abruptly: in 2011, 3,694 texts gained 1,000 points, and in 2020, only 28 texts were evaluated with the same grade. The objective of this research is to present a corpus of texts graded one thousand points by Enem's team, to describe them and to make brief considerations about their characteristics during the historical series from 2010 to 2020. The compilation was made manually, using the internet. We used Orange: Data Mining and the NILC-Metrix (Leal et al., 2022) textual complexity analyzer. The results suggest an expressive increase in the number of words and a decrease in the type/token ratio during the period. Finally, syntactic metrics were measured and confirmed the increase in textual complexity.

Keywords

essays; corpus linguistics; textual complexity

1. Introdução

O Exame Nacional do Ensino Médio (Enem) ocorre anualmente no Brasil desde 1998, tendo sido criado com a intenção de avaliar os estudantes da educação básica. A partir de 2009, com a aderência da maior parte das universidades brasileiras ao exame, houve busca crescente pela prova, acompanhada pelas revisões nos critérios de correção e nos assuntos a serem discutidos nas questões de múltipla escolha. No caso da redação, há, uma vez a cada ano, a disponibilização de materiais para os alunos por parte do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) — a Cartilha de Redação — e para os corretores — o Manual do Corretor — onde são arrolados os critérios para a correção dos textos dissertativo-argumentativos.

Boa parte do trabalho dos docentes, seja em cursinhos pré-vestibulares ou em escolas preparatórias, se relaciona com a análise detida dos critérios em cada um dos manuais prescritivos, de modo que se possa ter uma visão sucinta dos elementos necessários para a escrita da redação, que deve pertencer ao tipo textual dissertativo-argumentativo. As correções devem ser pautadas em cinco competências, que procuram tornar objetiva a análise (i) da norma padrão da Língua Portuguesa; (ii) da correspondência ao tipo textual; (iii) da consistência da argumentação; (iv) da coesão e (v) da apresentação de uma proposta de intervenção. Esses elementos e suas explicações oficiais podem ser visualizados nos tópicos abaixo, que relacionam cada

DOI: 10.21814/lm.15.1.393

competência com seus objetivos de avaliação, segundo a Cartilha do Participante. As notas de cada competência variam de 0 a 200 pontos e, somadas, constituem a nota final da prova de redação (Brasil, 2020).

- *Competência I*: Demonstrar domínio da modalidade escrita formal da Língua Portuguesa;
- *Competência II*: Compreender a proposta de redação e aplicar conceitos das áreas de conhecimento, dentro dos limites do texto dissertativo-argumentativo em prosa;
- *Competência III*: Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista;
- *Competência IV*: Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação;
- *Competência V*: Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos;

Além disso, na etapa do Sistema de Seleção Unificada (SISU), principal meio de alocação dos candidatos em vagas nas universidades públicas, a nota da prova de redação é inserida na média simples em relação às outras áreas do conhecimento. Em muitos casos, como salientam Cançado et al. (2020, pp. 64), essa nota "é responsável por uma grande parte da classificação de um candidato, fechando ou abrindo as portas de entrada em nossas universidades."

Curiosamente, dados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) mostram que, embora o número de participantes nas edições do ENEM tenha aumentado consideravelmente desde 1998, houve, a partir de 2011, queda acentuada na quantidade de textos avaliados em nota máxima — 1000 pontos. Partindo de um pico de 3.694 avaliações nota mil em 2011, em 2020, apenas 28 textos foram avaliados como nota máxima no exame, evidenciando a diminuição de textos exemplares (Figura 1).

O objetivo deste trabalho, portanto, é, a partir da ferramenta NILC-Metrix (Leal et al., 2022), desenvolvida pelo Núcleo Interinstitucional de Linguística Computacional (NILC) da Universidade de São Paulo (USP), descrever, de formas quantitativa e qualitativa, algumas das características linguísticas das redações nota mil ao longo dos anos de 2010 a 2020, período crítico em que houve diminuição expressiva da quantidade de textos avaliados em nota máxima. Dessa forma, evidenciar o perfil dos textos exemplares

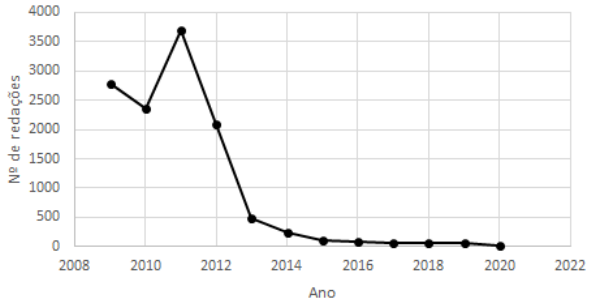

Figura 1: Quantidade de redações avaliadas em nota 1000 ao longo dos anos. Fonte: Sinopses Estatísticas do INEP.

no decorrer do período pode ajudar a reunir características linguísticas para futuros estudos em redações do Enem, além de revelar em que medida os textos da série histórica se aproximam e se distanciam. Para o caso de diferenças significativas entre os períodos, é possível pensar na mudança da estrutura dos textos coligada à mudança de critérios de correção. Este trabalho, portanto, oferece subsídios para se pensar a mudança estrutural para, posteriormente, se fazer a análise dos critérios e incrementar o corpus com textos de variados níveis.

Para isso, compilamos, sob o viés metodológico da Linguística de Corpus e Computacional, um corpus de 96 redações nota mil que passaram pela correção oficial do exame ao longo do período analisado. Como o corpus conta apenas com textos nota máxima, foi possível observar mudanças estruturais, ao longo do tempo, nas redações que foram avaliadas, a princípio, num mesmo patamar. Os dados foram arquitetados de modo que pudéssemos ter pelo menos um texto de cada um dos anos, totalizando, até o momento, 37.459 palavras, em uma razão type/token de 0,1606.

Para o estudo linguístico a partir da análise do corpus, lançamos mão da consagrada conceptualização histórica de Berber Sardinha (2000), que salienta que a descrição de um conjunto de dados está diretamente relacionada à representatividade e, portanto, ao caráter probabilístico do uso linguístico. Desse modo, as conclusões a serem observadas no estudo poderão assumir valores de verdade apenas em relação aos dados elencados na compilação. É importante salientar o fato de que, até o momento da confecção deste trabalho, não havia disponíveis corpora expressivos de redações nota mil corrigidas pela banca oficial do Enem.

Com o auxílio de métodos computacionais, há poucos trabalhos disponíveis que procuram correlacionar notas ou níveis de proficiência escrita em uma dada língua com métricas textuais. Esse é o caso de Crossley & McNamara (2012), que, através da ferramenta Coh-Metrix (Graesser et al., 2004), examinaram a relação entre as estratégias de coesão e de sofisticação linguística com o nível de proficiência conferido a aprendizes de inglês como segunda língua (L2). Como medidas de coesão, os autores observaram o uso de operadores lógicos, retomada lexical, coesão sequencial, referenciação semântica, causalidade, conectivos e diversidade lexical. Por sofisticação, os autores consideram como sendo a capacidade de produção de estruturas linguísticas menos frequentes e tomam como principais métricas medidas psicolinguísticas, frequência de hiperônimos, frequência de palavras e complexidade sintática. Como resultados, o estudo mostrou que as medidas de sofisticação se correlacionam com o grau de proficiência dos alunos.

Outro trabalho que parte de textos de aprendizes de inglês como L2 é o de Alexopoulou et al. (2017), que investigou os efeitos linguísticos nos textos dos alunos que podem ser gerados pela complexidade das tarefas e pelo enfoque instrucional dado por elas. Nesse sentido, diferentes instruções para o desenvolvimento de textos gera particularidades em suas estruturas, como diferentes distribuições de tempos verbais, complexidade sintática e uso de verbos irregulares. Para aprendizes de português como língua estrangeira, está disponível a plataforma LX-CEFR, apresentada por Branco et al. (2014), que calcula o grau de proficiência linguística a partir do Quadro Comum Europeu de Referência para Línguas (CEFR). O processamento do texto a partir do LX-CEFR enquadra nos níveis A1, A2, B1, B2 ou C1 as seguintes métricas textuais: índice Flesch, tamanho médio das frases (em palavras), tamanho médio das palavras (em sílabas) e proporção de nomes. Com as possibilidades de se compilar e de se treinar modelos de linguagem natural, é cada vez mais frequente o surgimento de ferramentas para avaliação automática. Como bem salientam Branco et al. (2014), é importante que esses recursos sejam utilizados de modo a complementarem a avaliação do aprendizado de línguas estrangeiras, já que, em muitos casos, os modelos são desenvolvidos com um modesto número de textos, fator limitante da acurácia.

Westerlund (2019), após compilar 30 ensaios produzidos em inglês para o Swedish National Exam por estudantes do segundo grau na Suécia, se utilizou de uma metodologia similar à de Crossley & McNamara (2012), processando os textos na ferramenta Coh-Metrix 3.0 e confirmando resultados semelhantes: textos com maior sofisticação, isto é, com uso de palavras menos frequentes, hiperônimos, voz passiva e alta densidade de sintagmas adverbiais foram aqueles que receberam notas mais altas.

Para análises que não lançam mão das métricas propostas por ferramentas de tratamento de textos como o Coh-Metrix, há, no Brasil, o desenvolvimento incipiente de trabalhos voltados para a avaliação automática de redações (Automatic Essay Scoring), como é o caso de Amorim & Veloso (2017). No estudo, os autores propõem um sistema de análise de múltiplos aspectos (multi-aspect analysis) para a correção automática de redações, considerando um dataset de 1840 textos corrigidos pelos corretores do website *Educação UOL* e métricas desenvolvidas para a língua inglesa. De forma similar, com objetivo de fornecer conjunto de dados para análises de redações do Enem, Marinho et al. (2021) criam um corpus com 4.570 redações coletadas das plataformas *Vestibular UOL* e *Educação UOL*. Com um conjunto de textos ainda maior, de 56.644 redações, Fonseca et al. (2018) criaram outro sistema de avaliações automático baseado em redes neurais. Nesse caso, o modelo considerou métricas de 5 grupos principais:

1. *Count metrics*: contagem de estatísticas básicas dos textos, como número de vírgulas, número de caracteres, número de parágrafos, tamanho médio das sentenças, dentre outras;
2. *Specific expressions*: contagem de grupos de palavras esperados em redações, como agentes sociais, conectores discursivos, palavras propositivas e marcas de oralidade;
3. *Token n-grams*: verifica a correlação entre a ocorrência de n-gramas e as notas altas;
4. *POS n-grams*: verifica a classe de palavras dos n-gramas;
5. *POS counts*: conta o número de classes de palavras no texto.

É importante evidenciar que os trabalhos elencados acima para redações em português brasileiro i) coletaram textos que não necessariamente passaram pela correção oficial do Enem e ii) e que possuem pontuações variadas - não apenas nota mil. Para este artigo, optamos pela coleta de redações que restritivamente passaram pela correção oficial e que receberam nota máxima (mil pontos) ao longo do período entre 2010 e 2020. Com isso, ao obtermos diferenças significativas em métricas textuais ao longo da série

estudada, teremos indícios que denotarão diferentes perfis linguísticos para textos avaliados num mesmo nível pelos corretores da banca oficial.

Nesse sentido, para o refinamento metodológico, foram utilizados outros artigos que se relacionam com a temática da linguística das redações do ENEM, dentre eles, citamos Buzato et al. (2021) e Cançado et al. (2020), que se assentam diretamente na Linguística de Corpus, Bertucci et al. (2020) e Cruz et al. (2021), que tratam, respectivamente, da ocorrência de anáforas encapsuladoras em textos do Exame e da argumentação em propostas de intervenções de redações nota mil. Na esteira da Linguística Computacional, lançamos mão do trabalho de Westerlund (2019), que correlaciona métricas textuais com as notas de avaliações do Exame Nacional Sueco (Swedish National Exam) a partir das métricas da ferramenta Coh-Metrix, da qual o NILC-Metrix (Leal et al., 2022) é uma das derivações. Além disso, as investigações de avaliação automática de textos para aprendizado de segunda língua, como Alexopoulou et al. (2017), para o inglês, e Branco et al. (2014), para o português foram importantes para a composição metodológica.

2. Metodologia

Diferentemente de boa parte dos trabalhos sobre redações do Enem, nosso objetivo foi levar em consideração apenas os textos que foram corrigidos pela banca oficial de corretores da prova. Nesse sentido, os passos metodológicos adotados passeiam pelas fases de:

1. pesquisa e compilação dos textos;
2. caracterização do corpus coletado;
3. escolha dos programas e das métricas de análise;
4. apresentação dos resultados e da análise linguística.

Na fase de pesquisa e compilação, foram coletados, de diferentes domínios da internet, textos de redações nota mil. A maior parte delas é proveniente de sites que veiculam boas práticas de escrita e que fornecem, de forma gratuita, instruções e cursos para a realização da prova, além de textos nota mil publicados em reconhecidos sites de notícias e reportagens.

Os textos foram retirados dos domínios na internet e dispostos em arquivos `.txt` com codificação UTF-8, nomeados conforme o ano em que foram escritos e o número associado a cada redação específica daquele ano. O exemplo 1 ilustra um trecho do texto 8, escrito e corrigido no ano de 2014.

> Exemplo 1
>
> *Durante o século XX, o estímulo à produção industrial, por Getúlio Vargas, e o incentivo à integração nacional, de Juscelino Kubitschek, foram fatores que possibilitaram a popularização dos meios de comunicação no Brasil. Com isso, cresceu também a publicidade infantil, que busca introduzir nas crianças, desde cedo, o princípio capitalista de consumo. No entanto, essa visão negativa pode ser significativamente minimizada, desde que acompanhada de uma forte base educacional que auxilia as crianças a discernir por meio do desenvolvimento de senso crítico próprio.*

Para a análise automática do corpus, foram selecionadas três vias principais, cada uma com suas métricas e bibliotecas particulares: (i) script em *Python*, para visualização das palavras mais frequentes; (ii) *Orange: data mining*, para aferição da razão type/token; (iii) NILC-Metrix, para o cálculo da complexidade textual.

Além de (i) e (ii), mecanismos extremamente difundidos para análise de dados em geral, a interface do NILC-Metrix foi desenvolvida para que os textos possam ser processados em relação a 200 métricas de complexidade textual, coerência e coesão. As métricas são divididas em 14 grandes grupos, considerando a natureza das análises e as técnicas de PLN empregadas, conforme Leal et al. (2022). Para este trabalho, foram escolhidas métricas de 7 grupos, a saber: duas métricas descritivas, uma morfossintática, uma de densidade de padrões sintáticos, uma de complexidade sintática, uma de conectivos e uma de leiturabilidade. Os nomes das métricas escolhidas e seus respectivos objetivos estão arrolados a seguir.

- *words per sentences*: métrica descritiva que calcula a quantidade média de palavras por sentença. Quanto maior a métrica, maior é o tamanho das sentenças no texto, e, portanto, a complexidade textual. É importante pontuar que a noção de sentença programada para o NILC-Metrix não leva em consideração a máxima projeção do sintagma verbal Jackendoff (1982), mas sim a unidade iniciada por letra maiúscula e finalizada por ponto final, ponto de exclamação, ponto de interrogação ou reticências;

- *sentences per paragraph*: métrica descritiva que calcula a quantidade média de sentenças por parágrafo. Para esta métrica, a relação com a complexidade textual é menos evidente, já que esta última dependerá tanto do tamanho do parágrafo quanto do tamanho das sentenças que o compõem. De toda forma, sua aferição será importante para a análise da série histórica de textos, já que, a princípio, redações exemplares mais divulgadas possuem quantidades próximas (e, em alguma medida, padronizada) de frases nos parágrafos.
- *content words*: métrica morfossintática que calcula a proporção de palavras de conteúdo (substantivos, verbos, adjetivos, advérbios e palavras denotativas) em relação ao número de palavras total. Quanto maior a métrica, maior é o vocabulário exigido para a compreensão do texto e, portanto, maior a complexidade;
- *mean noun phrase*: métrica de densidade de padrões sintáticos que calcula a quantidade média de palavras que compõem os sintagmas nominais. Para o cálculo desta métrica, o texto é processado pelo LX-Parser, que identifica os constituintes sintáticos, dentre eles, os sintagmas nominais (SNs). Quanto maior a métrica, maior é o tamanho médio dos SNs, e, portanto, maior a complexidade textual;
- *words before main verb*: métrica de complexidade sintática que calcula a quantidade média de palavras antes do verbo principal. Quanto maior a métrica, mais informações precisam ser armazenadas na memória de trabalho, aumentando a complexidade textual;
- *conn ratio*: métrica de conectivos que calcula a quantidade de conectivos em relação à quantidade de palavras total. Por meio de uma lista de palavras pré-determinada, a ferramenta procura por conectivos aditivos, temporais, causais e lógicos. Quanto maior o uso de conectivos, mais simples tende a se tornar o entendimento do texto, diminuindo a complexidade. O uso da métrica se justifica pelo fato de haver uma competência específica para a avaliação da coesão (competência IV), que se dá, dentre outros fatores, através da análise dos conectivos intra e interparagrafais. Quanto maior a métrica, maior a coesão entre as partes do texto;
- *flesch*: métrica de leiturabilidade que considera o tamanho médio das palavras (calculado em número médio de sílabas) e sentenças (calculado em número médio de palavras). Quanto maior a métrica, menor a complexidade do texto.

Embora o índice flesch possa representar uma medida grosseira da leiturabilidade do texto — já que depende da acurácia do segmentador silábico e do tokenizador de palavras e sentenças — o índice mostrou resultados intrigantes ao longo da série. Para o cálculo, são utilizadas a média de palavras por sentenças e a média do número de sílabas por palavra no texto. A fórmula empregada é ajustada empiricamente.

$$F = 248,835 - [1,015 \times MPS] - [84,6 \times MSP] \quad (1)$$

Sendo F o índice flesch, MPS a média de palavras por sentença e MSP a média de sentenças por parágrafo.

Por conta de problemas para o processamento de longas sequências de arquivos de texto na ferramenta, optamos por não calcular o Índice de Honoré, que também é uma medida de leiturabilidade, mas que leva em consideração o número de tokens e de hápax legomena, isto é, palavras que são utilizadas uma única vez no texto.

O fato de a quantidade de palavras compiladas a cada ano ser diferente ao longo da série histórica implica na impossibilidade de comparação direta entre os valores das métricas. Por conta disso, foi feita a normalização de cada um dos resultados para cada 100 (cem) palavras. Isso assegura que possamos comparar os valores de uma métrica relativa a um mesmo número de palavras entre dois ou mais períodos diferentes.

3. Resultados e discussões

O número de palavras coletadas no corpus, até o momento, é de 37.395, com razão type/token de 0,160602. A partir do ano de 2012, embora a quantidade de textos nota máxima diminua, boa parte deles se encontram disponíveis na internet e em compilações feitas por plataformas de vestibulares. A coleta manual e a dificuldade de se encontrarem redações de determinados períodos explicam os diferentes números de textos compilados para cada edição do exame na Tabela 1.

Os textos e seus respectivos temas estão organizados a seguir. O corpus está disponível publicamente.[1]

Os dados do NILC-Metrix, do script e do Orange foram dispostos em gráficos, com o objetivo de facilitar a visualização da série histórica das notas. O primeiro gráfico a ser notado é o do número médio de palavras por ano, calculado pelo script, e que quase duplicou, saindo

[1] https://github.com/atilavital/corpus-redacao

Ano	Tema	Quantidade
2010	O trabalho na construção da dignidade humana	4
2011	Viver em rede no século XXI: os limites entre o público e o privado	3
2012	O movimento imigratório para o Brasil no século XXI	9
2013	Efeitos da implantação da lei seca no Brasil	10
2014	Publicidade infantil em questão no Brasil	10
2015	A persistência da violência contra a mulher	10
2016	Caminhos para combater a intolerância religiosa	10
2017	Desafio para a formação educacional de surdos no Brasil	10
2018	Manipulação do comportamento do usuário pelo controle de dados	10
2019	Democratização do acesso ao cinema	10
2020	O estigma associado às doenças mentais na sociedade brasileira	10

Tabela 1: Quantidade de textos coletados para cada tema.

de 226,5, em 2010, para 441,5, em 2020, conforme a Figura 2. Como o espaço para a escrita das redações não foi alterado ao longo dos anos — mantendo-se um limite de 30 (trinta) linhas em uma área sem alterações significativas —, é de se esperar que, ao se incrementar gradativamente a quantidade de palavras, mudanças graduais devam ser observadas em outras métricas linguísticas.

A média da razão type/token (Figura 3), por outro lado, como métrica importante para indicação da riqueza lexical ao longo dos textos, foi calculada pelo Orange e sugeriu uma diminuição ao longo dos anos, embora com uma taxa não tão evidente quanto a da média de palavras, conforme o valor de R^2. A diminuição da razão type/token pode ser o reflexo do aumento gradual da quantidade de palavras (tokens), conforme a Figura 2, sem acréscimos, na mesma proporção, de tipos diferentes de palavras (types). Sob esse ponto de vista, a riqueza lexical tende a diminuir ao longo da série estudada.

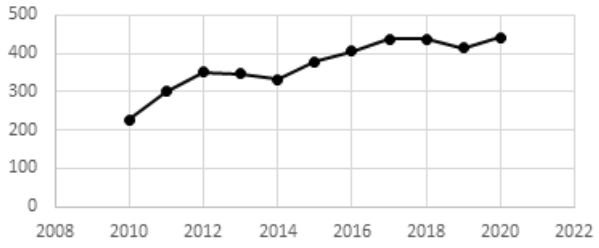

Figura 2: Média de palavras por ano. Fonte: elaborado pelo autor.

Como já salientado, a ferramenta NILC-Metrix compreende a noção de sentença como todos os elementos linguísticos posicionados entre uma letra maiúscula e um sinal de pontuação (ponto final, ponto de interrogação, exclamação ou reticências). As Figuras 4 e 5 nos mostram que tanto a média de sentenças por parágrafo

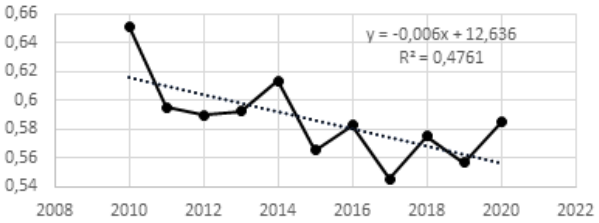

Figura 3: Evolução da média da razão type/token. Fonte: elaborado pelo autor.

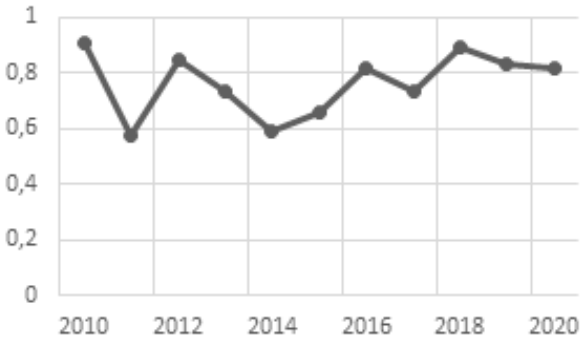

Figura 4: Média de sentenças por parágrafo.

quanto a média de palavras por sentenças não exibiram mudanças tão evidentes. A princípio, com um número de parágrafos inalterado e o mesmo espaço para a escrita do texto, o aumento do número de sentenças poderia indicar uma menor complexidade textual, já que a tendência seria a de haver mais sentenças, ainda que menores, no interior dos parágrafos; no entanto, essa conclusão só seria verdadeira caso a quantidade de palavras por sentença diminuísse, o que não se observa na Figura 5.

O tamanho médio dos sintagmas nominais, evidente na Figura 6, depende acurácia da anotação sintática do LX-Parser, ferramenta utilizada para a anotação sintática do NILC-Metrix.

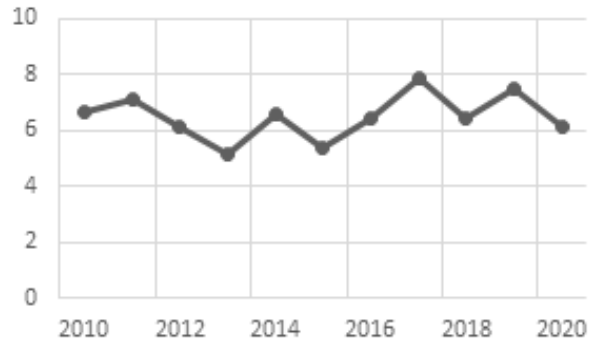

Figura 5: Média de palavras por sentença.

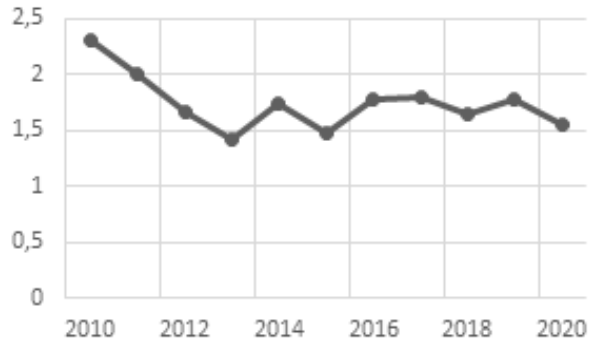

Figura 6: Tamanho médio dos SNs.

Pelo gráfico, podemos perceber que, a partir de 2015, o tamanho médio dos SNs tende a um descrescimento de 2010 a 2013, mas o padrão não se mantém até 2020.

Tanto a média de palavras de conteúdo (Figura 7) — medida pela biblioteca nlpnet (Fonseca & Rosa, 2013), do Python, modelo baseado em redes neurais e que faz etiquetagens semânticas e de classes de palavras — quanto a média de palavras antes do verbo principal (Figura 8) — que inclui medidas de anotação do LX-Parser (Silva et al., 2010) e de tokenização do Parser Palavras (Bick, 2000) — demonstraram um decrescimento.

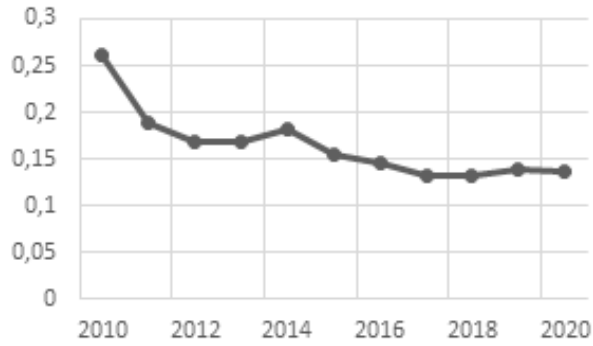

Figura 7: Média de palavras de conteúdo / palavras total.

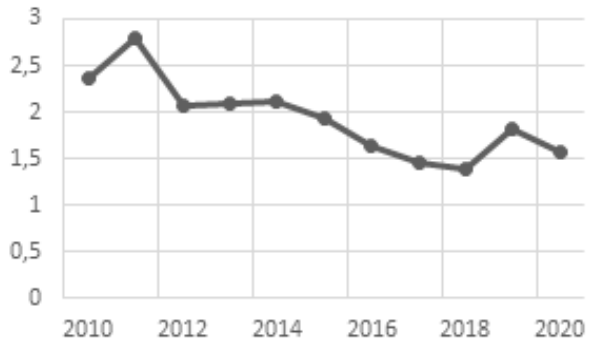

Figura 8: Média do nº de palavras antes do verbo principal.

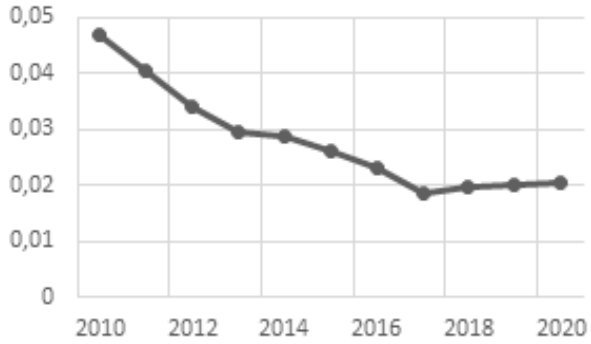

Figura 9: Média de conectivos em relação ao número de palavras.

No caso das palavras de conteúdo, é possível afirmar que o aumento do número de palavras total ao longo da série (Figura 2) desencadeia uma diminuição na proporção entre palavras significativas e gramaticais. As palavras acrescidas, por sua vez, parecem ocupar as posições sintáticas após o verbo principal, já que a média de palavras antes do verbo principal diminui ao longo dos anos (Figura 8). Neste ponto, é possível pensar em variadas explicações para a não variação no tamanho dos SNs: uma hipótese seria que o aumento de palavras no texto implicaria um maior número de sintagmas nominais com tamanhos próximos, o que evitaria um simples incremento da complexidade dentro dos SNs, cuja quantidade já parecia estar estável em relação ao tamanho e à informatividade do texto. É importante destacar que, para trabalhos futuros, o processamento de outros tipos de sintagmas seria valioso, de modo a esclarecer em quais constituintes, precisamente, as novas palavras tendem a se encaixar.

A Figura 9, por sua vez, representa a média de conectivos em relação ao número de palavras no texto, que diminuiu ao longo dos anos analisados, fazendo com que haja mais palavras para cada conectivo. Este dado não sugere que os textos passam a ser menos coesos, mas que o número de

elementos coesivos não aumenta na mesma taxa do número de palavras.

Além dos gráficos de complexidade textual que medem parâmetros sintáticos, semânticos e paragrafais, foi calculado o índice Flesch de leiturabilidade. Calculado automaticamente pelo NILC-Metrix, o índice tem o objetivo de verificar a correlação média entre tamanhos médios de palavras e sentenças. O comportamento da métrica ao longo dos anos pode ser visualizado na Figura 10.

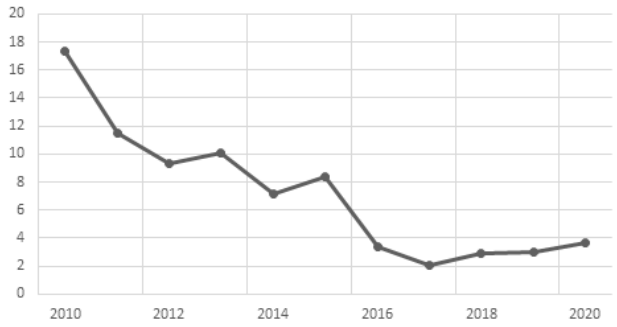

Figura 10: Evolução do índice de leiturabilidade Flesch.

Houve queda acentuada na média de leiturabilidade dos textos entre os anos de 2010 a 2017, o que implica numa maior complexidade textual e, portanto, numa menor facilidade de leitura. Embora o índice tenha apresentado resultados intrigantes e compatíveis com as métricas anteriores, é importante notar que os resultados dependem do desempenho do tokenizador, do sentenciador e do segmentador silábico, sendo que este último produz uma medida grosseira em relação ao tamanho médio das palavras do texto.

Ainda que incipientes, os resultados encontrados permitem que teçamos alguns comentários a respeito da série histórica de textos nota mil. O primeiro deles diz respeito ao aumento contínuo do número de palavras, que, em média, quase dobrou durante o período analisado, embora o espaço para a escrita do texto tenha se mantido constante. Os dados do aumento expressivo de tokens são enriquecidos com os dados de palavras de conteúdo, que, por sua vez, diminuem progressivamente, sugerindo que a relação entre as palavras significativas e o número total de palavras se dê em taxas diferentes: se o número de palavras total aumenta em taxa x , o número de palavras de conteúdo diminui em taxa y, tal que x é maior que y.

É importante notar que, aparentemente, nas condições em que há um limite rigoroso de linhas, como é o caso das redações, um aumento de palavras implica numa reorganização textual, impactando a estrutura das sentenças e o encadeamento coesivo. É precisamente nesse sentido, portanto, que seria esperado um aumento no número de palavras nos SNs, que, a princípio, enriqueceriam a descrição e os assuntos introduzidos nas sentenças. Por outro lado, a média de palavras antes do verbo principal se mostrou descrescente — mesmo que saibamos que a quantidade de palavras nas sentenças aumentou, em média. Esse fato pode sugerir que as sentenças se desenvolveram mais à direita do verbo, dentro do sintagma verbal e dos complementos.

Outro reflexo do aumento de palavras é a quantidade média de palavras para cada conectivo, que diminui na série histórica. Isso nos indica que o número de conectivos não cresce na mesma proporção do número de tokens, encaixando maiores quantidades de informação entre uma estratégia de conexão e outra, e, portanto, aumentando a complexidade coesiva. Buzato et al. (2021), ao analisar a ocorrência de operadores argumentativos em um corpus de redações de notas variadas, constatou alta frequência de uso de poucos operadores. Uma futura comparação poderia levar em conta os resultados de Buzato et al. (2021), na tentativa de encontrar aproximações e distanciamentos entre os dois corpora.

O índice de leiturabilidade Flesch, por sua vez, foi a métrica utilizada para verificar, sob um ponto de vista amplo, a complexidade de cada um dos textos. Como mostramos acima, acreditamos que boa parte dos comportamentos das métricas podem ser explicados, numa primeira análise, pelo aumento do número de palavras, embora esse fenômeno possa ter criado outras particularidades linguísticas que não foram captadas pelas medidas realizadas neste trabalho. Isso não é diferente no caso da leiturabilidade, já que, ao considerar os valores de *MPS* e *MSP* — dois valores que tendem a crescer com o aumento de tokens no texto —, é esperado que seu índice diminua ao longo dos anos analisados.

No sentido de tecer possíveis explicações para a diminuição drástica no número de redações nota mil entre os anos de 2011 e 2013, seria necessária, também, i) uma análise das métricas em conjunto com a evolução dos critérios de correção ao longo dos anos e ii) uma maior compilação de textos englobando, inclusive, redações de diferentes níveis. Sob o ponto de vista da logística das correções, as discrepâncias aceitas entre as notas dos dois primeiros corretores passaram de até 300 (trezentos) pontos, em 2011, para 200 (duzentos) pontos, em 2012, e, por fim, 100 (cem) pontos em 2013, valor considerado até os dias de hoje. Dessa forma, é possível que, com a diminuição

progressiva da tolerância, os textos passem a ser mais re-corrigidos, o que diminuiria a probabilidade de haver um elevado número de redações nota máxima. Embora a causalidade entre as duas variáveis não possa ser comprovada neste trabalho, é possível que a diminuição das notas mil não esteja relacionada com fatores puramente linguísticos, mas com elementos relativos à metodologia de correção.

4. Considerações finais

Este trabalho teve como objetivo fazer uma breve descrição de um corpus de redações nota mil a partir de métricas textuais, procurando tendências linguísticas ao longo do período de 2010 a 2020, em que houve a drástica diminuição na atribuição de notas máximas bela banca de correções do Enem. Neste momento, ainda não é possível afirmar, com precisão, os motivos de tal diminuição, mas são observáveis diferenças quantitativas entre textos considerados exemplares ao longo da série. Para isso, lançamos mão de ferramentas computacionais, dentre elas, o analisador de complexidade textual NILC-Metrix (Leal et al., 2022), disponível gratuitamente na internet. Como resultados encontrados, pudemos ressaltar o aumento de palavras e, de um modo geral, o aumento da complexidade textual, reportada pelas métricas de conectivos e de leiturabilidade.

Na esteira contemporânea da Linguística de Corpus, objetivamos, em momento oportuno, a disponibilização dos textos, que constituirão o primeiro corpus de redações nota mil avaliadas pela banca de corretores do Enem.

Agradecimentos

Este trabalho foi desenvolvido durante a disciplina de Linguística de Corpus e Computacional, da pós-graduação em Linguística da Universidade Federal de Minas Gerais. Agradeço à equipe do NILC pela disponibilização da plataforma NILC-Metrix, à Dr[a]. Heliana Mello pela oferta da disciplina e à revisão da Linguamática.

Referências

Alexopoulou, Theodora, Marije Michel, Akira Murakami & Detmar Meurers. 2017. Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques. *Language Learning* 67(S1). 180–208. 10.1111/lang.12232.

Amorim, Evelin & Adriano Veloso. 2017. A multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. Em *Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 94–102.

Berber Sardinha, Tony. 2000. Linguística de corpus: histórico e problemática. *D.E.L.T.A* 16(2). 323–367. 10.1590/S0102-44502000000200005.

Bertucci, Roberlei Alves, Andréa Jacqueline Malheiros & Wanderlei de Souza Lopes. 2020. Ocorrências de anáforas encapsuladoras em redações do enem. *Filologia e Linguística Portuguesa* 22(1). 81–102. 10.11606/issn.2176-9419.v22i1p81-102.

Bick, Eckhard. 2000. *The parsing system PALAVRAS: Automatic grammatical analysis of Portuguese in a constraint grammar framework*. Aarhus University Press.

Branco, António, João Rodrigues, Francisco Costa, João Silva & Rui Vaz. 2014. Rolling out text categorization for language learning assessment supported by language technology. Em *Computational Processing of the Portuguese Language (PROPOR)*, 256–261. 10.1007/978-3-319-09761-9_29.

Brasil. 2020. *A redação do Enem 2020: cartilha do participante*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Brasília, DF.

Buzato, Dalmo, Elias Victor de Jesus Cardoso Machado, Priscilla Tulipa da Costa & Suélen Érica Costa Silva. 2021. Operadores argumentativos em redações modelo enem: uma análise baseada em corpus. *Scientia Prima* 7(1). e97.

Cançado, Márcia, Luana Amaral, Amorin Evelin, Veloso Adriana & Heliana Mello. 2020. Subjetividade em correções de redações: detecção automática através de léxico de operadores de viés linguístico. *Linguamática* 12(1). 63–79. 10.21814/lm.12.1.313.

Crossley, Scoot A. & Danielle S. McNamara. 2012. Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35(2). 115–135. 10.1111/j.1467-9817.2010.01449.x.

Cruz, Daniel Ribeiro da, Rafaela Gonçalves Ulian, Robson Faleiros Ribeiro & Sheila Fernandes Pimenta e Oliveira. 2021. REDAÇÃO NOTA 1000: argumentos de propostas de intervenções em redações do ENEM 2009–2018. *Revista Eletrônica de Letras* 14(14). on–line.

Fonseca, Erick, Ivo Medeiros, Dayse Kamikawachi & Alessandro Bokan. 2018. Automatically grading Brazilian student essays. Em *Computational Processing of the Portuguese Language (PROPOR)*, 170–179.

Fonseca, Erick Rocha & João Luís G. Rosa. 2013. Mac-Morpho revisited: Towards robust part-of-speech tagging. Em *9th Brazilian Symposium in Information and Human Language Technology (STIL)*, 98–107.

Graesser, Arthur C., Danielle S. McNamara, Max M. Lowerse & Zhiqiang Cai. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36. 193–202. 10.3758/BF03195564.

Jackendoff, Ray. 1982. X syntax: A study of phrase structure. *Journal of Linguistics* 18. 409–497.

Leal, Sidney Evaldo, Magali Sanches Duran, Carolina Evaristo Scarton, Nathan Siegle Hartmann & Sandra Maria Aluísio. 2022. Nilc-metrix: assessing the complexity of written and spoken language in brazilian portuguese. ArXiv cs.CL.

Marinho, Jeziel, Rafael Anchiêta & Raimundo Moura. 2021. Essay-BR: a brazilian corpus of essays. Em *Anais do III Dataset Showcase Workshop*, 53–64. 10.5753/dsw.2021.17414.

Silva, João, António Branco, Sérgio Castro & Ruben Reis. 2010. Out-of-the-box robust parsing of Portuguese. Em *Computational Processing of the Portuguese Language (PROPOR)*, 75–85.

Westerlund, Marcus. 2019. *Correlations between textual features and grades on the Swedish national exam in English: A Coh-Metrix analysis*: Stockholms Universitet. Tese de Mestrado.

http://www.linguamatica.com/

linguaMATICA